Análise Discriminante com mistura de variáveis categóricas e contínuas

Renê Sanda

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA OBTENÇÃO DO GRAU DE MESTRE
EM
ESTATÍSTICA

Área de Concentração: Estatística
Orientador: Prof. Dr. Abraham Laredo Sicsu

– São Paulo, Maio de 1990 –

À memória
de minha mãe

A G R A D E C I M E N T O S

Desejo agradecer a todos aqueles que contribuíram para a concretização deste trabalho, em especial:

- ao meu orientador Abraham Laredo Sicsu pelo apoio, dedicação e valiosas sugestões;
- ao Banco do Brasil S.A. pela confiança e incentivos proporcionados;
- e, finalmente, à minha conselheira, TEXpert, revisora e, na fase final deste trabalho, para a minha felicidade, esposa Meire.

PREFÁCIO

Análise Discriminante compõe-se de técnicas estatísticas multivariadas com dois objetivos básicos:

- determinar uma regra para classificar uma observação de origem desconhecida em uma das populações previamente definidas;
- discriminar (separar) as populações determinando quais foram as variáveis mais influentes.

Neste trabalho vamos nos restringir ao problema de classificação. Consideráveis avanços têm sido feitos nos últimos anos para a solução deste importante problema em Estatística. A maior parte dos livros e artigos que tratam desse assunto assume, implicitamente, que as variáveis discriminadoras (variáveis associadas às observações) são exclusivamente contínuas, ou exclusivamente categóricas. Entretanto, freqüentemente, é necessário trabalhar com os dois tipos de variáveis simultaneamente.

O objetivo deste trabalho é apresentar os métodos mais importantes de classificação quando temos uma mistura de variáveis contínuas e categóricas. Por simplicidade nos referiremos a isso como variáveis misturadas.

Grande parte dos exemplos considerados neste trabalho é da área financeira em virtude da maior experiência e interesse do autor nesta área. Atente-se, porém, ao fato de que a quase totalidade dos dados analisados são fictícios.

Este trabalho está organizado da seguinte forma:

O Capítulo 1 introduz o problema da Análise Discriminante apresentando alguns resultados clássicos. Os principais métodos de avaliação das regras discriminantes também são apresentados. Uma breve discussão sobre a necessidade de métodos específicos para variáveis misturadas encerra o capítulo.

O Capítulo 2 trata do *location model* desenvolvido principalmente por Krzanowski em diversos artigos. Este método consiste em utilizar diferentes regras discriminantes, uma para cada combinação das variáveis categóricas.

No Capítulo 3, o modelo logístico é descrito. Várias distribuições, inclusive algumas de variáveis misturadas, satisfazem a este modelo. A simplicidade da regra discriminante, aliada à disponibilidade de *software*, são fatores que fazem deste método um dos mais utilizados na prática.

O Capítulo 4 apresenta, de forma sucinta, três métodos não-paramétricos que podem ser utilizados com variáveis misturadas: o método do Núcleo (*Kernel*), o método dos k-Vizinhos Mais Próximos (k-*NN*) e o método da Árvore Binária de Classificação.

Finalmente, no Capítulo 5, são dados alguns resultados de comparações entre os métodos e um exemplo de aplicação da Análise Discriminante, denominada na área financeira de *credit scoring*.

Incluímos, no apêndice, a listagem de alguns programas que foram desenvolvidos para este trabalho. Da união destes programas nasceu o que ousamos chamar de "*software*" AD. Encorajamos os leitores a solicitarem uma cópia.

Maio de 1990

Renê Sanda

ÍNDICE

Pág.

Capítulo 1 - Idéias Básicas em Análise Discriminante 1

1.1 - Planejamento Amostral 3

1.2 - Métodos de Avaliação de uma Regra Discriminante 4

1.3 - Algumas Regras Discriminantes 7

1.3.1 - Regra de Bayes 7

1.3.2 - Regra do Mínimo Custo Total Esperado 9

1.4 - A Necessidade de Métodos Específicos para Variáveis Misturadas 11

1.5 - Sugestões para Futuros Trabalhos 13

Capítulo 2 - O Método de Krzanowski 15

2.1 - O Método de Krzanowski e a Regra de Classificação 16

2.1.1 - A Distância de Matusita 22

2.2 - Estimação dos Parâmetros 25

2.2.1 - Estimadores Usuais 25

2.2.2 - Estimação por Modelos 26

2.3 - Seleção de Variáveis Categóricas 29

Capítulo 3 - Discriminação Logística 33

3.1 - Formulação do Problema 34

3.2 - Funções de Verossimilhança 38

3.2.1 - Amostas Conjuntas 38

3.2.2 - Amostras Separadas 40

3.3 - Estimação dos Parâmetros 42

3.3.1 - Máxima Verossimilhança 42

3.3.2 - Procedimento Discriminante Normal 44

3.3.3 - Transformação Logística Empírica 46

3.4 - Existência e Unicidade do Estimador de Máxima Verossimilhança 53

3.4.1 - Separação Completa 54

3.4.2 - Separação Quase Completa 56

3.4.3 - Sobreposição 57
3.4.4 - Separação Parcial e Quase Parcial 59
3.5 - Adequação do Modelo 60
3.6 - Seleção de Variáveis 66

Capítulo 4 - Métodos Não-Paramétricos 69

4.1 - O Método do Núcleo 70
4.1.1 - Variáveis Contínuas 71
4.1.2 - Variáveis Qualitativas Nominais 78
4.1.3 - Variáveis Qualitativas Ordinais 82
4.1.4 - Variáveis Misturadas 84
4.2 - Método dos k-Vizinhos Mais Próximos 86
4.3 - Método da Árvore Binária de Classificação 90
4.3.1 - Regra de Partição 94
4.3.2 - O Tamanho Ótimo da Árvore 101
4.3.3 - Exemplo 105

Capítulo 5 - Comparação dos Métodos e um Exemplo de Aplicação 113

5.1 - Comparação dos Métodos 114
5.2 - Um Exemplo de Aplicação: *Credit Scoring* 116
5.2.1 - Descrição dos Dados e Critérios Utilizados 119
5.2.2 - Resultados 122
5.2.3 - Comentários e Conclusões 123

Apêndice A - Detalhes sobre a Aplicação do Método de Newton-Raphson em Discriminação Logística

Apêndice B - Análise Descritiva das Variáveis que Compõem o Exemplo da Seção 5.2

Apêndice C - Distribuição dos Escores para os Métodos Abordados no Exemplo da Seção 5.2

Apêndice D - Listagem do Programa AD

Referências Bibliográficas

Capítulo 1

IDÉIAS BÁSICAS EM ANÁLISE DISCRIMINANTE

O objetivo principal da Análise Discriminante é construir regras que possibilitem classificar uma observação multivariada $\mathbf{x}$ em uma das g populações $\Pi_1, \ldots, \Pi_g$, conhecidas *a priori*, a partir de amostras de observações classificadas corretamente.

Este é um problema bastante freqüente em diversas áreas, como podemos constatar pelos seguintes exemplos:

(i) **Medicina.** Com base nos sintomas do paciente e em alguns exames preliminares ($\mathbf{x}$), um médico precisa decidir se o paciente é portador de uma certa doença (Π_1) ou não (Π_2).

(ii) **Finanças.** Um analista de crédito deve decidir, com base nas informações ($\mathbf{x}$) de uma ficha cadastral, se concede ou não o crédito ao cliente. O crédito será concedido se o cliente for classificado como bom pagador (Π_1).

(iii) **Botânica.** Uma certa planta deve ser classificada em uma das espécies pré-determinadas ($\Pi_1, \ldots, \Pi_g$), baseando-se apenas nas suas características morfológicas ($\mathbf{x}$).

Outros exemplos de aplicações podem ser encontrados a partir de uma leitura atenta dos títulos das 579 referências contidas no livro de Lachenbruch (1975).

O primeiro trabalho conhecido nesta área deve-se a Fisher (1936) que introduziu o termo "discriminação" para o problema. Atualmente, "discriminação" é utilizado mais no sentido de "separação", sendo mais apropriado o termo "classificação" para o nosso trabalho.

Finalizamos esta seção apresentando a notação geral que será utilizada neste trabalho. Notações específicas de cada seção ou capítulo serão introduzidas oportunamente.

Notação geral:

g = número de populações;

p = número de variáveis discriminadoras contínuas;

q = número de variáveis discriminadoras categóricas;

r = número total de variáveis discriminadoras $(p+q)$;

n = tamanho da amostra;

c_j = número de categorias diferentes que a j-ésima variável categórica pode assumir;

Π_i = variável indicativa da i-ésima população;

$\mathbf{Z}$ = vetor $(q \times 1)$ de variáveis aleatórias categóricas;

$\mathbf{Y}$ = vetor $(p \times 1)$ de variáveis aleatórias contínuas;

$\mathbf{X}$ = vetor $(r \times 1)$ de variáveis aleatórias $\left(\mathbf{X}' = (\mathbf{Z}' \ \mathbf{Y}')\right)$;

$\mathbf{z}$ = vetor $(q \times 1)$ de observações categóricas;

$\mathbf{y}$ = vetor $(p \times 1)$ de observações contínuas;

$\mathbf{x}$ = vetor $(r \times 1)$ de observações $\left(\mathbf{x}' = (\mathbf{z}' \ \mathbf{y}')\right)$;

$\mathbf{D}$ = matriz $(n \times r)$ de dados $\left(\mathbf{D} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')\right)$;

Ω = espaço amostral de $\mathbf{X}$;

E_i = conjunto dos índices das observações de $\mathbf{D}$ provenientes da população Π_i;

n_i = número de elementos de E_i;

p_i = probabilidade *a priori* de $\mathbf{x}$ pertencer à população Π_i;

R_i = região de Ω cujos pontos $\mathbf{x}$ serão classificados como pertencentes à população Π_i.

1.1 - Planejamento Amostral

A forma como as amostras são obtidas influi diretamente na construção de um regra de classificação. Neste trabalho vamos considerar dois métodos de amostragem:

- amostragem conjunta;
- amostragem separada.

Diremos que o planejamento é de amostragem conjunta se a amostra for selecionada aleatoriamente da mistura das populações, ou seja, se os tamanhos das amostras de cada população $(n_1, \ldots, n_g)$ forem variáveis aleatórias. Por outro lado, denominaremos de amostragem separada se os n_i $(i = 1, \ldots, g)$ tiverem sido fixados antes do experimento ter sido realizado.

Vamos definir p_i como sendo a probabilidade *a priori* da observação $\mathbf{x}$ pertencer à população Π_i, isto é:

$$p_i = \Pr(\mathbf{x} \in \Pi_i) \qquad (i = 1, \ldots, g)$$

com $\sum_{i=1}^{g} p_i = 1$.

A estimação destas probabilidades no caso de amostragem conjunta é dada por $\widehat{p_i} = n_i/n \quad (i = 1, \ldots, g)$. No caso de amostragem separada a estimação de p_i não é possível a partir das amostras. Iremos supor que a estimação dessas probabilidades foi feita subjetivamente pelo pesquisador, ou a partir de amostras conjuntas obtidas anteriormente. Na prática, quando não existe nenhuma informação sobre essas probabilidades, é comum utilizar $\widehat{p_i} = 1/g$ $(i = 1, \ldots, g)$.

Para facilitar a notação omitiremos o sinal de estimador em $\widehat{p_i}$ denotando-o simplesmente de p_i $(i = 1, \ldots, g)$.

1.2 - Métodos de Avaliação de uma Regra Discriminante

Uma regra de classificação é uma função $d(\mathbf{x})$ definida em Ω (espaço amostral de $\mathbf{X}$) tal que para cada $\mathbf{x} \in \Omega$, $d(\mathbf{x})$ é igual a um dos números $1, 2, \ldots, g$. Dado $\mathbf{x}$, se $d(\mathbf{x}) = i$ então $\mathbf{x}$ é classificado como sendo pertencente à população Π_i. Esta função define uma partição de Ω em $R_1, \ldots, R_g$ com

$$R_i = \{\mathbf{x} \in \Omega \mid d(\mathbf{x}) = i\} \qquad (i = 1, \ldots, g)$$

O problema é verificar quão boa é a regra $d(\mathbf{x})$, isto é, qual a é taxa de classificação incorreta que $d(\mathbf{x})$ provoca. Esta taxa que denominaremos de taxa de erro real é definida como sendo a probabilidade de que uma observação $\mathbf{x}$ não seja classificada corretamente, isto é,

$$e_{real} = \sum_{i=1}^{g} P\,(\,d(\mathbf{x}) \neq i\,,\,\mathbf{x} \in \Pi_i\,) \qquad (1-1)$$

Há várias formas de estimar a taxa de erro real. As mais utilizadas são:

- taxa de erro aparente (e_{ap});
- taxa de erro da amostra teste (e_{at});
- taxa de erro da validação cruzada $(e_{vc}(V))$.

A e_{ap} é obtida reclassificando-se a própria amostra utilizada para construir a regra. Definindo m_i como o número de observações de Π_i classificadas incorretamente, temos:

$$\begin{aligned} e_{ap} &= \sum_{i=1}^{g} \widehat{P}\,(\,d(\mathbf{x}) \neq i, \mathbf{x} \in \Pi_i\,) \\ &= \sum_{i=1}^{g} \widehat{P}\,(\,d(\mathbf{x}) \neq i \,|\, \mathbf{x} \in \Pi_i\,)\; p_i \\ &= \sum_{i=1}^{g} \frac{m_i}{n_i}\, p_i \qquad (1-2) \end{aligned}$$

De acordo com Efron (1986) a taxa de erro aparente tende a subestimar a taxa de erro real e este problema é mais grave a medida que o tamanho da amostra é menor. Isto ocorre, basicamente, porque as mesmas observações utilizadas para construir a regra de classificação são utilizadas para avaliá-la.

Uma forma para estimar a taxa de erro real, viável só quando temos amostras suficientemente grandes, é através da taxa de erro da amostra teste e_{at}. Neste caso, a amostra é dividida (de preferência aleatoriamente) em duas partes que denominaremos de amostra-desenvolvimento e amostra-teste. A amostra-desenvolvimento é utilizada para gerar a re-

gra de classificação enquanto que a amostra-teste é utilizada somente para avaliá-la. A expressão é semelhante a obtida para se calcular e_{ap}:

$$e_{at} = \sum_{i=1}^{g} \frac{m_{i2}}{n_{i2}} p_i \qquad (1-3)$$

onde n_{i2} é o número de observações na amostra-teste pertencentes à população Π_i e m_{i2} é o número de observações da população Π_i nessa amostra classificadas incorretamente.

A taxa de erro da validação cruzada é uma das formas mais precisas para estimar a taxa de erro real no caso de amostras pequenas. O procedimento é uma extensão do método desenvolvido para a obtenção da e_{at}. Divide-se a amostra $\{L\}$ em V subamostras $\{L_1\},\ldots,\ \{L_V\}$ de tamanhos $N_1,\ldots,\ N_V$ com $N_i \approx n/V$. Para cada v, $v = 1, \ldots, V$, constrói-se a regra a partir de $\{L\} - \{L_v\}$ e avalia-se esta regra classificando-se $\{L_v\}$. Assim, a expressão da taxa de erro da validação cruzada é:

$$e_{vc}(V) = \frac{1}{V} \sum_{v=1}^{V} \sum_{i=1}^{g} \frac{m_{iv}}{n_{iv}} p_i \qquad (1-4)$$

onde n_{iv} é o número de observações da população Π_i na amostra $\{L_v\}$ e m_{iv} é o número de observações da população Π_i em $\{L_v\}$ classificadas incorretamente.

Quando $V = n$, a taxa de erro da validação cruzada é a taxa de erro *leaving one out* desenvolvido por Lachenbruch e Mickey (1968). Referiremo-nos a esta taxa como taxa de erro do método da exclusão.

A utilização de uma ou outra forma para estimar a taxa de erro real depende, essencialmente, do tamanho da amostra e da regra de classificação. Regras geradas pelo método do Núcleo (Seção 4.1) e pelo método da Árvore Binária de Classificação (Seção 4.3) são obtidas iterativamente, em virtude disto, a estimação da taxa de erro real pela taxa de erro da validação cruzada poderá, eventualmente, requerer um tempo de processamento computacional impraticável.

1.3 - Algumas Regras Discriminantes

Na seção anterior vimos como estimar a taxa de erro real provocado por uma regra $d(\cdot)$. O problema agora é determinar qual deve ser a regra $d(\cdot)$ a partir de algum critério. Dois critérios serão abordados neste trabalho: minimização da taxa de erro real (Regra de Bayes) e minimização do custo total esperado de classificação incorreta.

1.3.1 - Regra de Bayes

Este trabalho está centrado na regra de Bayes (ver, por exemplo: Mardia, Kent e Bibby (1979) - pág. 305) que define as partições da seguinte forma:

$$R_i = \{\mathbf{x} \mid \Pr(\mathbf{x} \mid \Pi_i)\, p_i = \max_{h=1,\ldots,g} \Pr(\mathbf{x} \mid \Pi_h)\, p_h\} \qquad (i = 1, \ldots, g)$$

ou seja, a regra de classificação de Bayes, que denominaremos de regra ótima de classificação, é a seguinte:

$$\text{Dado } \mathbf{x}, \quad \text{se } \Pr(\mathbf{x} \mid \Pi_i)\, p_i = \max_{h=1,\ldots,g} \Pr(\mathbf{x} \mid \Pi_h)\, p_h \quad \text{então, } d(\mathbf{x}) = i$$

Se o máximo for atingido por várias populações classificamos $\mathbf{x}$ de forma aleatória nestas populações.

Pode-se mostrar que esta regra é equivalente à obtida minimizando-se a taxa de erro real (Breiman, Friedman, Olshen e Stone (1984) - pág. 13).

Alguns casos particulares importantes desta regra são os seguintes:

(i) $(\mathbf{X} \mid \Pi_i) \sim \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$.

Como $\Pr(\mathbf{x} \mid \Pi_i) = |2\,\Pi\,\boldsymbol{\Sigma}|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\}$, decorre que a regra ótima de classificação é:

$$\text{Dado} \quad \mathbf{x}, \qquad \text{se} \quad d_i^2 - 2\,ln\,p_i = \min_{h=1,\ldots,g}(d_h^2 - 2\,ln\,p_h) \qquad \text{então,} \quad \mathbf{x} \in R_i$$

onde $d_i^2 = (\mathbf{x}-\boldsymbol{\mu}_i)'\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}-\boldsymbol{\mu}_i)$.

Por ser uma abreviatura consagrada na literatura nos referiremos a esta regra como LDF (*Linear Discriminant Function*).

No caso particular em que $g = 2$ esta regra simplifica-se para:

$$\begin{aligned} &\text{Dado} \quad \mathbf{x}, \qquad \text{se} \quad (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}\left[\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)\right] \geq \ln(p_2/p_1) \qquad \text{então,} \quad \mathbf{x} \in R_1 \\ &\qquad\qquad \text{caso contrário,} \quad \mathbf{x} \in R_2 \end{aligned} \tag{1-5}$$

Esta última regra é matematicamente idêntica à proposta por Fisher (1936). A regra de Fisher para $g > 2$ coincide com a regra da LDF somente quando as médias populacionais forem colineares. Não apresentamos a generalização da Regra de Fisher por não ser comumente usada como regra de classificação. Ao leitor interessado sugerimos Johnson e Wichern (1982) - páginas 512 a 517.

(ii) $(\mathbf{X} \mid \Pi_i) \sim \mathrm{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.

$$\text{Dado} \quad \mathbf{x}, \qquad \text{se} \quad \ln|\Sigma_i| + D_i^2 - 2\,ln\,p_i = \min_{h=1,\ldots,g}(ln\,|\Sigma_h| + D_h^2 - 2\,ln\,p_h) \qquad \text{então,} \quad \mathbf{x} \in R_i$$

onde $D_i^2 = (\mathbf{x}-\boldsymbol{\mu}_i)'\,\boldsymbol{\Sigma}_i^{-1}\,(\mathbf{x}-\boldsymbol{\mu}_i)$.

Esta regra é conhecida como regra da discriminação quadrática (QDF - *Quadratic Discriminant Function*).

Os estimadores de máxima verossimilhança são comumente utilizados nos casos em que os parâmetros são desconhecidos.

1.3.2 - Regra do Mínimo Custo Total Esperado

Há situações em que o objetivo não é criar uma regra que minimize o número de classificações incorretas. Estas situações se caracterizam quando os custos de se classificar incorretamente em uma ou outra população são diferentes. Por exemplo, o custo financeiro de se deferir um crédito a um inadimplente é maior do que o custo de se negar a um bom pagador.

Para estas situações recomenda-se a utilização da regra que minimiza o custo total esperado (ver, por exemplo, Choi (1986)) cujas partições são dadas por:

$$R_h = \left\{ \mathbf{x} \mid \sum_{\substack{i=1 \\ i \neq h}}^{g} p_i \, c(h|i) \, \Pr(\mathbf{x} \mid \Pi_i) \leq \sum_{\substack{i=1 \\ i \neq t}}^{g} p_i \, c(t|i) \, \Pr(\mathbf{x} \mid \Pi_i), \qquad (t = 1, \ldots, g; \; t \neq h) \right\}$$

onde $c(h|i)$ é o custo de se classificar na população Π_h uma observação que, na verdade, provem de Π_i $(h, i = 1, \ldots, g)$.

A regra de classificação consiste em classificar em Π_i toda observação pertencente a R_i.

No caso particular em que $g = 2$ esta regra simplifica-se para:

$$\text{Dado} \quad \mathbf{x}, \quad \text{se} \quad \frac{\Pr(\mathbf{x}\,|\,\Pi_1)}{\Pr(\mathbf{x}\,|\,\Pi_2)} \geq \frac{p_2\, c(1|2)}{p_1\, c(2|1)} \quad \text{então,} \quad \mathbf{x} \in R_1 \qquad (1-6)$$

$$\text{caso contrário,} \quad \mathbf{x} \in R_2$$

Uma forma didática de apresentar esta regra de classificação é reescrever $(1-6)$ em função da probabilidade *a posteriori* de $\mathbf{x} \in \Pi_1$, ou seja, como:

$$\frac{\Pr(\mathbf{x}\,|\,\Pi_1)}{\Pr(\mathbf{x}\,|\,\Pi_2)} = \frac{\Pr(\Pi_1\,|\,\mathbf{x})\,p_2}{\Pr(\Pi_2\,|\,\mathbf{x})\,p_1} = \frac{\Pr(\Pi_1\,|\,\mathbf{x})}{1-\Pr(\Pi_1\,|\,\mathbf{x})}\,\frac{p_2}{p_1}$$

Substituindo em (1-6) temos:

$$\frac{\Pr(\Pi_1\,|\,\mathbf{x})}{1-\Pr(\Pi_1\,|\,\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)} \qquad \Rightarrow$$

$$\Rightarrow \quad \text{se} \quad \Pr(\Pi_1\,|\,\mathbf{x}) \geq \frac{c(1|2)}{c(1|2)+c(2|1)} \quad \text{então,} \quad \mathbf{x} \in R_1 \qquad (1-7)$$

$$\text{caso contrário,} \quad \mathbf{x} \in R_2$$

Denominaremos de ponto de corte ao valor assumido pela expressão à direita da inequação anterior.

Observe que se os custos de classificação incorreta forem todos iguais a uma constante, a regra do mínimo custo total esperado torna-se idêntica à regra de Bayes.

1.4 - A Necessidade de Métodos Específicos para Variáveis Misturadas

Um procedimento comum quando um pesquisador defronta com um problema de Análise Discriminante com variáveis misturadas tem sido o de transformar as variáveis qualitativas em binárias * e utilizar um programa baseado na LDF disponível em vários pacotes estatísticos (BMDP, SAS, SPSS, etc.). Este procedimento pode levar a resultados satisfatórios, entretanto, procedimentos específicos para variáveis misturadas tendem a suplantá-los.

Outro procedimento consiste em categorizar as variáveis quantitativas e utilizar técnicas discriminantes específicas para este tipo de variável (ver Hills (1967), Goldstein e Dillon (1978)). Uma forma de categorização que preserva o máximo poder discriminante das variáveis contínuas é descrito em Cochran e Hophins (1961). A maior crítica a estes procedimentos é a perda de informação envolvida ao se categorizar.

Alternativamente, pode-se efetuar uma Análise Discriminante para as variáveis quantitativas e outra para as variáveis qualitativas, mas a interpretação final fica extremamente dificultada. Além disso, este procedimento desconsidera eventuais correlações existentes entre variáveis quantitativas e variáveis qualitativas.

Enfim, como nenhum desses procedimentos é totalmente satisfatório, é de interesse verificar como atuam os métodos que foram criados ou adaptados para serem utilizados com variáveis misturadas.

Na realidade os métodos constantes neste trabalho podem ser utilizados, a princípio, nos seguintes tipos de misturas:

---

* Quando a estrutura de ordem é considerada muito importante, atribui-se valores às categorias das variáveis qualitativas ordinais (em geral 0, 1, 2,...), e trata-se essas variáveis como se fossem contínuas. O procedimento é altamente discutível.

- categóricas e contínuas
  (métodos de Krzanowski, do Núcleo e dos k-Vizinhos Mais Próximos);

- binárias e contínuas
  (LDF e QDF);

- binárias, discretas e contínuas
  (Discriminação Logística);

- categóricas, discretas e contínuas
  (método da Árvore Binária de Classificação).

Variáveis qualitativas ordinais poderão ser incorporadas aos métodos tratando-as como nominais (categóricas) perdendo-se, dessa forma, a informação contida na estrutura de ordem. Se for necessário, variáveis qualitativas assumindo c valores distintos ($c > 2$) poderão ser transformadas em $c - 1$ variáveis binárias. Denominaremos a este procedimento, comum em regressão múltipla, de reparametrização.

Finalmente, variáveis discretas poderão ser tratadas como contínuas ou como categóricas dependendo do número de valores possíveis que cada variável puder assumir e da importância dada a sua estrutura de ordem.

É importante mencionar que a possibilidade de uma variável ser transformada em sua estrutura para poder ser incorporada a um determinado método, não significa, absolutamente, que este método vá utilizar toda a informação possível desta variável.

Os métodos, por nós considerados como os mais importantes, foram classificados conforme o nível de suposições feitas sobre as distribuições das variáveis em três tipos:

- métodos paramétricos;

- métodos parcialmente paramétricos;

- métodos não paramétricos.

O método de Krzanowski (Capítulo 2) é, tipicamente, um método paramétrico uma vez que a distribuição de $(\mathbf{X}|\Pi_i)$ será especificada a menos de alguns parâmetros. Tam-

bém a LDF e QDF são métodos paramétricos. Já a Discriminação Logística (Capítulo 3) é um método considerado parcialmente paramétrico (Anderson (1982)), pois somente o $\ln\{\Pr(\mathbf{x} \mid \Pi_i)/\Pr(\mathbf{x} \mid \Pi_g)\}$ será suposto linear. Finalmente, os métodos não paramétricos não fazem nenhuma suposição sobre tais distribuições. O método do Núcleo (Seção 4.1) e método dos k-Vizinhos Mais Próximos (Seção 4.2) procuram obter estimativas das densidades. O método da Árvore Binária de Classificação (Seção 4.3) utiliza uma abordagem heurística.

1.5 Sugestões para Futuros Trabalhos

Apesar de ser um problema bastante freqüente na prática, métodos multivariados para variáveis misturadas só recentemente se tornaram assunto relativamente constante nas revistas especializadas. Em Análise Discriminante, muitas questões ainda permanecem em aberto, especialmente no que se refere ao desempenho de algumas extensões naturais dos métodos apresentados.

A seguir, apresentamos algumas sugestões para futuros trabalhos, recomendando ao leitor, para maior clareza, que as leia somente após tomar conhecimento dos capítulos indicados entre parênteses no início de cada tópico.

- (1) Efron (1986) apresentou artigo sobre o vício da taxa de erro aparente na Discriminação Logística (Capítulo 3) com variáveis contínuas. A sugestão é estudar o mesmo vício num caso simples de mistura de variáveis, por exemplo, X_1 variável binária e X_2 normal, independentes;

- (2 e 4) O modelo de Krzanowski assume que em cada casela, determinadas pelas variáveis categóricas, as variáveis contínuas tenham distribuição normal multivariada. Nos ca-

sos em que as variáveis contínuas (em cada casela) tenham uma distribuição muito assimétrica, provavelmente, a utilização de outras distribuições ou de um método não-paramétrico para a estimação dessas densidades produziria resultados mais eficientes;

- (2) Seleção de variáveis no modelo de Krzanowski. Daudin (1986) apresentou um método baseado no critério de Akaike que, em seu exemplo, não teve um resultado satisfatório. Estudar aplicabilidade ou desenvolver outros métodos;

- (2) Estender o modelo de Krzanowski de forma a tratar, sem perda de informação, as variáveis qualitativas ordinais;

- (3) Demonstrar, para variáveis contínuas ou misturadas, que o estimador de máxima verossimilhança dos parâmetros $\boldsymbol{\alpha}$ em amostragem separada na Discriminação Logística é o mesmo (a menos de uma correção em α_0) que o obtido se supusermos amostragem conjunta. Anderson (1972) demonstrou para o caso em que $\mathbf{X}$ é composto por variáveis discretas;

- (4) A regra de partição para variáveis quantitativas no Método da Árvore Binária de Classificação é obtida escolhendo-se em cada variável a melhor entre $n-1$ partições onde n é o tamanho da amostra. Para reduzir o tempo de processamento, uma das principais desvantagens deste método, poderia-se determinar que a regra de partição seria escolhida entre os decis dessa variável. Estudar a confiabilidade do método proposto;

- (2, 3 e 4) Estender cada um dos métodos para os casos em que há uma estrutura de ordem nas populações. Para a Discriminação Logística, Albert e Lesaffre (1986) citam três referências.

Capítulo 2

O MÉTODO DE KRZANOWSKI

Neste capítulo veremos uma das técnicas mais recentes para a solução do problema de discriminação quando as variáveis são misturadas.

O método de Krzanowski denominado, em inglês, de *location model*, foi introduzido por Olkin e Tate (1961) e utilizado em Análise Discriminante inicialmente por Chang e Afifi (1974). Posteriormente, Krzanowski (1975, 1980, 1982, 1986) escreveu diversos artigos que viabilizaram a sua aplicação (daí denominarmos o método com o seu nome).

Para termos uma idéia preliminar do método, consideremos um exemplo simplificado: Seja $\mathbf{X} = (X_1,\ X_2,\ X_3)'$, onde:

X_1 = variável categórica cujas categorias serão representadas por 0, 1 ou 2;

X_2 = variável binária com seus resultados representados por 0 ou 1;

X_3 = variável contínua.

Podemos representar as variáveis X_1 e X_2 a partir da seguinte tabela composta por seis caselas:

	X_1		
X_2	0	1	2
0	casela 1	casela 2	casela 3
1	casela 4	casela 5	casela 6

Para cada casela contrói-se uma regra discriminante baseada em X_3 e na probabilidade de que uma observação da população Π_i $(i = 1, \ldots, g)$ pertença a esta casela.

Assim, para que uma nova observação seja classificada, devemos determinar a casela correspondente às variáveis categóricas e a seguir, classificá-la em uma das populações utilizando-se da regra específica dessa casela.

A principal desvantagem deste método é que o tamanho mínimo da amostra necessário cresce exponencialmente com o número de caselas. Desse modo, a aplicação do método é operacionalmente viável somente quando as variáveis discriminadoras categóricas formarem poucas caselas ou quando pudermos dispor de grandes amostras.

Na Seção 2.1 formalizaremos o método de Krzanowski e apresentaremos a expressão de sua regra de classificação. A estimação dos parâmetros pelos estimadores usuais, via modelos log-lineares e via regressão múltipla multivariada será assunto da Seção 2.2. Na Seção 2.3 apresentaremos um algoritmo *backward* para seleção de variáveis categóricas.

2.1 O Método de Krzanowski e a Regra de Classificação

Vamos supor, sem perda de generalidade, que das r variáveis discriminadoras, as q primeiras são variáveis categóricas e as p $(p = r - q)$ restantes são variáveis contínuas.

Dessa forma, podemos particionar $\mathbf{X}$ em $\mathbf{Z}$ e $\mathbf{Y}$, onde $\mathbf{Z} = (X_1, \ldots, X_q)'$ e $\mathbf{Y} = (X_{q+1}, \ldots, X_r)'$.

Seja c_j $(j = 1, \ldots, q)$ o número de categorias de Z_j, categorias estas que serão representadas por $0, 1, \ldots, c_j - 1$.

Cada uma das possíveis combinações das categorias de $Z_1, \ldots, Z_q$ de $\mathbf{Z}$ definem de forma única uma casela de uma tabela de contingência q-dimensional. Seja M a variável indicativa desta casela. Isto é, M assume valores $1, 2, \ldots, c$ onde c é o número total de caselas dada por:

$$c = \prod_{j=1}^{q} c_j \qquad (2-1)$$

A função (bijetora) que relaciona um particular valor $\mathbf{z}$ (realização da variável $\mathbf{Z}$) com o valor m (realização da variável M), é :

$$m = 1 + z_1 + \sum_{j=2}^{q} z_j \left(\prod_{v=1}^{j-1} c_v \right) \qquad (2-2)$$

Esta função simplesmente rotula as caselas. Por exemplo: se $\mathbf{Z} = (Z_1,\ Z_2,\ Z_3)'$ com $c_1 = 3$ e $c_2 = c_3 = 2$, então, aplicando $(2-2)$:

$$m = 1 + z_1 + 3\, z_2 + 6\, z_3$$

que gera os seguintes valores para M:

$z_3 = 0$

z_2	z_1 0	1	2
0	1	2	3
1	4	5	6

$z_3 = 1$

z_2	z_1 0	1	2
0	7	8	9
1	10	11	12

O método de Krzanowski supõe que a distribuição de $\mathbf{Y}$ condicionada em $(M = m)$ e na população Π_i é conhecida a menos de alguns parâmetros. Em geral, supõe-se que a distribuição seja normal multivariada com matriz de dispersão constante, ou seja:

$$(\mathbf{Y} \,|\, M = m,\ \Pi_i) \sim \mathrm{N}_p(\boldsymbol{\mu}_i^{(m)}, \boldsymbol{\Sigma}) \qquad (m = 1, \ldots, c;\ i = 1, \ldots, g)$$

Vamos admitir também que $\Pr(M = m \,|\, \Pi_i)$ $(m = 1, \ldots, c;\ i = 1, \ldots, g)$ seja conhecida.

Como $\Pr(\mathbf{x} \,|\, \Pi_i) = \Pr(\mathbf{y} \,|\, \mathbf{z}, \Pi_i)\ \Pr(\mathbf{z} \,|\, \Pi_i)$ e, utilizando $(2-2)$,

$$\begin{aligned} \Pr(\mathbf{z} \,|\, \Pi_i) &= \Pr(m \,|\, \Pi_i) \\ \text{e} \quad \Pr(\mathbf{y} \,|\, \mathbf{z}, \Pi_i) &= \Pr(\mathbf{y} \,|\, m, \Pi_i) \end{aligned}$$

Temos,

$$\Pr(\mathbf{x} \,|\, \Pi_i)\, p_i = \Pr(\mathbf{y} \,|\, m, \Pi_i)\ \Pr(m \,|\, \Pi_i)\, p_i \qquad (m = 1, \ldots, c;\ i = 1, \ldots, g)$$

onde $\Pr(\mathbf{y} \,|\, m, \Pi_i)$ é a função densidade $\mathbf{y}$ em Π_i e $M = m$.

Portanto, a regra ótima de classificação pelo método de Krzanowski torna-se:

Dado $\mathbf{x}' = (\mathbf{z}', \mathbf{y}')$, localizo m e

$$\begin{array}{ll} \text{se} & \Pr(\mathbf{y} \mid m, \Pi_i)\, \Pr(m \mid \Pi_i)\, p_i \;=\; \max\limits_{h=1,\ldots,g} \Pr(\mathbf{y} \mid m, \Pi_h)\, \Pr(m \mid \Pi_h)\, p_h \\ \text{então,} & \mathbf{x} \in R_i \end{array} \qquad (2-3)$$

Krzanowski (1975) restringe o método para o caso em que $\Pr(\mathbf{y} \mid m, \Pi_i)$ é a densidade de uma distribuição normal p-variada com média $\boldsymbol{\mu}_i^{(m)}$ e matriz de variâncias e covariâncias constante $\boldsymbol{\Sigma}$. Referiremo-nos a isto, simplesmente, como hipótese de normalidade e a utilizaremos aqui e nas demais seções deste capítulo.

Sob a hipótese de normalidade, pode-se mostrar (através de simples manipulações algébricas) que a regra $(2-3)$ é equivalente a:

Dado $\mathbf{x}' = (\mathbf{z}', \mathbf{y}')$, localizo m e

$$\begin{array}{ll} \text{se} & d_{im}^2 - 2 \ln [p_i \Pr(m \mid \Pi_i)] \;=\; \min\limits_{h=1,\ldots,g} \{ d_{hm}^2 - 2 \ln [p_h \Pr(m \mid \Pi_h)] \} \\ \text{então,} & \mathbf{x} \in R_i \end{array} \qquad (2-4)$$

onde $d_{im}^2 = (\mathbf{y} - \boldsymbol{\mu}_i^{(m)})\, \boldsymbol{\Sigma}^{-1}\, (\mathbf{y} - \boldsymbol{\mu}_i^{(m)}) \quad (i = 1, \ldots, g)$.

Considere a situação em que as variáveis categóricas são não informativas, isto é, $\Pr(m \mid \Pi_i)$ é igual a uma constante para todo m e para todo i. Neste caso, a regra do método de Krzanowski, numa casela qualquer, é equivalente a regra da LDF calculada exclusivamente com os dados pertencentes a esta casela.

Um caso particular do método de Krzanowski em que se pode observar, com clareza, a superioridade deste em relação à LDF é descrito a seguir.

Chang e Afifi (1974) introduziram a idéia do método de Krzanowski apresentando o caso de duas populações e uma variável binária Z (admitiram, por simplicidade, que $p_1 = p_2$). Obtiveram duas funções discriminantes, utilizadas conforme o valor de

Z. Sob a restrição de normalidade, isto é, $(\mathbf{Y} \mid Z = z,\ \Pi_i) \sim \mathrm{N}_p(\boldsymbol{\mu}_i^{(z)}, \boldsymbol{\Sigma})$, a regra de classificação torna-se:

Dado $\mathbf{x}$, localizo z e

se $\left(\boldsymbol{\mu}_1^{(z)} - \boldsymbol{\mu}_2^{(z)}\right)' \boldsymbol{\Sigma}^{-1} \left[\mathbf{y} - \frac{1}{2}\left(\boldsymbol{\mu}_1^{(z)} + \boldsymbol{\mu}_2^{(z)}\right)\right] > \ln\left[\Pr(Z = z \mid \Pi_2)/\Pr(Z = z \mid \Pi_1)\right]$

então, $\mathbf{x} \in R_1$;

caso contrário, $\mathbf{x} \in R_2$

Observe que mesmo quando a variável Z , isoladamente, não tem nenhum poder discriminatório, isto é, $\Pr(Z \mid \Pi_1) = \Pr(Z \mid \Pi_2)$, a discriminação baseada no método de Krzanowski ainda tende a suplantar a função discriminante linear (LDF) calculada com as mesmas variáveis. Isto ocorre de maneira expressiva quando existe uma forma de "reversão" entre Z e Y. A Figura 2.1 abaixo, ilustra um caso extremo de reversão:

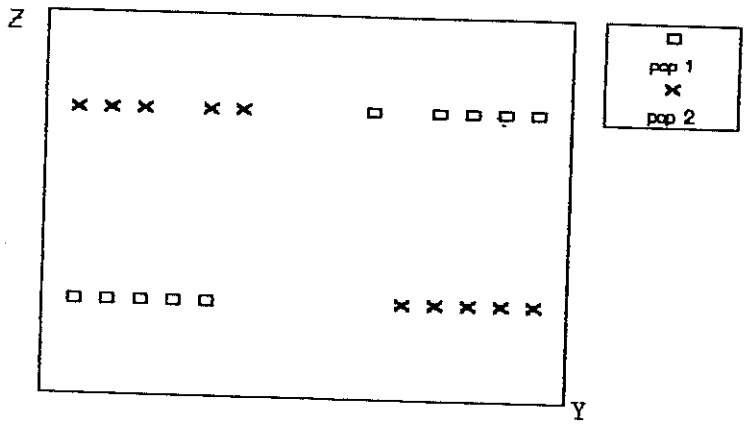


Figura 2.1 - Variáveis discriminadoras: Y (contínua) e Z (binária).

Neste exemplo, a regra de classificação pelo método de Krzanowski terá um desempenho ótimo considerando que se fixarmos $Z = 0$ ou 1 as observações das populações estarão nitidamente separadas. Já a LDF ou a QDF terão desempenhos péssimos, pois os vetores de médias e as matrizes de dispersão estimadas de Π_1 e Π_2 são praticamente coincidentes.

Krzanowski (1977) comparou através de simulações o desempenho do método de Krzanowski com o método da LDF e constatou que na presença de reversão este último método é francamente inferior.

Finalizamos esta seção observando que a hipótese de normalidade do método de Krzanowski pode ser alterada, permitindo que as matrizes de variâncias e covariâncias variem em cada população ou até em cada casela.

No primeiro caso, temos:

$$(\mathrm{Y} \mid M = m, \Pi_i) \sim \mathrm{N}_p(\boldsymbol{\mu}_i^{(m)}, \boldsymbol{\Sigma}_i) \qquad (m = 1, \ldots, c;\ i = 1, \ldots, g)$$

Gerando a seguinte regra de classificação:

Dado $\mathbf{x}$, localizo m e

$$\text{se} \quad \ln|\Sigma_i| + D_{im}^2 - 2\ln[p_i \Pr(m \mid \Pi_i)] = \min_{h=1,\ldots,g} \{\ln|\Sigma_h| + D_{hm}^2 - 2\ln[p_h \Pr(m \mid \Pi_h)]\}$$

$$\text{então,} \quad \mathbf{x} \in R_i$$

onde $D_{im}^2 = (\mathbf{y} - \boldsymbol{\mu}_i^{(m)})\, \boldsymbol{\Sigma}_i^{-1}\, (\mathbf{y} - \boldsymbol{\mu}_i^{(m)}) \qquad (m = 1, \ldots, c;\ i = 1, \ldots, g)$.

Pode-se mostrar que se as variáveis categóricas forem não informativas, esta regra, calculada numa casela m qualquer, é equivalente a regra da QDF calculada exclusivamente com os dados relacionados a esta casela.

No segundo caso, a suposição é:

$$(\mathrm{Y} \mid M = m, \Pi_i) \sim \mathrm{N}_p(\boldsymbol{\mu}_i^{(m)}, \boldsymbol{\Sigma}_i^{(m)}) \qquad (m = 1, \ldots, c;\ i = 1, \ldots, g)$$

A regra de classificação torna-se:

Dado $\mathbf{x}$, localizo m e

$$\text{se} \quad \ln |\Sigma_i^{(m)}| + D_{im}^2 - 2 \ln [p_i \Pr(m \,|\, \Pi_i)] = \min_{h=1,\ldots,g} \left\{ \ln |\Sigma_h^{(m)}| + D_{hm}^2 - 2 \ln [p_h \Pr(m \,|\, \Pi_h)] \right\}$$

$$\text{então,} \quad \mathbf{x} \in R_i$$

onde $D_{im}^2 = (\mathbf{y} - \boldsymbol{\mu}_i^{(m)})\, \Sigma_i^{(m)\,-1} (\mathbf{y} - \boldsymbol{\mu}_i^{(m)}) \qquad (m = 1,\ldots, c;\ i = 1,\ldots, g)$.

É importante considerar que estas extensões criam modelos com muito mais parâmetros e a estimação destes pode, eventualmente, requerer um tamanho de amostra impraticável.

2.1.1 A Distância de Matusita

O objetivo desta seção é apresentar uma medida de distância entre variáveis misturadas que, quando aplicada em Análise Discriminante, produz os mesmos resultados que (2–4) quando as probabilidades *a priori* são não informativas. Por "*priori* não informativa" entenda-se $p_i = 1/g \quad (i = 1,\ldots, g)$. Esta suposição é necessária em todas as proposições desta seção.

No caso clássico em que se supõe que $(\mathbf{X} \,|\, \Pi_i) \sim \mathrm{N}_r(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, a regra de classificação LDF é matematicamente igual a classificar $\mathbf{x}$ na população cuja distância de Mahalanobis à média desta população for mínima (Mardia *et al* (1979) - pág. 303).

Fazendo analogia com este caso, Krzanowski (1986) utilizou a distância de Matusita, definida a seguir, e obteve resultados idênticos a (2–4). Assim, para variáveis misturadas, classificar $\mathbf{x}$ pela regra (2–4) é equivalente a classificar $\mathbf{x}$ na população cuja distância de Matusita for mínima.

Inicialmente, a distância de Matusita foi proposta para ser calculada entre duas populações (Krzanowski (1983-A)). Denotando por $\Pr(\mathbf{x}\,|\,\Pi_i)$ a função densidade de $\mathbf{X}$ na população Π_i ($i = 1,\ldots, g$), temos que a distância de Matusita entre as populações Π_i e Π_j é dada por:

$$\Delta^2(\Pi_i,\Pi_j) = 2\,(1-\rho_{ij})$$

onde $\rho_{ij} = \int_{\Omega} [\,\Pr(\mathbf{x}\,|\,\Pi_i)\,\Pr(\mathbf{x}\,|\,\Pi_j)]^{1/2} d\mathbf{x}$ é a afinidade entre Π_i e Π_j.

No método de Krzanowski ,

$$\Pr(\mathbf{x}\,|\,\Pi_i) = \Pr(\mathbf{z},\mathbf{y}\,|\,\Pi_i) = \Pr(m,\mathbf{y}\,|\,\Pi_i) = \Pr(\mathbf{y}\,|\,m,\Pi_i)\,\Pr(m\,|\,\Pi_i)$$

Logo, denotando por Ω_y o espaço amostral de $\mathbf{y}$ temos,

$$\rho_{ij} = \int_{\Omega_y} \sum_{m=1}^{c} [\,\Pr(\mathbf{y}\,|\,m,\Pi_i)\,\Pr(m\,|\,\Pi_i)\,\Pr(\mathbf{y}\,|\,m,\Pi_j)\,\Pr(m\,|\,\Pi_j)]^{1/2} d\mathbf{y}$$

$$= \sum_{m=1}^{c} [\,\Pr(m\,|\,\Pi_i)\,\Pr(m\,|\,\Pi_j)\,]^{1/2} \int_{\Omega_y} [\,\Pr(\mathbf{y}\,|\,m,\Pi_i)\,\Pr(\mathbf{y}\,|\,m,\Pi_j)\,]^{1/2} d\mathbf{y}$$

Sob a suposição de normalidade pode-se verificar (Matusita (1972)) que:

$$\int_{\Omega_y} [\,\Pr(\mathbf{y}\,|\,m,\Pi_i)\,\Pr(\mathbf{y}\,|\,m,\Pi_j)\,]^{1/2} d\mathbf{y} = \exp\left(-\frac{1}{8}\,d^2\right)$$

onde $d^2 = (\mu_i^{(m)} - \mu_j^{(m)})'\,\Sigma^{-1}\,(\mu_i^{(m)} - \mu_j^{(m)})$.

Logo,

$$\Delta^2(\Pi_i,\Pi_j) = 2\left\{1 - \sum_{m=1}^{c} [\,\Pr(m\,|\,\Pi_i)\,\Pr(m\,|\,\Pi_j)\,]^{1/2}\,\exp\left(-\frac{1}{8}\,d^2\right)\right\} \qquad (2-5)$$

Devido a sua simplicidade e generalidade, esta expressão pode ser aplicada em muitas técnicas multivariadas que necessitam do cálculo de distâncias entre populações representadas por variáveis misturadas.

A adoção dessa expressão em Análise Discriminante foi feita por Krzanowski (1986) que propôs classificar uma observação $\mathbf{x}_0$ (tratada como população degenerada) na população Π_i se a distância de Matusita entre $\mathbf{x}_0$ e Π_i for mínima.

É possível tratar $\mathbf{x}_0' = (\mathbf{z}_0', \mathbf{y}_0')$ como uma população degenerada, Π_j digamos, fazendo-se com que o valor da casela (m_0) relacionada com a parte qualitativa $\mathbf{z}_0$ tenha probabilidade 1. Isto é,

$$\Pr(\, m \,|\, \Pi_j \,) = \begin{cases} 1, & \text{se } m = m_0; \\ 0, & \text{c. c.} \end{cases}$$

E a parte contínua $\mathbf{Y}$ tenha massa de probabilidade unitária no valor observado $\mathbf{y}_0$ e probabilidade zero em qualquer outro lugar.

Conseqüentemente, a afinidade entre $\mathbf{x}_0$ e Π_i que denotaremos por ρ_i $(i = 1, \ldots, g)$, é dada por:

$$\begin{aligned} \rho_i &= [\Pr(\, m_0 \,|\, \Pi_i \,)]^{1/2} \; [\Pr(\, \mathbf{y}_0 \,|\, m_0, \Pi_i \,)]^{1/2} \\ &= \Pr(\, m_0 \,|\, \Pi_i \,)^{1/2} \; |2\,\Pi\,\Sigma|^{-1/4} \; \exp\left\{ -\tfrac{1}{4} \left(\mathbf{y}_0 - \boldsymbol{\mu}_i^{(m)} \right)' \; \boldsymbol{\Sigma}^{-1} \; \left(\mathbf{y}_0 - \boldsymbol{\mu}_i^{(m)} \right) \right\} \end{aligned}$$

Como minimizar $\Delta^2(\mathbf{x}_0, \Pi_i) = 2(1 - \rho_i)$ é equivalente a maximizar ρ_i, a regra de classificação torna-se, dado $\mathbf{x}_0$

$$\text{se} \quad \rho_i = \max_{h=1,\ldots,\, g} \rho_h \qquad \text{então,} \quad \mathbf{x}_0 \in R_i \tag{2-6}$$

O fato notável é que a regra acima é exatamente a mesma que a regra (2–4) quando as probabilidades *a priori* p_i $(i = 1, \ldots, g)$ são não informativas. Entre-

tanto, a regra de classificação baseado na distância de Matusita tem vantagens descritivas. Por exemplo, pode-se verificar se uma observação $\mathbf{x}$ é um *outlier* analisando-se a magnitude das distâncias $\Delta^2(\mathbf{x}, \Pi_i)$ $(i = 1, \ldots, g)$.

2.2 Estimação dos Parâmetros

2.2.1 Estimadores Usuais

Em geral, a regra (2–4) não é aplicável uma vez que os parâmetros p_i, $\Pr(m \mid \Pi_i)$, $\mu_i^{(m)}$ e Σ $(\Sigma_i, \Sigma_i^{(m)})$ $(m = 1, \ldots, c;\ i = 1, \ldots, g)$ são desconhecidos. A informação disponível, geralmente, consiste em amostras de tamanho n_i de cada população Π_i $(i = 1, \ldots, g)$.

A estimação das probabilidades *a priori* p_i $(i = 1, \ldots, g)$ foi discutida na Seção 1.1. As estimativas dos demais parâmetros independem do planejamento amostral e podem ser feitas pelos estimadores usuais:

$$\widehat{\Pr}(m \mid \Pi_i) = n_{im}/n_i$$

$$\hat{\mu}_i^{(m)} = \bar{\mathbf{y}}_i^{(m)} = \frac{1}{n_{im}} \sum_{k \in E_{im}} \mathbf{y}_k$$

$$\hat{\Sigma}_i^{(m)} = \frac{1}{n_{im} - 1} \sum_{k \in E_{im}} (\mathbf{y}_k - \bar{\mathbf{y}}_i^{(m)})\,(\mathbf{y}_k - \bar{\mathbf{y}}_i^{(m)})'$$

$$\hat{\Sigma}_i = \frac{1}{n_i - c} \sum_{m=1}^{c} (n_{im} - 1)\, \hat{\Sigma}_i^{(m)}$$

$$\hat{\Sigma} = \frac{1}{n - g\,c} \sum_{i=1}^{g} (n_i - c)\, \hat{\Sigma}_i \qquad (2-7)$$

onde n_{im} é o número de elementos da amostra pertencentes à população Π_i e à casela m, $\mathbf{y}_k$ é o vetor de variáveis contínuas associado à k-ésima observação e E_{im} é o conjunto dos índices das observações que pertencem à casela m e à população Π_i.

Recomenda-se que estes estimadores sejam utilizados somente quando amostras correspondentes a cada casela e a cada população estiverem disponíveis em tamanho suficiente para uma estimação segura.

A regra de classificação amostral é obtida substituindo-se os parâmetros pelas suas respectivas estimativas.

2.2.2 Estimação por Modelos

Quando a amostra não é muito grande em relação ao número de caselas c, é bastante provável que algumas caselas tenham poucas ou, eventualmente, nenhuma observação. Logo, as estimativas serão pouco precisas ou até impossíveis de serem calculadas. Neste caso, para que a análise via método de Krzanowski sob a hipótese de normalidade seja feita, torna-se necessário assumirmos a existência de estruturas adicionais entre as distribuições das variáveis.

Consideremos inicialmente o problema de se estimar através de modelos as probabilidades $\Pr(m \mid \Pi_i)$ $(m = 1, \ldots, c;\ i = 1, \ldots, g)$. Como já foi descrito, podemos representar as q variáveis qualitativas na forma de tabelas de contingência, uma para cada população.

Uma forma utilizada para se obter estimativas de $\Pr(m \mid \Pi_i)$ é supor que o logaritmo desta probabilidade seja linear nos parâmetros determinados a partir das variáveis categóricas. Desse modo é possível obter estimativas para $\Pr(m \mid \Pi_i)$ mesmo quando a

freqüência observada n_{im} for nula. Evidentemente, a qualidade dessas estimativas está diretamente ligada à qualidade do ajuste do modelo log-linear considerado.

A estimação por modelos log-lineares é um assunto extenso e foge dos objetivos deste trabalho. Referências básicas sobre este assunto são Knoke e Burke (1980) e Bishop, Fienberg e Holland (1975). Recentemente, Reis (1989) apresentou tese de mestrado sobre este tema.

Um resultado empírico que pode ser utilizado no ajuste dos modelos log-lineares é reter no modelo apenas os efeitos principais e as interações de primeira ordem. Krzanowski (1975) argumenta que tal procedimento tem dado resultados satisfatórios.

Consideremos agora o problema da estimação de $\boldsymbol{\mu}_i^{(m)}$ e $\boldsymbol{\Sigma}$ $(\boldsymbol{\Sigma}_i)$. Krzanowski (1975) sugere fazermos isto através de regressão múltipla multivariada. As variáveis independentes deste modelo são as variáveis categorizadas $Z_1, \ldots, Z_q$ assumindo $c_1, \ldots, c_q$ "valores" diferentes, respectivamente; e o vetor de variáveis dependentes é formado pelas variáveis contínuas. A idéia é supor que a média das variáveis contínuas de cada casela e de cada população possa ser escrita como uma combinação linear dos valores das variáveis categóricas.

Contudo, a aplicação direta do modelo de regressão em variáveis categóricas não é recomendável, uma vez que se fizéssemos isso estaríamos introduzindo uma métrica que poderia não ser adequada.

O artifício utilizado consiste em reparametrizar cada variável categórica com c_j $(c_j > 2)$ categorias de resposta em $c_j - 1$ variáveis binárias.

Assim, a notação "$\boldsymbol{\alpha}_j z_j$" no modelo que iremos propor significa soma dos produtos dos coeficientes com as variáveis binárias. Por exemplo, se a variável Z_j apresenta quatro categorias $(c_j = 4)$ então:

$$\boldsymbol{\alpha}_j\, z_j = \boldsymbol{\alpha}_{j1}\, d_{j1} + \boldsymbol{\alpha}_{j2}\, d_{j2} + \boldsymbol{\alpha}_{j3}\, d_{j3}$$

onde $\boldsymbol{\alpha}_j$ é um vetor de parâmetros de dimensão p e as variáveis binárias d_{j1}, d_{j2} e d_{j3} assumem valores conforme a tabela abaixo:

z_j	d_{j1}	d_{j2}	d_{j3}
0	1	0	0
1	0	1	0
2	0	0	1
3	0	0	0

No modelo de regressão poderão ser incorporadas novas variáveis independentes resultantes de produtos de variáveis a fim de possibilitar a inclusão de interações. Assim, o modelo para a média da casela m da população Π_i pode ser escrito na forma geral como:

$$\begin{aligned}\mu_i^{(m)} &= \boldsymbol{\alpha}_i + \sum_{j=1}^{q} \boldsymbol{\alpha}_{ij}\, z_j^{(m)} + \sum_{j=1}^{q} \sum_{v=j+1}^{q} \boldsymbol{\alpha}_{ijv}\, z_j^{(m)}\, z_v^{(m)} + \\ &+ \cdots + \boldsymbol{\alpha}_{i12\ldots q}\, z_1^{(m)}\, z_2^{(m)} \ldots z_q^{(m)} \qquad (m = 1, \ldots, c;\ i = 1, \ldots, g)\end{aligned}$$

onde $\boldsymbol{\alpha}_i$, $\boldsymbol{\alpha}_{ij}$, $\ldots$, $\boldsymbol{\alpha}_{i12\ldots q}$ são os coeficientes multivariados do modelo e $z_j^{(m)}$ é o valor da j-ésima variável categórica na casela m.

O lado direito da expressão acima pode ter um número menor de termos, gerando, por exemplo, modelos de primeira ordem, segunda ordem, etc.

A estimação dos coeficientes e, conseqüentemente de $\boldsymbol{\mu}_i^{(m)}$, pode ser obtida através de regressão múltipla multivariada. Detalhes computacionais podem ser vistos em Mardia *et al* (1979) - Capítulo 6. Após o ajuste do modelo, feito separadamente para cada população, resulta uma matriz de resíduos que permite obter uma estimativa para $\boldsymbol{\Sigma}_i$.

A estimativa de Σ, necessária para atender a hipótese de normalidade é feita da forma usual (2–7).

2.3 Seleção de Variáveis Categóricas

A grande desvantagem do método de Krzanowski em relação a outros métodos discriminantes é o excessivo número de parâmetros que necessitam ser estimados. Sob a hipótese de normalidade e considerando q variáveis binárias e p contínuas, o número de parâmetros (v) que necessitam ser estimados é:

$$\begin{array}{rcccccc} v & = & 2^q\, g\, p & + & p^2 - \binom{p}{2} & + & (2^q - 1)\, g \\ & & \text{(médias)} & + & \begin{pmatrix}\text{matrizes de}\\ \text{dispersão}\end{pmatrix} & + & \begin{pmatrix}\text{“probabilidades” de}\\ \text{cada casela}\end{pmatrix} \end{array}$$

Observando a expressão acima, verifica-se que o número de variáveis binárias (q) exerce uma influência muito grande na magnitude de v.

Não raramente, o número de variáveis contínuas é de tal ordem que impossibilita a aplicação do método de Krzanowski . Mesmo quando grandes amostras (em relação a c) são disponíveis, a excessiva parametrização pode influir negativamente no desempenho do modelo.

Assim é de grande interesse prático um procedimento que selecione as variáveis discriminadoras categóricas mais importantes. Krzanowski (1983-B) sugere um procedimento simples para o caso de duas populações, baseado na distância de Matusita (2–5) que ora reproduzimos:

$$\begin{aligned} \Delta^2(\Pi_1, \Pi_2) \;=\; & 2\Big\{1 - \sum_{m=1}^{c} [\Pr(\,m\,|\,\Pi_1\,)\ \Pr(\,m\,|\,\Pi_2\,)]^2 \\ & \exp\Big[-\tfrac{1}{8}\,(\mu_1^{(m)} - \mu_2^{(m)})'\ \Sigma^{-1}\ (\mu_1^{(m)} - \mu_2^{(m)})\Big]\Big\} \end{aligned}$$

A idéia do procedimento consiste em eliminar as variáveis categóricas que não provoquem um decréscimo expressivo na distância original entre as duas populações.

O ideal seria calcular a distância $\Delta^2(\Pi_1, \Pi_2)$ para todas as possíveis combinações de variáveis categóricas. Tal procedimento, entretanto, oneraria em muito o tempo de processamento. Dessa forma o seguinte procedimento *backward* foi sugerido:

(i) Seja $A = \{Z_1, \ldots, Z_q\}$. Calcule $\Delta^2(\Pi_1, \Pi_2)$;

(ii) Para cada Z_j $(Z_j \in A)$, calcule $\Delta^2(\Pi_1, \Pi_2)$ sem esta variável;

(iii) Retire a variável do conjunto A cuja exclusão no passo anterior tiver resultado no máximo valor de $\Delta^2(\Pi_1, \Pi_2)$ ou, equivalentemente, a que tiver gerado a menor redução na distância;

(iv) Repita os passos (ii) e (iii) até que $A = \emptyset$, ocasião em que $\Delta^2(\Pi_1, \Pi_2)$ será calculada conforme a distância de Matusita para variáveis contínuas:

$$\Delta^2(\Pi_1, \Pi_2) = 2\left\{1 - \exp\left[-\frac{1}{8}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2)\right]\right\}$$

(v) A escolha do melhor conjunto de variáveis é feita descritivamente.

No exemplo considerado por Krzanowski (1983-B), um problema com seis variáveis categóricas e sete variáveis contínuas foi submetido ao procedimento acima. Apresentamos um resumo dos resultados na Tabela 2.1.

Não existe um critério formal para determinar quando a redução na distância é expressiva. Descritivamente, observa-se na Tabela 2.1 que a exclusão das variáveis Z_4 e Z_5 provocaram uma redução muito pequena na distância entre as populações. Se, além destas, excluirmos a variável Z_3, a redução já não será tão inexpressiva. Portanto, o "melhor" subconjunto de variáveis discriminadoras qualitativas é $\{Z_1,\ Z_2,\ Z_3,\ Z_6\}$.

A	$\Delta^2(\Pi_1, \Pi_2)$	variável excluída	redução na distância
Z_1, Z_2, Z_3, Z_4, Z_5, Z_6	1,305	Z_4	0,019
Z_1, Z_2, Z_3, Z_5, Z_6	1,286	Z_5	0,054
Z_1, Z_2, Z_3, Z_6	1,232	Z_3	0,164
Z_1, Z_2, Z_6	1,068	Z_1	0,205
Z_2, Z_6	0,863	Z_6	0,225
Z_2	0,638	Z_2	0,423

Tabela 2.1 - Resultados extraídos de Krzanowski (1983-B).

É importante frisar que este procedimento seleciona um particular subconjunto de variáveis categóricas que tende a ser ótimo no sentido de preservar a distância entre as populações devida a estas variáveis e não ótimo para minimizar a probabilidade de classificação incorreta de futuras observações, que é o nosso objetivo.

Uma crítica que poderia ser feita ao procedimento refere-se ao mesmo tratamento que é dado às variáveis qualitativas com diferentes números de categorias possíveis. Como o objetivo básico é diminuir o número de parâmetros a serem estimados, não é lógico excluir uma variável (por exemplo binária) se a redução na distância que ela provocar for pouco inferior à causada por outra variável com, digamos, 5 categorias. A redução no número de parâmetros a serem estimados excluindo-se a variável com 5 categorias "compensa" o maior decréscimo provocado na distância entre as populações. O mais adequado seria considerar como critério de exclusão de variáveis a minimização de uma função que relacionasse redução na distância e número de parâmetros remanescentes.

Finalizamos esta seção observando que se a magnitude de $\Delta^2(\Pi_1, \Pi_2)$ com $A = \emptyset$ for comparável a $\Delta^2(\Pi_1, \Pi_2)$ com A contendo todas as variáveis categóricas, então es-

tas variáveis não têm aparentemente poder discriminatório. Mas, mesmo neste caso, o método LDF calculado com as variáveis quantitativas poderá não ter um desempenho tão bom quanto o do método de Krzanowski com todas as variáveis, em virtude da possibilidade de haver reversão (Figura 2.1).

Capítulo 3

DISCRIMINAÇÃO LOGÍSTICA

Desde que foi proposto, separadamente, por Cox (1966) e por Day e Kerridge (1967), o modelo logístico tem sido estudado com bastante freqüência nas áreas de regressão e discriminação.

A utilização deste modelo em Análise Discriminante (que denominaremos de Discriminação Logística) tem vários atrativos, dentre os quais destacam-se:

(i) as suposições sobre as distribuições das variáveis discriminadoras são bastante gerais;

(ii) pode ser aplicado em variáveis binárias, contínuas ou numa mistura de binárias e contínuas;

(iii) disponibilidade de *software* (por exemplo: BMDP, SAS);

(iv) uma vez estimados os parâmetros, a sua utilização é bastante simples.

As distribuições a que se refere o item (i) são das variáveis discriminadoras condicionadas na população Π_i $(i = 1, \ldots, g)$. Veremos na Seção 3.1 qual é a suposição necessária sobre estas distribuições. Por ora, adiantamos que as distribuições mais importantes que satisfazem ao modelo logístico são as seguintes:

(i) distribuição normal multivariada com mesma matriz de dispersão em cada população;

(ii) distribuições multivariadas binárias seguindo um modelo log-linear com as mesmas interações em cada população;

(iii) uma mistura das situações (i) e (ii), não necessariamente independentes.

Entretanto, não é necessário que as variáveis discriminadoras sejam de uma das distribuições acima para que o método tenha um bom desempenho. Day e Kerridge (1967) afirmam, sem demonstrar, que a Discriminação Logística terá um bom desempenho sempre que uma regra discriminante linear for adequada para o problema, independentemente da distribuição das variáveis discriminadoras.

3.1 - Formulação do Problema

Como nos capítulos anteriores, consideremos o problema de classificar uma observação $\mathbf{x}$, da variável aleatória $\mathbf{X}$ em uma das g populações Π_i $(i = 1, \ldots, g)$. As variáveis que compõem $\mathbf{X}$ poderão ser binárias ou contínuas. Existe a possibilidade de se incorporar outros tipos de variáveis (ver Seção 1.4).

Neste capítulo, $\mathbf{X}$ é um vetor aleatório de dimensão $r + 1$ com $X_0 = 1$, isto é , $\mathbf{X} = (1, X_1, \ldots, X_r)'$ com densidade denotada por $\Pr(\mathbf{x} \mid \Pi_i)$.

A suposição fundamental da Discriminação Logística é que o logaritmo da razão das densidades seja linear nos parâmetros. Isto é, para todo $i = 1, \ldots, g-1$, temos:

$$\ln\left[\frac{\Pr(\mathbf{x}\,|\,\Pi_i)}{\Pr(\mathbf{x}\,|\,\Pi_g)}\right] = \boldsymbol{\beta}_i'\,\mathbf{x} \qquad (3-1)$$

onde $\Pr(\mathbf{x}\,|\,\Pi_g)$ funciona como referência e $\boldsymbol{\beta}$ é o vetor de parâmetros de dimensão $r+1$.

A partir de (3–1) podemos determinar uma expressão para a probabilidade *a posteriori* $\Pr(\Pi_i\,|\,\mathbf{x})$. De (3–1) temos,

$$\Pr(\mathbf{x}\,|\,\Pi_i) = \exp(\boldsymbol{\beta}_i'\,\mathbf{x})\Pr(\mathbf{x}\,|\,\Pi_g) \qquad (i = 1, \ldots, g-1)$$

e utilizando o Teorema de Bayes,

$$\Pr(\mathbf{x}\,|\,\Pi_i) = \frac{\Pr(\Pi_i\,|\,\mathbf{x})}{p_i}\Pr(\mathbf{x}) \qquad (i = 1, \ldots, g-1)$$

onde p_i é a probabilidade *a priori* de $\mathbf{X}$ pertencer à população Π_i.

Logo,

$$\frac{\Pr(\Pi_i\,|\,\mathbf{x})}{p_i} = \exp\left(\boldsymbol{\beta}_i'\,\mathbf{x}\right)\frac{\Pr(\Pi_g\,|\,\mathbf{x})}{p_g} \qquad (i = 1, \ldots, g-1) \qquad (3-2)$$

somando-se ambos os lados para i de 1 até $g-1$, temos

$$\sum_{i=1}^{g-1}\Pr(\Pi_i\,|\,\mathbf{x}) = \sum_{i=1}^{g-1}\exp\left[\ln\left(\frac{p_i}{p_g}\right) + \boldsymbol{\beta}_i'\,\mathbf{x}\right]\Pr(\Pi_g\,|\,\mathbf{x}) \qquad \Rightarrow$$

$$\Rightarrow \qquad 1 - \Pr(\Pi_g\,|\,\mathbf{x}) = \Pr(\Pi_g\,|\,\mathbf{x})\sum_{i=1}^{g-1}\exp\left[\ln\left(\frac{p_i}{p_g}\right) + \boldsymbol{\beta}_i'\,\mathbf{x}\right] \qquad \Rightarrow$$

$$\Rightarrow \qquad \Pr(\Pi_g\,|\,\mathbf{x}) = \frac{1}{1 + \sum_{i=1}^{g-1}\exp\left[\ln\left(p_i/p_g\right) + \boldsymbol{\beta}_i'\,\mathbf{x}\right]}$$

que substituído em (3–2) nos dá a expressão para $\Pr(\Pi_i \,|\, \mathbf{x})$:

$$\Pr(\Pi_i \,|\, \mathbf{x}) = \frac{\exp\left[\ln(p_i/p_g) + \boldsymbol{\beta}_i' \mathbf{x}\right]}{1 + \sum_{h=1}^{g-1} \exp\left[\ln(p_h/p_g) + \boldsymbol{\beta}_h' \mathbf{x}\right]} \qquad (i = 1, \ldots, g-1)$$

Uma notação mais elegante e funcional é obtida definindo-se $\boldsymbol{\alpha}_i' = (\alpha_{i0}\ \alpha_{i1}, \ldots, \alpha_{ir})$ $(i = 1, \ldots, g)$ da seguinte forma:

$$\begin{array}{lcll} \alpha_{i0} & = & \beta_{i0} + \ln(p_i/p_g) & (i = 1, \ldots, g-1) \\ \alpha_{ij} & = & \beta_{ij} & (i = 1, \ldots, g-1;\ j = 1, \ldots, r) \\ \boldsymbol{\alpha}_g & = & \mathbf{0} & \end{array} \qquad (3-3)$$

Logo,

$$\Pr(\Pi_i \,|\, \mathbf{x}) = \frac{\exp(\boldsymbol{\alpha}_i' \mathbf{x})}{\sum_{h=1}^{g} \exp(\boldsymbol{\alpha}_h' \mathbf{x})} \qquad (i = 1, \ldots, g) \qquad (3-4)$$

Para classificar uma nova observação utilizamos a regra que minimiza a probabilidade de classificação incorreta (equivalente à regra bayesiana) vista na Seção 1.2.1. Esta regra é equivalente a classificar $\mathbf{x}$ na população cuja probabilidade *a posteriori* for maior. Isto é,

$$\text{Dado} \quad \mathbf{x}, \qquad \text{se} \quad \Pr(\Pi_i \,|\, \mathbf{x}) = \max_{h = 1, \ldots, g} \Pr(\Pi_h \,|\, \mathbf{x}) \qquad \text{então,} \quad \mathbf{x} \in R_i$$

onde R_i é tal que se $\mathbf{x} \in R_i$ então $\mathbf{x}$ é classificada em Π_i.

A partir de $(3-4)$, podemos reescrever esta regra de classificação:

$$\text{Dado} \quad \mathbf{x}, \qquad \text{se} \quad \frac{\exp(\boldsymbol{\alpha}_i' \mathbf{x})}{\sum_{t=1}^{g} \exp(\boldsymbol{\alpha}_t' \mathbf{x})} = \max_{h = 1, \ldots, g} \frac{\exp(\boldsymbol{\alpha}_h' \mathbf{x})}{\sum_{t=1}^{g} \exp(\boldsymbol{\alpha}_t' \mathbf{x})} \qquad \text{então,} \quad \mathbf{x} \in R_i$$

ou, simplesmente, classificamos $\mathbf{x}$ em Π_i se

$$\boldsymbol{\alpha}_i' \mathbf{x} \geq \boldsymbol{\alpha}_h' \mathbf{x} \qquad (h = 1, \ldots, g) \tag{3-5}$$

Portanto, a regra de classificação da Discriminação Logística é linear em $\mathbf{x}$.

Na prática, os parâmetros são desconhecidos. É preciso estimá-los a partir de amostras classificadas corretamente. Veremos na Seção 3.3.1 que, usualmente, métodos numéricos serão necessários para isso.

No caso geral, $\boldsymbol{\alpha}$ é um vetor de dimensão $(g-1)(r+1)$. Logo, o número de parâmetros a serem estimados é $v = (g-1)(r+1)$, um número relativamente pequeno quando comparado com as regras paramétricas. Estamos admitindo também que $n > v$ para que seja possível a estimação de $\boldsymbol{\alpha}$.

Para algumas distribuições de $(\mathbf{X} \mid \Pi_i)$, a suposição $(3-1)$ não é satisfeita. Muitas vezes a linearidade só é obtida para um conjunto de funções de $\mathbf{x}$ e não para $\mathbf{x}$, ou seja,

$$\ln \left[\frac{\Pr(\mathbf{x} \mid \Pi_i)}{\Pr(\mathbf{x} \mid \Pi_g)} \right] = \boldsymbol{\gamma}_i \, \mathbf{t}(\mathbf{x}) \qquad (i = 1, \ldots, g-1)$$

onde $\boldsymbol{\gamma}_i$ é um vetor de parâmetros e $\mathbf{t}(\mathbf{x})$ é um vetor de funções. A dimensão de $\mathbf{t}(\mathbf{x})$ é determinada pelo número de funções necessárias para que a linearidade ocorra.

Um exemplo clássico deste problema é o caso univariado com duas populações e $(X \mid \Pi_i) \sim \mathrm{N}(\mu_i, \sigma_i^2)$.

Supondo $\sigma_1 \neq \sigma_2$ temos

$$\ln \left[\frac{\Pr(x \mid \Pi_1)}{\Pr(x \mid \Pi_2)} \right] = \gamma_0 + \gamma_1 x + \gamma_2 x^2$$

Assim, a suposição $(3-1)$ é satisfeita se introduzirmos uma nova variável $x_2 = x^2$.

Um outro exemplo é o caso em que X pertence a um intervalo fechado, digamos, $[a, b]$. A variável $Z = (X - a)/(b - a)$ varia entre 0 e 1. Supondo uma distribuição Beta $(\alpha_i,\ \beta_i)$ para $(Z \mid \Pi_i)$ $(i = 1,\ 2)$, temos

$$\ln\left[\frac{\Pr(x \mid \Pi_1)}{\Pr(x \mid \Pi_2)}\right] = \gamma_0 + \gamma_1 \ln z + \gamma_2 \ln(1 - z)$$

$$= \gamma_0 + \gamma_1 \ln\left(\frac{x - a}{b - a}\right) + \gamma_2 \ln\left(\frac{b - x}{b - a}\right)$$

Se ao invés de x utilizarmos x_1 e x_2 dados por

$$x_1 = \ln\left(\frac{x - a}{b - a}\right) \qquad \text{e} \qquad x_2 = \ln\left(\frac{b - x}{b - a}\right)$$

então o logaritmo da razão das densidades será linear em x_1 e x_2 satisfazendo, dessa forma, a suposição do modelo logístico.

Kay e Little (1987) estudaram o problema da determinação de $\mathbf{t}(\mathbf{x})$ a partir das distribuições de $(X \mid \Pi_i)$ supostas da família exponencial. Concluíram que soluções explícitas para $\mathbf{t}(\mathbf{x})$ só são possíveis nos casos simples, como por exemplo, quando $(X_1, \ldots, X_r)$ são independentes entre si ou em alguns casos com $r \leq 3$ envolvendo dependência.

3.2 - Funções de Verossimilhança

3.2.1 - Amostras Conjuntas

Quando a amostragem é conjunta, o logaritmo da verossimilhança da amostra é dado por

$$\ln \mathrm{L}_C(\boldsymbol{\alpha}) = \sum_{i=1}^{g} \sum_{k \in E_i} \ln \Pr(\Pi_i, \mathbf{x}_k)$$

que pode ser escrito como

$$\ln \mathrm{L}_C(\boldsymbol{\alpha}) = \sum_{i=1}^{g} \sum_{k \in E_i} \ln \Pr(\Pi_i \,|\, \mathbf{x}_k) + \sum_{i=1}^{g} \sum_{k \in E_i} \ln \Pr(\mathbf{x}_k)$$

Conforme argumenta Anderson (1982), a única restrição a ser satisfeita por $\Pr(\mathbf{x})$ é que ela seja uma densidade. Esta restrição independe dos valores dos parâmetros $\boldsymbol{\alpha}$. Logo, o estimador de máxima verossimilhança de $\boldsymbol{\alpha}$ é obtido maximizando-se apenas a primeira parcela da soma acima.

Dessa forma, o estimador de máxima verossimilhança de $\boldsymbol{\alpha}$, $\hat{\boldsymbol{\alpha}}$ é obtido maximizando-se:

$$\mathrm{L}(\mathbf{D}, \boldsymbol{\alpha}) = \sum_{i=1}^{g} \sum_{k \in E_i} \ln \Pr(\Pi_i \,|\, \mathbf{x}_k)$$

onde $\mathbf{D}$ é a matriz de dados de dimensão $n \times r$.

Podemos escrever esta expressão usando (3–4):

$$\Pr(\Pi_i \,|\, \mathbf{x}_k) = \frac{\exp(\boldsymbol{\alpha}_i' \mathbf{x}_k)}{\sum_{h=1}^{g} \exp(\boldsymbol{\alpha}_h' \mathbf{x}_k)} = \frac{1}{\sum_{h=1}^{g} \exp\left[(\boldsymbol{\alpha}_h - \boldsymbol{\alpha}_i)' \mathbf{x}_k\right]}$$

logo,

$$\mathrm{L}(\mathbf{D}, \boldsymbol{\alpha}) = \sum_{i=1}^{g} \sum_{k \in E_i} \ln \left\{ \frac{1}{\sum_{h=1}^{g} \exp\left[(\boldsymbol{\alpha}_h - \boldsymbol{\alpha}_i)' \mathbf{x}_k\right]} \right\} \qquad (3-6)$$

Neste tipo de amostragem n é fixo e os n_i $(i = 1, \ldots, g)$ são aleatórios, o que possibilita a estimativa das probabilidades *a priori* p_i : $\hat{p}_i = n_i/n$.

Além disso, uma vez obtido $\hat{\boldsymbol{\alpha}}$ (caso exista) podemos estimar $\boldsymbol{\beta}_i' = (\beta_{i0}, \beta_{i1}, \ldots, \beta_{ir})$ a partir de (3–3):

$$\hat{\beta}_{i0} = \hat{\alpha}_{i0} - \ln\left(\frac{\hat{p}_i}{\hat{p}_g}\right) = \hat{\alpha}_{i0} - \ln\left(\frac{n_i}{n_g}\right) \qquad (i = 1, \ldots, g)$$

$$\hat{\beta}_{ij} = \hat{\alpha}_{ij} \qquad (i = 1, \ldots, g-1;\ j = 1, \ldots, r)$$

3.2.2 - Amostras Separadas

Amostras de tamanhos fixos n_i são retiradas da distribuição condicional $(\mathbf{X} \mid \Pi_i)$ para $i = 1, \ldots, g$. Decorre que as estimativas de p_i não podem ser estimados a partir da amostra. Vamos assumir que $p_1, \ldots, p_g$, com $\sum_{i=1}^{g} p_i = 1$, são conhecidas ou foram estimadas anteriormente.

O logaritmo da função de verossimilhança amostral é dada por

$$\ln \mathrm{L}_S(\boldsymbol{\alpha}) = \sum_{i=1}^{g} \sum_{k \in E_i} \ln \Pr(\mathbf{x}_k \mid \Pi_i) \qquad (3-7)$$

A estimação de $\boldsymbol{\alpha}$ na expressão acima deve preservar o fato de que $\Pr(\mathbf{x}_k \mid \Pi_i)$ é uma densidade. Para isto são necessárias algumas restrições que relacionam os parâmetros $\boldsymbol{\alpha}$ com a verossimilhança marginal de $\mathbf{X}$ (Albert e Lesaffre(1986)). O problema se resume a maximizar $(3-7)$ sujeito a tais restrições.

A estimação de $\boldsymbol{\alpha}$, neste caso, não é trivial. Anderson (1972), em artigo que serve como referência básica sobre o tema, estudou o caso em que $\mathbf{X}$ é composto somente por variáveis discretas e mostrou que o estimador de máxima verossimilhança de

$\boldsymbol{\alpha} : \hat{\boldsymbol{\alpha}}^*$, é obtido maximizando-se a mesma função (3–6) deduzida no caso de amostras conjuntas, a menos da seguinte correção no primeiro elemento de $\hat{\boldsymbol{\alpha}}_i$:

$$\hat{\alpha}^*_{i0} = \hat{\alpha}_{i0} + \ln\left(\frac{p_i\, n_g}{p_g\, n_i}\right) \qquad (i = 1, \ldots, g-1) \qquad (3-8)$$

A demonstração dada por Anderson (1972) foi criticada por Farewell (1979) e por Prentice e Pyke (1979) pela falta de simplicidade. Ao leitor interessado, sugerimos consultar o trabalho de Anderson e Blair (1982), onde consta uma demontração elegante para o caso em que $\mathbf{X}$ é composto por variáveis binárias.

Não encontramos na literatura prova formal para o caso em que apenas variáveis contínuas ou variáveis misturadas são consideradas. Em artigo relativamente recente, Albert e Lesaffre (1986) consideravam este problema ainda em aberto.

Entretanto, o resultado (3–8) é considerado aproximadamente válido por muitos autores (Prentice e Pyke (1979), Anderson (1982), Anderson e Blair (1982)). Uma justificativa intuitiva de que o resultado é, pelo menos, aproximadamente válido é a possibilidade de se "categorizar" as variáveis contínuas.

Em resumo, qualquer que seja a natureza das variáveis (binárias, contínuas ou uma mistura de ambos) e o tipo de amostragem efetuado (conjunta ou separada), temos que maximizar (3–6):

$$\mathrm{L}(\mathbf{D}, \boldsymbol{\alpha}) = \sum_{i=1}^{g} \sum_{k \in E_i} \ln \left[\frac{1}{\sum_{h=1}^{g} \exp\left(\boldsymbol{\alpha}_h - \boldsymbol{\alpha}_i\right)' \mathbf{x}_k} \right]$$

Os valores que maximizam a função acima são os estimadores de máxima verossimilhança de $\boldsymbol{\alpha}$ quando as amostras são conjuntas (independentemente do tipo das variáveis consideradas).

Quando a amostragem é feita separadamente e as variáveis são discretas, $\hat{\boldsymbol{\alpha}}^*$ é o estimador de máxima verossimilhança de $\boldsymbol{\alpha}$ onde $\hat{\boldsymbol{\alpha}}^*$ é dado por

$$\hat{\alpha}^*_{i0} = \hat{\alpha}_{i0} + \ln\left(\frac{p_i\, n_g}{p_g\, n_i}\right) \qquad (i = 1,\ldots, g-1)$$

$$\hat{\alpha}^*_{ij} = \hat{\alpha}_{ij} \qquad (i = 1,\ldots, g-1;\, j = 1,\ldots, p)$$

com $\quad \hat{\boldsymbol{\alpha}}_i'^* = (\hat{\alpha}^*_{i0},\ldots, \hat{\alpha}^*_{ip}) \quad$ e $\quad \hat{\boldsymbol{\alpha}}_i' = (\hat{\alpha}_{i0},\ldots, \hat{\alpha}_{ip}) \qquad (i = 1,\ldots, g-1).$

Se amostras separadas são utilizadas e se há pelo menos uma variável contínua no modelo, o resultado acima é tido como aproximado, aproximação esta considerada satisfatória.

3.3 - Estimação dos Parâmetros

3.3.1 - Máxima Verossimilhança

Vimos que para a estimação de $\boldsymbol{\alpha}$ é necessário maximizar $\mathrm{L}(\mathbf{D},\boldsymbol{\alpha})$ em (3–6). Isto é equivalente a determinar $\hat{\boldsymbol{\alpha}}$ tal que

$$\frac{\partial \mathrm{L}(\mathbf{D}, \hat{\boldsymbol{\alpha}})}{\partial \alpha_{ij}} = 0 \qquad (i = 1,\ldots, g-1;\, j = 0,\ldots, r)$$

Para a solução desta equação foi proposto, inicialmente, o método Newton-Raphson (Day e Kerridge (1967), Anderson (1972)). Com a evolução dos algoritmos de otimização, o método *quasi-Newton* (Gill e Murray (1972)) é o mais recomendado atualmente, por aliar rapidez e convergência mesmo com valores iniciais pouco favoráveis.

Pelo método Newton-Raphson, detalhado no Apêndice A, as estimativas de máxima verossimilhança de $\boldsymbol{\alpha}$ são obtidas através do seguinte procedimento iterativo (Anderson (1972)):

$$\hat{\boldsymbol{\alpha}}_{n+1} = \hat{\boldsymbol{\alpha}}_n - [\mathrm{I}(\hat{\boldsymbol{\alpha}}_n)]^{-1}\,\mathrm{S}(\hat{\boldsymbol{\alpha}}_n) \qquad (3-9)$$

onde $\mathrm{I}(\hat{\boldsymbol{\alpha}}_n)$ é a matriz hessiana calculada em $\hat{\boldsymbol{\alpha}}_n$ e $\mathrm{S}(\hat{\boldsymbol{\alpha}}_n)$ é a derivada de $\mathrm{L}(\mathbf{D}, \boldsymbol{\alpha})$ com relação a $\boldsymbol{\alpha}$ calculada em $\hat{\boldsymbol{\alpha}}_n$.

Anderson (1972) sugere utilizar o vetor nulo como valor inicial. Com exceção de alguns casos que veremos na Seção 3.4, a aplicação desta sugestão tem se mostrado eficiente no sentido de provocar uma convergência rápida no processo iterativo.

Usualmente a regra de parada é acionada quando o estimador $\hat{\boldsymbol{\alpha}}$ atinge a precisão desejada, isto é para um ε pré-fixado,

$$\| \hat{\boldsymbol{\alpha}}_{n+1} - \hat{\boldsymbol{\alpha}}_n \| < \varepsilon$$

Outras regras de parada são necessárias quando o procedimento $(3-9)$ não converge. Elas serão discutidas na Seção 3.4.

Uma vez obtido o vetor de estimadores de máxima verossimilhança $\hat{\boldsymbol{\alpha}}$ pode-se construir intervalos de confiança e testar hipóteses sobre os parâmetros utilizando-se de um resultado bastante conhecido da teoria assintótica dos estimadores de máxima verossimilhança (ver, por exemplo, Mardia (1979) - pág. 102):

$$\hat{\boldsymbol{\alpha}} \overset{a}{\sim} \mathrm{N}_{r+1}\left(\boldsymbol{\alpha}, -\mathrm{I}^{-1}(\boldsymbol{\alpha})\right)$$

Para a estimação da matriz de variâncias e covariâncias, o planejamento amostral é relevante. Para amostras conjuntas, a matriz hessiana calculada no passo final do procedimento iterativo gera uma boa estimativa para $\mathrm{Var}(\hat{\boldsymbol{\alpha}})$ (Albert e Lesaffre (1986)), ou seja:

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\alpha}}) = -\frac{1}{\mathrm{I}(\hat{\boldsymbol{\alpha}})}$$

Já no caso de amostras separadas, Anderson (1972) mostrou que o resultado acima é válido a menos dos seguintes ajuste nas variâncias dos interceptos:

$$\widehat{\mathrm{Var}}(\hat{\alpha}^*_{i0}) = \widehat{\mathrm{Var}}(\hat{\alpha}_{i0}) - \frac{1}{n_i} - \frac{1}{n_g} \qquad (i = 1,\ldots, g-1)$$

onde $\widehat{\mathrm{Var}}(\hat{\alpha}_{i0})$ é o elemento (h, h) da matriz $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\alpha}})$ com $h = (i-1)(r+1)+1$.

Para pequenas amostras os estimadores de máxima verossimilhança possuem um vício não desprezível. Anderson e Richardson (1979) apresentaram técnicas para a correção desse vício e, conseqüentemente, da regra discriminante.

3.3.2 - Procedimento Discriminante Normal

Se as distribuições $(\mathrm{X} \mid \Pi_i)$ $(i = 1,\ldots, g)$ forem normais multivariadas com média $\boldsymbol{\mu}_i$ e mesma matriz de variâncias e covariâncias $\boldsymbol{\Sigma}$, então os parâmetros $\boldsymbol{\alpha}_j$ $(j = 1,\ldots, g-1)$ podem ser estimados diretamente sem a necessidade de métodos iterativos.

Para verificarmos isso, lembremo-nos da suposição básica do modelo logístico (3–1):

$$\ln\left[\frac{\Pr(\mathrm{x} \mid \Pi_i)}{\Pr(\mathrm{x} \mid \Pi_g)}\right] = \boldsymbol{\beta}'_i\,\mathrm{x} \qquad (i = 1,\ldots, g-1)$$

$$\text{Se} \quad \Pr(\mathrm{x} \mid \Pi_i) = |2\,\Pi\,\Sigma|^{-p/2} \exp\left\{-\frac{1}{2}(\mathrm{x}-\boldsymbol{\mu}_i)'\,\Sigma^{-1}\,(\mathrm{x}-\boldsymbol{\mu}_i)\right\}$$

$$\text{então,} \quad \ln\frac{\Pr(\mathrm{x} \mid \Pi_i)}{\Pr(\mathrm{x} \mid \Pi_g)} = -\frac{1}{2}(\mathrm{x}-\boldsymbol{\mu}_i)'\,\Sigma^{-1}\,(\mathrm{x}-\boldsymbol{\mu}_i) + \frac{1}{2}(\mathrm{x}-\boldsymbol{\mu}_g)'\,\Sigma^{-1}\,(\mathrm{x}-\boldsymbol{\mu}_g)$$

$$= -\frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_g) + (\boldsymbol{\mu}_i - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}^{-1} \mathbf{x}$$

Assim,

$$\beta_{i0} = -\frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_g)$$

$$(\beta_{i1}, \ldots, \beta_{ir})' = \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_g) \qquad (i = 1, \ldots, g-1)$$

A partir de $(3-3)$, podemos determinar $\boldsymbol{\alpha}$,

$$\alpha_{i0} = -\frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_g) + \ln\left(\frac{p_i}{p_g}\right)$$

$$(\alpha_{i1}, \ldots, \alpha_{ip})' = \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_g) \qquad (i = 1, \ldots, g-1)$$

Observe que, neste caso, a regra de classificação coincide com a regra dada pela LDF (Seção 1.3.1).

Finalmente, utilizando os estimadores usuais de $\boldsymbol{\mu}_i$ e $\boldsymbol{\Sigma}$, $\bar{\mathbf{x}}_i$ e S_p, respectivamente, podemos estimar $\boldsymbol{\alpha}$:

$$\begin{aligned} \hat{\alpha}_{i0} &= -\tfrac{1}{2}(\bar{\mathbf{x}}_i + \bar{\mathbf{x}}_g)' S_p^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_g) + \ln(p_i/p_g) \\ (\hat{\alpha}_{i1}, \ldots, \hat{\alpha}_{ip})' &= S_p^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_g) \qquad (i = 1, \ldots, g-1) \end{aligned}$$

onde $\bar{\mathbf{x}}_i = \frac{1}{n_i} \sum_{k \in E_i} \mathbf{x}_k$ e $S_p = \frac{1}{n-g} \sum_{i=1}^{g} \sum_{k \in E_i} (\mathbf{x}_k - \bar{\mathbf{x}}_i)(\mathbf{x}_k - \bar{\mathbf{x}}_i)'$.

A utilização destas estimativas no modelo logístico é chamada de procedimento discriminante normal e sua eficiência foi estudada empiricamente por Press e Wilson (1978)

e teoricamente para $g = 2$ por Efron (1975). Recentemente, Bull e Donner (1987) generalizaram os resultados de Efron (1975). Todos estes artigos recomendam a utilização da Discriminação Logística via máxima verossimilhança se houver pelo menos uma variável binária presente.

Atualmente, o procedimento discriminante normal é recomendável principalmente para estudo exploratório. Trabalhos importantes em medicina já foram feitos utilizando-se este método, sendo um dos mais conhecidos o de Truett, Cornfield e Kannel (1967). Neste trabalho os fatores de risco relacionados com doenças cardiovasculares foram pela primeira vez medidos de forma simples.

3.3.3 - Transformação Logística Empírica

Uma das desvantagens da Discriminação Logística é a necessidade de métodos iterativos para a estimação dos parâmetros. Existe, entretanto, uma forma alternativa para a estimação de $\boldsymbol{\alpha}$ quando temos duas populações e os dados estão resumidos em uma tabela de contingência. Está implícita a necessidade das variáveis contínuas serem categorizadas e no caso de variáveis categóricas, o modelo requer que elas sejam previamente reparametrizadas (Seção 1.4).

Outra condição importante para que o método tenha um bom desempenho é que o número de observações seja suficientemente grande para que cada uma das c possíveis caselas, resultantes da combinação das categorias das diferentes variáveis, tenha um número significativo de observações (digamos, pelo menos cinco). Isso é necessário para possibilitar estimativas confiáveis dos parâmetros, haja visto que as propriedades dos estimadores foram obtidos para condições assintóticas.

Na prática, esta condição implica que o método só é viável quando temos um número não muito grande de variáveis discriminadoras. Além disso, o número de categorias diferentes de cada variável deve ser levado em consideração.

Lembrando que a suposição básica do modelo logístico (3−1) para $g = 2$ é

$$\ln\left[\frac{\Pr(\mathbf{x}_k \,|\, \Pi_1)}{\Pr(\mathbf{x}_k \,|\, \Pi_2)}\right] = \boldsymbol{\beta}'\,\mathbf{x}_k \qquad (k = 1,\ldots, c)$$

Utilizando o teorema de Bayes e a expressão (3−2), obtemos

$$\ln\left[\frac{\Pr(\Pi_1 \,|\, \mathbf{x}_k)}{\Pr(\Pi_2 \,|\, \mathbf{x}_k)}\right] = \ln\frac{p_1}{1-p_1} + \boldsymbol{\beta}'\,\mathbf{x}_k = \boldsymbol{\alpha}'\,\mathbf{x}_k$$

A regra discriminante classifica uma observação $\mathbf{x}_k$ em Π_1 se $\boldsymbol{\alpha}'\mathbf{x}_k$ for maior ou igual a 0.

A idéia é estimarmos $\Pr(\Pi_i \,|\, \mathbf{x}_k)$* diretamente da tabela de contingência e, posteriormente, considerarmos

$$y_k = \ln\left[\frac{\Pr(\Pi_1 \,|\, \mathbf{x}_k)}{\Pr(\Pi_2 \,|\, \mathbf{x}_k)}\right]$$

como variável dependente de um modelo de regressão múltipla, onde as variáveis discriminadoras $(x_1,\ldots, x_r)$ são as variáveis independentes.

Dessa forma, temos o seguinte modelo

$$y_k = \alpha_0 + \alpha_1\, x_{k1} + \cdots + \alpha_r\, x_{kr} + \varepsilon_k \qquad (k = 1,\ldots, c) \qquad (3-10)$$

Ocorre, porém, que a suposição de homocedasticidade não é satisfeita, isto é $\mathrm{Var}(y_k)$ $(k = 1,\ldots, c)$ não é constante. Assim, o método de mínimos quadrados ponderados é recomendável.

---

* Teoricamente, poderia se estimar $\ln\,[\Pr(\Pi_1 \,|\, \mathbf{x}_k)/\Pr(\Pi_2 \,|\, \mathbf{x}_k)]$.

Vamos denotar por m_k o número de observações na casela k e por f_k o número de observações pertencentes à população Π_1 na casela k ($k = 1, \ldots, c$). A Tabela 3.1 utiliza a notação empregada.

Casela	Π_1	Π_2	Total
1	f_1	$m_1 - f_1$	m_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
c	f_c	$m_c - f_c$	m_c
Total	n_1	n_2	n

Tabela 3.1 - Notação.

Cox (1970) - pág. 33 recomenda utilizar os seguintes resultados para se estimar y_k,

$$\hat{y}_k = \ln\left(\frac{f_k + 0,5}{m_k - f_k + 0,5}\right) \qquad (k = 1, \ldots, c) \qquad (3-11)$$

cuja variância assintótica é dada por

$$\mathrm{Var}(\hat{y}_k) = \frac{1}{m_k \, \Pr(\Pi_1 \,|\, \mathbf{x}_k) \, \Pr(\Pi_2 \,|\, \mathbf{x}_k)} \qquad (k = 1, \ldots, c)$$

que é, segundo Cox (1970) - pág. 34, convenientemente estimada por

$$\widehat{\mathrm{Var}}(\hat{y}_k) = \frac{(m_k + 1)(m_k + 2)}{m_k (f_k + 1)(m_k - f_k + 1)} \qquad (k = 1, \ldots, c) \qquad (3-12)$$

A estimação de $\boldsymbol{\alpha}$ em $(3-10)$ pode ser obtida através de um programa de regressão múltipla ponderada, com os pesos estimados por:

$$\hat{w}_k = \left[\widehat{\mathrm{Var}}(\hat{y}_k)\right]^{-1/2} \qquad (k = 1, \ldots, c) \qquad (3-13)$$

Se o planejamento amostral for o de amostras separadas, torna-se necessário, após a obtenção de $\hat{\boldsymbol{\alpha}}$, efetuar a correção $(3-8)$ no intercepto:

$$\hat{\alpha}_0^* = \hat{\alpha}_0 + \ln\left(\frac{p_1\, n_2}{p_2\, n_1}\right)$$

Exemplo

Um gerente de um banco quer determinar quais clientes emitem, freqüentemente, cheques de valores baixos. A intenção é fazer uma campanha contra o mau uso dos cheques com o objetivo de racionalizar os serviços. Por razões de custo ele quer determinar quais são esses clientes baseando-se em apenas três variáveis:

X_1 = número de talões de cheques requeridos por mês;

$$X_2 = \begin{cases} 1, & \text{possui cheque especial;} \\ 0, & \text{não possui cheque especial.} \end{cases}$$

$$X_3 = \begin{cases} 1, & \text{possui cartão de crédito;} \\ 0, & \text{não possui cartão de crédito.} \end{cases}$$

Isto é baseado em X_1, X_2 e X_3 ele quer classificar o correntista em

Π_1 : emite freqüentemente cheques de valores muito baixos;

Π_2 : não emite freqüentemente cheques de valores muito baixos.

Os dados extraídos de amostras conjuntas estão resumidos na Tabela 3.2 dada a seguir.

k	x_1	x_2	x_3	Π_1	Π_2	m_k
1	1	0	0	20	40	60
2	1	0	1	5	43	48
3	1	1	0	18	37	55
4	1	1	1	6	85	91
5	2	0	0	14	8	22
6	2	0	1	5	25	30
7	2	1	0	9	6	15
8	2	1	1	13	61	74
9	3	0	1	17	28	45
10	3	1	0	77	17	94
11	3	1	1	9	16	25
Total				$n_1 = 193$	$n_2 = 366$	$n = 559$

Tabela 3.2 - Dados hipotéticos.

O estimador de y_k $(k = 1, \ldots, 11)$ é dado em $(3-11)$. Por exemplo, para a primeira casela,

$$\hat{y}_1 = \ln\left[\frac{20+0,5}{(60-20+0,5)}\right] = -0,68$$

A estimação de $\hat{w}_k$ é obtida substituindo-se (3–12) em (3–13), logo,

$$\hat{w}_1 = \sqrt{\frac{60\,(20+1)\,(60-20+1)}{(60+1)\,(60+2)}} = 3,70$$

Os demais valores de $\hat{y}_k$ e $\hat{w}_k$ foram calculados e estão apresentados na Tabela 3.3.

k	$\hat{y}_k$	$\hat{\omega}_k$
1	-0,68	3,70
2	-2,07	2,27
3	-0,71	3,53
4	-2,58	2,53
5	0,53	2,32
6	-1,53	2,17
7	0,38	1,96
8	-1,52	3,36
9	-0,49	3,30
10	1,49	3,80
11	-0,55	2,46

Tabela 3.3 - Valores estimados da variável dependente (y_k) e de seu respectivo peso (w_k).

A estimação de $\hat{\boldsymbol{\alpha}} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3)$, com os respectivos erros-padrões (entre parênteses) resultou em :

$$\underset{(0,119)}{\hat{\alpha}_0 = -1,621}, \quad \underset{(0,053)}{\hat{\alpha}_1 = 1,031}, \quad \underset{(0,091)}{\hat{\alpha}_2 = -0,072} \quad \text{e} \quad \underset{(0,091)}{\hat{\alpha}_3 = -1,880}$$

Como o coeficiente da variável X_2 foi considerado não significativo, resolveu-se excluí-la do modelo, o que resultou em:

$$\underset{(0,119)}{\hat{\alpha}_0 = -1,654}, \quad \underset{(0,053)}{\hat{\alpha}_1 = 1,024} \quad \text{e} \quad \underset{(0,088)}{\hat{\alpha}_3 = -1,872} \qquad (3-14)$$

Logo, a regra discriminante (3–5) torna-se:

$$\text{classificamos} \quad \mathbf{x} \quad \text{em} \quad \Pi_1 \quad \text{se} \quad -1,654 + 1,024\, x_1 - 1,872\, x_3 \geq 0$$

Ao reclassificarmos a própria amostra obtivemos a classificação em Π_1 para $k = 5$, 7 e 10, ou seja, quando a freqüência da casela k da população Π_1 superou a da população Π_2.

É evidente que, como regra discriminante, o método multinomial (Goldstein e Dillon (1978)) que consiste em classificar $\mathbf{x}$ na população que for predominante em sua casela, é muito mais simples de ser aplicado. Entretanto, a Discriminação Logística permite obter interessantes conclusões. Por exemplo,

(i) O coeficiente de X_2 é não significativo, o que vale dizer que o fato do cliente ter ou não o cheque especial é irrelevante para a classificação uma vez que as outras variáveis estão incluídas no modelo;

(ii) É possível classificar observações que não foram representadas na amostra. Por exemplo, nenhuma das 559 observações dadas na Tabela 3.2 tem observações do tipo: $x_1 = 3$, $x_2 = 0$ e $x_3 = 0$. Contudo, matematicamente, é possível utilizar as estimativas dos coeficientes dadas em $(3-14)$ e classificar este tipo de observação: Dado $\mathbf{x} = (1 \;\; 3 \;\; 0 \;\; 0)'$, como $\boldsymbol{\alpha}'\mathbf{x} = \mathbf{1,419}$ então $\mathbf{x} \in \mathbf{R}_1$.

Certos "pacotes" computacionais estatísticos possuem o programa para regressão múltipla mas sem a opção de ponderação. Mesmo neste caso é possível obter uma estimativa de $\boldsymbol{\alpha}$ diretamente. Muitos livros de regressão apresentam este artifício que consiste em padronizar a variância dividindo-se y_k, $x_{k0}, \ldots, x_{kr}$ por $\sqrt{\mathrm{Var}(y_k)}$.

A principal vantagem no procedimento visto nesta seção é que programas de regressão múltipla são mais acessíveis que programas de regressão logística. Além disso, técnicas de seleção de variáveis, detecção de observações aberrantes , etc. desenvolvidos para problemas de regressão múltipla poderiam ser diretamente aplicadas em Discriminação Logística.

3.4 - Existência e Unicidade do Estimador de Máxima Verossimilhança

Na Seção 3.3.1 vimos que a estimação de $\boldsymbol{\alpha}$, vetor de $v = (g-1)(p+1)$ parâmetros, era obtida através da maximização da função (3–6), reproduzida abaixo,

$$\mathrm{L}(\mathbf{D}, \boldsymbol{\alpha}) = \sum_{i=1}^{g} \sum_{k \in E_i} \ln \left[\frac{1}{\sum_{h=1}^{g} \exp\left(\alpha_h - \alpha_i \right)' x_k)} \right]$$

Ocorre, entretanto, que certas configurações dos pontos amostrais levam a não existência de um estimador de máxima verossimilhança. Para nós, seguindo a literatura (Haberman (1974), Albert e Lesaffre (1986)), não existência de um estimador de máxima verossimilhança significa ausência de solução finita na maximização de $\mathrm{L}(\mathbf{D}, \boldsymbol{\alpha})$. Dessa forma, torna-se necessário detectar essas configurações com o objetivo de evitar iterações desnecessárias no algoritmo utilizado ou interpretações errôneas.

Albert e Anderson (1984) classificaram as configurações em três categorias: separação completa, separação quase completa e sobreposição. Eles verificaram que o estimador de máxima verossimilhança de $\boldsymbol{\alpha}$ existe somente no último caso. Albert e Lesaffre (1986) apresentaram novas configurações em que o estimador de máxima verossimilhança não existe. Eles as chamaram de separação parcial e separação quase parcial.

Constata-se que a não existência do estimador de máxima verossimilhança de $\boldsymbol{\alpha}$ em situações práticas é bastante rara estando praticamente restrita a duas situações: tamanhos de amostras pequenos e variáveis discriminadoras exclusivamente categóricas.

3.4.1 - Separação Completa

Dizemos que ocorre separação completa nos pontos amostrais se existir $\boldsymbol{\alpha} \in I\!R^v$ tal que para todo $k \in E_i$ e para todo $i \neq h$ $(i,\ h = 1, \ldots, g)$ tivermos:

$$(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_h)' \, \mathbf{x}_k > 0 \qquad (3-15)$$

isto é, todas observações são corretamente classificadas.

Uma representação gráfica para o caso em que $g = 2$ e temos apenas 2 variáveis $(X_1$ e $X_2)$ é :

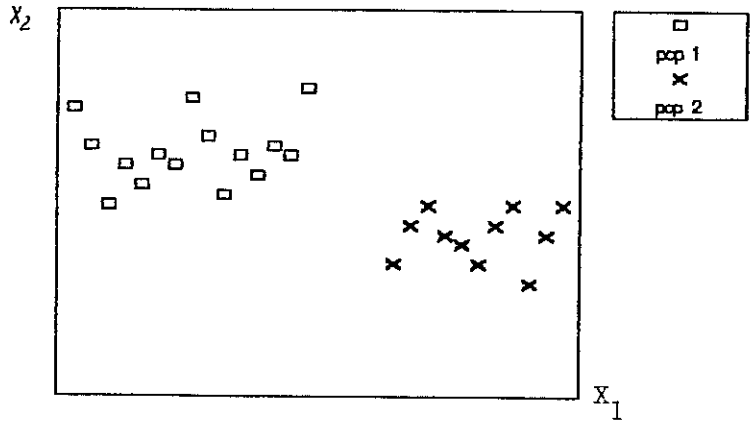


Figura 3.1 - Separação completa.

Fica evidente ao observarmos a Figura 3.1 que existem infinitos $\boldsymbol{\alpha}$ que classificam corretamente todas as observações, no caso os $\boldsymbol{\alpha}$ definem as retas que separam completamente as duas amostras.

Vamos denotar por A^c o conjunto dos vetores $\boldsymbol{\alpha}$, que satisfazem à definição acima. (A^c é um cone convexo em $I\!R^v$). Se $\boldsymbol{\alpha} \in A^c$ é sempre possível construir um ve-

tor $\boldsymbol{\Delta}$ ($\boldsymbol{\Delta} \neq t\,\boldsymbol{\alpha}$), $t \in I\!R$, tal que $(\boldsymbol{\alpha} + \boldsymbol{\Delta}) \in A^c$. O vetor $\boldsymbol{\Delta}$ deve ter norma tão pequena quanto for necessária para que $\boldsymbol{\alpha} + \boldsymbol{\Delta}$ ainda satisfaça $(3-16)$.

Deste modo, existem infinitos vetores de direções distintas em A^c. Note também, que para cada $\boldsymbol{\alpha} \in A^c$, $t\,\boldsymbol{\alpha} \in A^c$ $(t > 0)$. Para verificarmos isso basta multiplicar por t ambos os lados em (3–15). Vamos provar que o vetor que maximiza (3–6) entre os estimadores da "família" $t\,\boldsymbol{\alpha}$ é obtido quando $t \to \infty$. Se considerarmos outras "famílias" dentro do cone A^c o resultado é análogo. Como a maximização de $(3-6)$ é obtida para vetores não finitos, o estimador de máxima verossimilhança não existe.

Teorema 3.4.1: Se existir uma separação completa nos pontos amostrais, então o estimador de máxima verossimlhança $\hat{\boldsymbol{\alpha}}$ não existe e

$$\max_{\boldsymbol{\alpha} \in R^v} \mathrm{L}(\mathbf{D}, \boldsymbol{\alpha}) = 0$$

Dem: Seja $t\,\boldsymbol{\alpha} \in A^c$ com $t \in I\!R, t > 0$. Substituindo na função a ser maximizada (3–5), temos

$$\mathrm{L}(\mathbf{D}, t\,\boldsymbol{\alpha}) = \sum_{i=1}^{g} \sum_{k \in E_i} \ln \left\{ \frac{1}{\sum_{h=1}^{g} \exp\left[t\,(\boldsymbol{\alpha}_h - \boldsymbol{\alpha}_i)'\,\mathbf{x}_k\right]} \right\}$$

Observe que quando t tende a infinito, todos as parcelas da somatória de exponenciais tendem a zero (ver (3–15)) exceto para $h = i$ quando assume o valor 1 qualquer que seja o índice k. Assim, quando t tende a infinito, $\mathrm{L}(\mathbf{D}, \boldsymbol{\alpha})$ tende a zero, que é o seu valor máximo.

Portanto, como a solução $\widehat{\boldsymbol{\alpha}}$ obtida não é finita, dizemos que não existe estimador de máxima verossimilhança neste caso. ■

Configurações completamente separadas levam à divergência do procedimento iterativo (Seção 3.3.1). Anderson (1972) mostrou que, neste caso, qualquer que seja o procedimento iterativo utilizado as soluções $(\boldsymbol{\alpha}_n)$ a partir de algumas iterações são soluções que levam à completa separação das populações. Assim é necessário incluir uma regra de parada logo que um $\boldsymbol{\alpha}$ classificar corretamente todas as observações.

3.4.2 Separação Quase Completa

Dizemos que ocorre separação quase completa em $\mathbf{D}$ se existir $\boldsymbol{\alpha} \in I\!R^v, \boldsymbol{\alpha} \neq \mathbf{0}$, tal que para todo $k \in E_i$ e para todo $i \neq h$ $(i = 1,\ldots, g;\ h = 1,\ldots, g)$:

$$(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_h)' \mathbf{x}_k \geq 0 \qquad (3-16)$$

com igualdade dada para pelo menos um $(i,\ h,\ k)$.

Em outras palavras, existe um vetor $\boldsymbol{\alpha}$ que classifica corretamente todas as observações amostrais mas, para pelo menos uma observação $\mathbf{x}_k$ esta classificação não é única. Isto é, $\mathbf{x}_k$ pode ser classificada corretamente ou ser classificada em uma outra população. As observações $\mathbf{x}_k$ para os quais esta classificação não é única implicam na igualdade em (3–16).

Por exemplo, para $g = 2$ e $r = 2$, possíveis configurações de dados com separação quase completa são dadas pela Figura 3.2. No caso da Figura 3.2a $\boldsymbol{\alpha}$ é único, já a Figura 3.2b exemplifica uma situação em que existem infinitos $\boldsymbol{\alpha}$ que satisfazem à $(3-16)$.

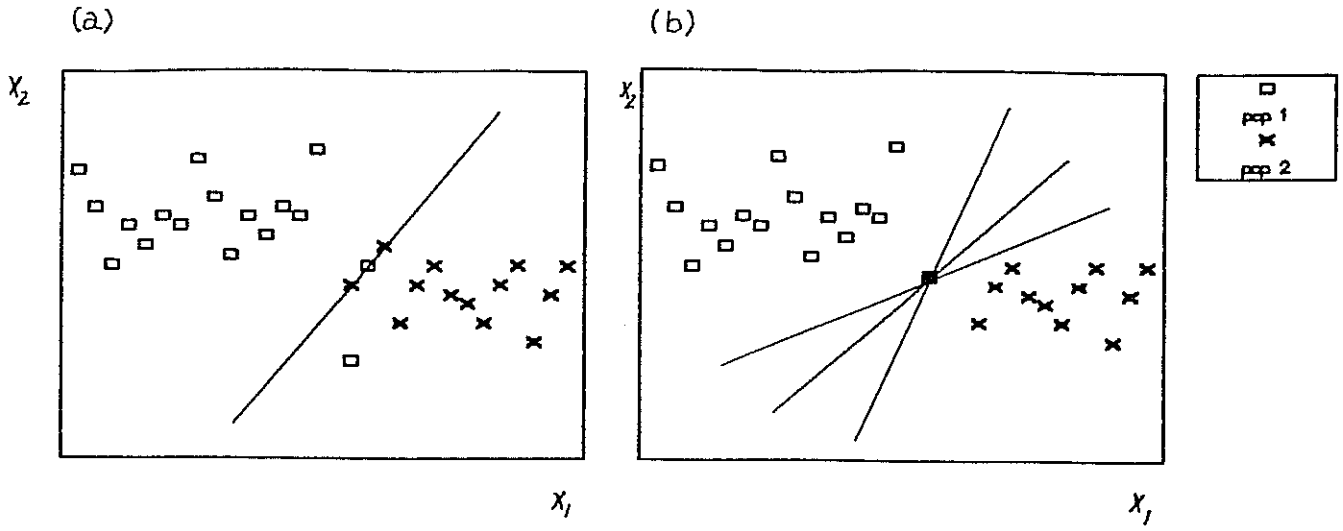


Figura 3.2 - Separação quase completa ($g = 2$ e $p = 2$).

Pode-se provar que também neste caso o estimador de máxima verossimilhança não existe. A demonstração será omitida neste trabalho em virtude de sua complexidade do ponto de vista matemático. O leitor poderá encontrá-la em Albert e Anderson (1984). Contudo, várias passagens desta demonstração foram contestadas e modificadas por Santner e Duffy (1986).

Um método empírico para detectar este tipo de configuração dos dados é parar o processo iterativo de maximização de $\mathrm{L}(\mathbf{D}, \boldsymbol{\alpha})$ se $|(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_h)' \mathbf{x}_k| < \varepsilon$, ε pré-fixado. Para evitar alarmes falsos recomendamos utilizar esta regra somente após 3 ou 4 iterações.

3.4.3 Sobreposição

Dizemos que ocorre sobreposição se existir uma trinca $(k,\ i,\ h)$ tal que para todo $\boldsymbol{\alpha} \in \mathbb{R}^v$ $(i,\ h = 1, \ldots,\ g;\ i \neq h$ e $k \in E_i)$ tivermos,

$$(\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_h)' \, \mathbf{x}_k < 0$$

ou seja, pelo menos uma observação é classificada incorretamente.

Não é necessário que todas as populações se sobreponham. Observe que na Figura 3.3, em que há um exemplo para o caso $p = 2$, não há sobreposição entre as amostras de Π_1 e Π_3.

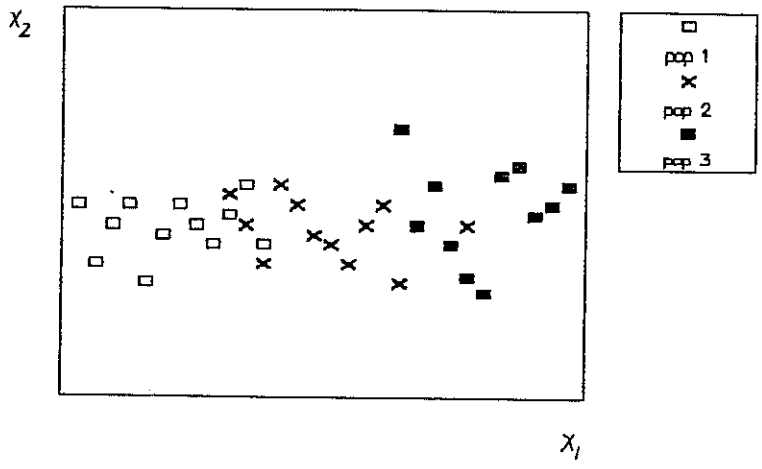


Figura 3.3 - : Sobreposição.

Teorema 3.3: Se houver sobreposição em $\mathbf{D}$, então o estimador de máxima verossimilhança existe e é único.

A demonstração deste teorema pode ser vista em Haberman (1974) - Capítulo 8 que se baseia no fato desta função ter limite $-\infty$ para $\boldsymbol{\alpha} \to +\infty$ ou $\boldsymbol{\alpha} \to -\infty$ e ser estritamente côncava.

3.4.4 Separação Parcial e Quase Parcial

Albert e Lesaffre (1986) descrevem novos tipos de amostras em que o estimador $\hat{\boldsymbol{\alpha}}$ não existe. Trata-se de uma generalização dos tipos discutidos nas Seções 3.4.1 e 3.4.2. A idéia destes novos casos podem ser facilmentes intuídas pelas figuras a seguir:

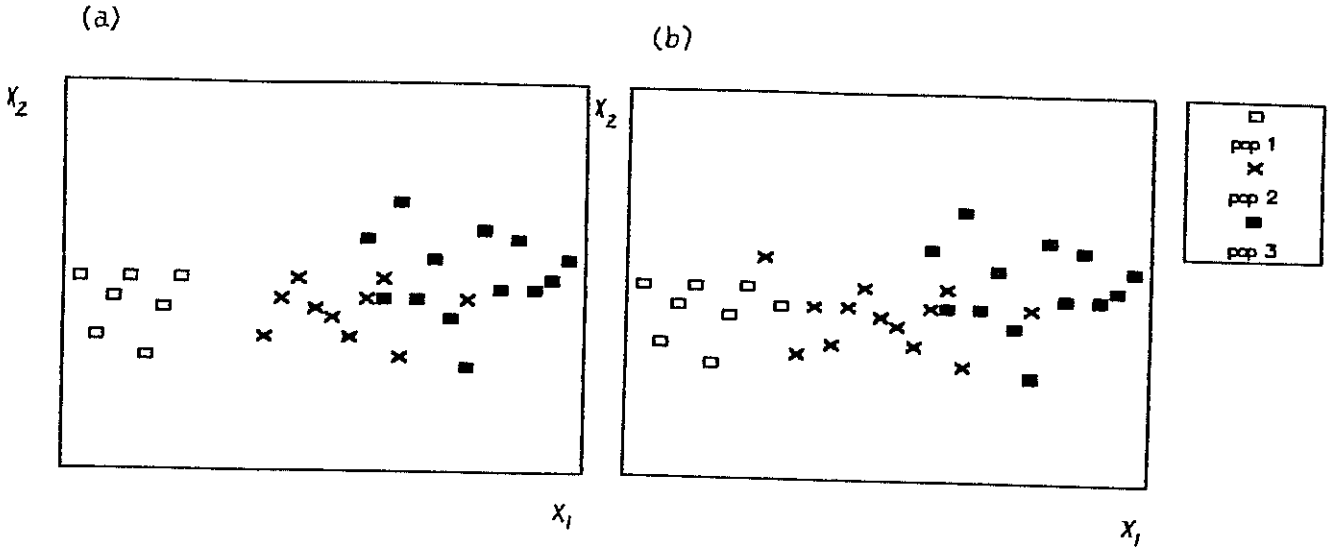


Figura 3.4. - (a) Separação parcial. (b) Separação quase parcial.

A separação parcial ocorre quando a amostra de uma população está completamente separada de outras que se sobrepõem entre si. Na Figura 3.4a a amostra da população Π_1 está completamente separada de Π_2 e Π_3.

Já na separação quase parcial uma das populações está quase completamente separada de pelo menos uma das outras. Na Figura 3.4b observa-se que Π_1 está quase completamente separada de Π_2.

A justificativa formal incluindo alguns teoremas sobre a inexistência de $\hat{\boldsymbol{\alpha}}$ nestes casos estarão em artigo de Lesaffre e Albert submetido a publicação, conforme consta em Albert e Lesaffre (1986).

3.5 Adequação do Modelo

O método mais consagrado na literatura para verificar se a Discriminação Logística para duas populações é adequado é o método de Hosmer (Hosmer e Lemeshow (1980))*.

Neste método, a hipótese nula é equivalente à suposição básica da Discriminação Logística para $g = 2$:

$$\mathrm{H}_0 \quad : \quad \mathrm{Pr}(\,\Pi_1 \,|\, \mathbf{x}\,) = \exp(\boldsymbol{\alpha}'\,\mathbf{x})/\,[\,1 + \exp(\boldsymbol{\alpha}'\,\mathbf{x})\,]$$

Para verificar se os dados amostrais obedecem a essa hipótese, o método de Hosmer compara a distribuição dos valores estimados das probabilidades *a posteriori* com os seus respectivos valores esperados sob H_0. A comparação é feita através de um teste χ^2.

O primeiro passo é fixar as constantes $c_0,\ c_1, \ldots,\ c_m$ com m fixado *a priori*. Estes c_j serão utilizados para definir as m classes que agruparão as probabilidades *a posteriori* (estimadas) de cada observação.

A escolha do conjunto de constantes deve ser tal que:

$$0 = c_0 < c_1 < \cdots < c_{m-1} < c_m = 1$$

Usualmente utiliza-se valores para m entre 5 e 10 e amplitudes constantes. Nada impede, porém, que os c_j sejam escolhidos de acordo com convenientes percentis dos valores estimados de $\mathrm{Pr}(\,\Pi_1 \,|\, \mathbf{x}_k\,)$ $(k = 1, \ldots,\ n)$.

Para simplificar a notação vamos definir:

$$\theta_k = \widehat{\mathrm{Pr}}(\,\Pi_1 \,|\, \mathbf{x}_k\,) \qquad (k = 1, \ldots, n)$$

* Este método é encontrado no BMDP (LR).

Para construir a tabela de valores observados o_{ij} é necessário definir a variável aleatória W_k que assume valores em $\{1, 2, \ldots, m\}$ da seguinte forma:

$$W_k = j \qquad \text{se} \quad c_{j-1} \leq \theta_k < c_j \qquad (j = 1, \ldots m; \quad k = 1, \ldots, n)$$

Dessa forma, o_{ij} $(i = 1, 2; \quad j = 1, \ldots, m)$ é o número de ocorrências do par $(\Pi_i, W = j)$ na amostra.

A notação empregada está tabulada abaixo:

			W		
Π	1	2	...	m	Total
1	o_{11}	o_{12}	...	o_{1m}	n_1
2	o_{21}	o_{22}	...	o_{2m}	n_2
Total	$o_{.1}$	$o_{.2}$	...	$o_{.m}$	n

O segundo passo do método de Hosmer é estimar a mesma tabela só que com os valores esperados sob H_0. Os valores esperados que denotaremos por e_{ij} são obtidos da seguinte maneira:

$$e_{ij} = n\, \widehat{\Pr}(W = j \,|\Pi_i)\, p_i \qquad (i = 1, 2; j = 1, \ldots, m) \qquad (3-17)$$

Estendemos os resultados obtidos por Hosmer e Lemeshow (1980) para poder incorporar, também, os resultados oriundos de amostragem separada e obtivemos as seguintes fórmulas:

$$\text{(i)} \qquad \widehat{\Pr}(W = j \,|\, \Pi_1) = \frac{1}{n_1} \sum_{k \in I_j} \theta_k \qquad (j = 1, \ldots, m) \qquad (3-18)$$

(ii)
$$\widehat{\Pr}(W = j \mid \Pi_2) = \frac{\widehat{\Pr}(W = j, \Pi_2)}{p_2}$$

$$= \frac{\widehat{\Pr}(W = j) - \widehat{\Pr}(W = j, \Pi_1)}{p_2}$$

$$= \frac{\widehat{\Pr}(W = j) - \widehat{\Pr}(W = j \mid \Pi_1)\, p_1}{p_2}$$

$$= \frac{1}{p_2}\left[\frac{o_{.j}}{n} - \frac{p_1}{n_1}\sum_{k \in \mathrm{I}_j} \theta_k\right] \qquad (j = 1, \ldots, m) \qquad (3-19)$$

onde $\mathrm{I}_j = \{\, k \in E \mid c_{j-1} \leq \theta_k < c_j \,\}$.

Quando a amostragem é conjunta os valores esperados se reduzem para:

$$e_{1j} = \sum_{k \in \mathrm{I}_j} \theta_k$$

$$\text{e} \quad e_{2j} = o_{.j} - e_{1j} \qquad (j = 1, \ldots, m)$$

No caso de amostragem separada basta aplicar as fórmulas (3–18) e (3–19) em (3–17).

O teste χ^2 é comumente usado para testar se existe diferença significante entre os valores observados e os previstos pelo modelo. A estatística proposta é

$$\mathrm{T}_{HL} = \sum_{i=1}^{2}\sum_{j=1}^{m} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

que, assintoticamente, segue uma distribuição χ^2 com (aproximadamente) $m - 2$ graus de liberdade.

Assim, rejeito H_0 se $\mathrm{T}_{HL} > \chi^2_{\text{crítico}}$.

Exemplo

Consideremos o exemplo da Seção 3.3.3. Naquele exemplo, a estimativa da probabilidade *a posteriori* $\Pr(\Pi_1 \mid \mathbf{x})$ era dada por,

$$\theta_k = \widehat{\Pr}(\Pi_1 \mid \mathbf{x}_k) = \frac{\exp(-1,65 + 1,02\, x_1 - 1,87\, x_3)}{1 + \exp(-1,65 + 1,02\, x_1 - 1,87\, x_3)}$$

Vamos fixar $m = 5$; $c_0 = 0$; $c_1 = 0,2$; $c_2 = 0,4$;...; $c_5 = 1,0$.

A determinação de w_k é feita verificando-se em qual intervalo θ_k está. Por exemplo

$$\text{se} \quad \theta_k = 0,3 \quad \Rightarrow \quad c_1 \leq \theta_k < c_2 \quad \Rightarrow \quad w_k = 2$$

A tabela dos dados incluindo θ_k e w_k é dada abaixo,

k	x_1	x_3	Π_1	Π_2	n_k	θ_k	w_k
1	1	0	20	40	60	0,348	2
2	1	1	5	43	48	0,076	1
3	1	0	18	37	55	0,348	2
4	1	1	6	85	91	0,076	1
5	2	0	14	8	22	0,597	3
6	2	1	5	25	30	0,186	1
7	2	0	9	6	15	0,597	3
8	2	1	13	61	74	0,186	1
9	3	1	17	28	45	0,389	2
10	3	0	77	17	94	0,805	5
11	3	1	9	16	25	0,389	4

Tabela 3.4 - Valores de $\theta_k = \Pr(\Pi_1 \mid x_k)$ e w_k definido no texto.

Que pode ser reestruturada originando a tabela dada abaixo:

Π			W			Total
	1	2	3	4	5	
1	29	55	23	9	77	193
2	214	105	14	16	17	366
Total	243	160	37	25	94	559

Tabela 3.5 - Tabela de freqüências observadas.

Os conjuntos I_j são determinados através de uma simples inspeção na Tabela 3.4.

$$I_1 = \{2,4,6,8\} \qquad I_3 = \{5,7\} \qquad I_5 = \{10\}$$

$$I_2 = \{1,3,9\} \qquad I_4 = \{11\}$$

Considerando que a amostragem foi obtida conjuntamente podemos obter os valores esperados a partir da Tabela 3.4 diretamente.

Por exemplo,

$$e_{13} = \sum_{k \in I_3} \theta_k = \sum_{k \in \{5,7\}} \theta_k = (22 + 15)\ 0,597 = 22,09$$

Procedendo de forma análoga para as demais caselas, podemos construir a seguinte tabela:

Π			W			Total
	1	2	3	4	5	
1	29,91	57,53	22,09	9,73	75,67	194,93
2	213,09	102,47	14,91	15,27	18,33	364,07
Total	243,00	160,00	37,00	25,00	94,00	559,00

Tabela 3.6 - Tabela de freqüências esperadas.

O cálculo do qui-quadrado (χ^2) resultou em

$$\mathrm{T}_{HL} = 0,508$$

cujo nível descritivo em relação a uma distribuição χ^2 com 3 g.l. é de aproximadamente 90%. Portanto, não há evidências para a rejeição de H_0.

A extensão para o caso $g > 2$ não é imediata. Sugerimos escolher uma população base, por exemplo Π_g, e aplicar o método Hosmer para Π_j *versus* Π_g ($j = 1, \ldots, g-1$), e aceitar a hipótese se todos os $g-1$ testes forem não significantes. Métodos como o de Bonferroni podem ser utilizados para garantir um nível de confiança global pré-fixado.

Ainda para o caso de duas populações, Copas (1983) e Landwehr, Pregibon e Shoemaker (1984) desenvolveram métodos gráficos para checar a adequacidade de modelos logísticos.

3.6 Seleção de Variáveis

Muitas vezes o número de variáveis discriminadoras é excessivo sendo necessário que o estatístico selecione um subconjunto de variáveis dentre todas as variáveis discriminadoras. Não raramente, a taxa de erro real decresce ao retirarmos do modelo variáveis com pouco (ou nenhum) poder discriminatório.

Nesta seção, veremos um método *stepwise* de seleção de variáveis proposto por Peduzzi, Hardy e Holford (1980). Neste método, a hipótese nula é que o parâmetro da j-ésima variável que será considerada para entrar no modelo, $\mathbf{X}_{(j)}$, é nulo. No modelo já estão as variáveis $X_{(1)},\ldots, X_{(j-1)}$ cujos parâmetros foram estimados via máxima verossimilhança.

Antes de apresentarmos a estatística do teste, vamos definir algumas notações: Seja $\mathbf{x}^{k(j)} = (x_{k(1)},\ldots, x_{k(j)})$ $(k = 1,\ldots, n;\ j = 1,\ldots, r)$ o vetor contendo os valores de j variáveis aleatórias (não necessariamente $X_1,\ldots, X_j$) para a k-ésima observação e $\boldsymbol{\alpha}^j$ o vetor contendo os parâmetros $\alpha_{(1)}, \ldots, \alpha_{(j)}$ correspondentes.

Vimos na Seção 3.2 que a obtenção dos estimadores de máxima verossimilhança era feita maximizando-se $\mathrm{L}(\mathbf{D}, \boldsymbol{\alpha}^j)$ (3–6) que aqui será denotado por:

$$\mathrm{L}(\mathbf{D}, \boldsymbol{\alpha}^j) = \sum_{i=1}^{g} \sum_{k \in E_i} \ln \Pr(\,\Pi_i, \mathbf{x}^{k(j)}\,)$$

Além disso, adaptamos as funções $\mathrm{I}(\boldsymbol{\alpha})$ e $\mathrm{S}(\boldsymbol{\alpha})$ definidas na Seção 3.3.1 da seguinte forma:

$$\mathrm{S}_j(\gamma) = \left[\frac{\partial \mathrm{L}(\mathbf{D}, \boldsymbol{\alpha}^j)}{\partial \alpha_{(j)}} \bigg|_{\boldsymbol{\alpha}^{j-1},\, \alpha_{(j)} = 0} \right]_{1 \times 1}$$

$$\mathrm{I}_j(\gamma) = \left[\frac{\partial^2 \mathrm{L}(\mathbf{D}, \boldsymbol{\alpha}^j)}{\partial \alpha_{(i)}\, \partial \alpha_{(h)}} \bigg|_{\boldsymbol{\alpha}^{j-1},\, \alpha_{(j)} = 0} \right]_{j \times j} \qquad (i,\ h = 1,\ldots, j)$$

A estatística, que denominaremos de estatística P, é dada por:

$$P(\gamma) = [\,\mathbf{O}_{j-1} \;\; \mathrm{S}_j(\gamma)]\;\mathrm{I}_j^{-1}(\gamma)\;[\,\mathbf{O}_{j-1} \;\; \mathrm{S}_j(\gamma)]' \qquad (3-19)$$

onde 0_{j-1} é um vetor linha de zeros de dimensão $(j-1)$ e $\boldsymbol{\gamma} = \begin{pmatrix} \hat{\boldsymbol{\alpha}}^{j-1} \\ 0 \end{pmatrix}$ é um vetor de dimensão j.

A estatística P é inspirada num resultado da teoria dos estimadores de máxima verossimilhança que iremos descrever sucintamente a seguir: Rao (1973) - pág. 418 mostrou que se $\hat{\boldsymbol{\alpha}}$ são os estimadores de máxima verossimilhança e $\hat{\boldsymbol{\alpha}}^*$ são os estimadores de máxima verossimilhança sujeitos, porém, a m restrições, então, sob certas condições de regularidade,

$$\mathrm{S}(\hat{\boldsymbol{\alpha}}^*)'\;\mathrm{I}^{-1}(\hat{\boldsymbol{\alpha}}^*)\;\mathrm{S}(\hat{\boldsymbol{\alpha}}^*) \quad \sim \quad \chi^2_m \qquad (3-20)$$

onde S é a matriz hessiana e I é a matriz de escores correspondentes.

Por exemplo, se $\hat{\boldsymbol{\alpha}} = \left(\hat{\boldsymbol{\alpha}}_1', \hat{\boldsymbol{\alpha}}_2'\right)'$ e $\hat{\boldsymbol{\alpha}}^* = \left(\hat{\boldsymbol{\alpha}}^{*\prime}_1, \mathbf{O}'\right)'$ podemos testar a hipótese $\boldsymbol{\alpha}_2 = 0$ aplicando o teste $(3-20)$ acima. A estatística P é semelhante a $(3-20)$ se considerarmos

$$\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}^j \qquad \text{e} \qquad \hat{\boldsymbol{\alpha}}^* = \boldsymbol{\gamma}.$$

Observe que ao contrário da estatística apresentada por Rao, os parâmetros $\hat{\boldsymbol{\alpha}}^{j-1}$não são estimados novamente na estatística P.

A distribuição da estatística P ainda não é conhecida, contudo, devido a sua similaridade com $(3-20)$ Peduzzi *et al* (1980) sugerem que se compare o valor desta estatística com o valor de uma distribuição $\chi^2_{(1)}$.

O algoritmo *stepwise* é apresentado a seguir:

(i) calcule $(3-19)$ para cada uma das variáveis $X_1, \ldots, X_r$ sem nenhuma outra variável no modelo. Seja $\boldsymbol{\alpha}_{(1)}$ o parâmetro referente à variável que gerou o valor máximo em $(3-20)$. Pare se este valor for menor que $\chi^2_{1(\text{crítico})}$;

(ii) calcule o EMV (estimador de máxima verossimilhança) de $\boldsymbol{\alpha}_{(1)}$;

(iii) calcule $(3-19)$ para cada uma das variáveis que ainda não entraram no modelo. A variável associada ao valor máximo obtido é a próxima a entrar no modelo desde que esse valor seja maior que $\chi^2_{1(\text{crítico})}$. Pare caso contrário;

(iv) calcule o EMV considerando todas as variáveis que entraram no modelo. Usando a teoria assintótica dos EMV teste a "significância" de cada parâmetro. Remova todas as variáveis cujos parâmetros foram considerados não significativos. Retorne ao passo (iii).

Observe que o método proposto não considera o intercepto. Dessa forma, após a seleção das variáveis discriminadoras deve-se modelar novamente com as variáveis selecionadas e o intercepto para a obtenção dos EMV definitivos. Outra solução seria adaptar o método de tal forma que o intercepto fosse considerado inicialmente.

Qualquer que seja a solução adotada, convém recordar que se o planejamento amostral for o de amostras separadas, o ajuste $(3-8)$ deverá ser efetuado.

Capítulo 4

MÉTODOS NÃO-PARAMÉTRICOS

O objetivo deste capítulo é descrever alguns métodos discriminantes não-paramétricos que podem ser utilizados em problemas contendo variáveis misturadas.

Descartando aqueles métodos que requerem a prévia "categorização" das variáveis quantitativas, consideramos relevantes os seguintes:

- método do Núcleo (*Kernel*);
- método dos k-Vizinhos Mais Próximos (k-*Nearest Neighbor*);
- método da Árvore Binária de Classificação.

Em geral, métodos não-paramétricos requerem amostras maiores e um tempo de processamento superior quando comparados com os métodos vistos nos capítulo anteriores.

Contudo, tais métodos são particularmente interessantes na presença de muitas variáveis categóricas ou na presença de variáveis contínuas com distribuições muito assimétricas.

Em nossa opinião, o método da Árvore desenvolvido por Breiman *et al* (1984) terá dentro de poucos anos uma popularidade superior aos outros métodos não-paramétricos, haja visto que a sua principal desvantagem (excessivo tempo de processamento) tende a desaparecer com o desenvolvimento dos computadores.

Observamos, também, a existência de um grande número de artigos recentes sobre o método do Núcleo, levando-nos a conjeturar numa maior divulgação e utilização deste método. O mesmo não ocorre com o método dos k-Vizinhos Mais Próximos que tem sido citado, em geral, apenas devido a sua importância histórica.

4.1 O Método do Núcleo

O método do Núcleo, introduzido por Parzen (1962), consiste em obter a partir (exclusivamente) da amostra uma estimativa para a densidade das variáveis discriminadoras. A sua aplicação em Análise Discriminante é feita calculando-se esta estimativa para cada população e classificando $\mathbf{x}$ de acordo com a regra ótima de classificação (Seção 1.3.1):

$$\text{Dado} \quad \mathbf{x}, \qquad \text{se} \quad \Pr(\mathbf{x} \,|\, \Pi_i)\, p_i = \max_{h=1,\ldots,g} \Pr(\mathbf{x} \,|\, \Pi_h)\, p_h \qquad \text{então,} \quad \mathbf{x} \in R_i$$

O método do Núcleo permite obter estimativas da $\Pr(\mathbf{x} \,|\, \Pi_i)$ a partir, exclusivamente, de amostras de Π_i $(i = 1, \ldots, g)$. Deste modo vamos estudar o problema genérico de estimar a densidade de uma variável aleatória multidimensional $\mathbf{X}$ a partir de uma amostra $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. A estimação dessa densidade por este método depende de $\mathbf{X}$ através de uma função (função Núcleo) que denotaremos por $K(\mathbf{x})$ con-

truída a partir da amostra. Para o nosso problema de variáveis misturadas * utilizaremos a proposição de Aitchison e Aitken (1976) que consiste em definir $K(\mathbf{x})$ como sendo o produto $K_1(\mathbf{z})\, K_2(\mathbf{y})$, onde $K_1(\mathbf{z})$ e $K_2(\mathbf{y})$ são as funções Núcleo das partes categóricas e contínuas do vetor $\mathbf{x}$ respectivamente.

Este enfoque permite tratarmos separadamente cada tipo de variável.

4.1.1 Variáveis Contínuas

Para termos uma primeira noção de como funciona este método, consideremos o caso univariado. Seja $\mathrm{Pr}(y)$ a função densidade de probabilidade de uma variável aleatória contínua Y e $\mathrm{P}\,(\,a < Y < b\,)$ a probabilidade de que Y pertença ao intervalo (a, b). Podemos aproximar $\mathrm{Pr}(y)$ para λ suficientemente pequeno da seguinte forma:

$$\mathrm{Pr}(y) \approx \frac{1}{2\,\lambda}\,\mathrm{P}\,(\,y - \lambda < Y < y + \lambda\,)$$

Assim, um estimador natural para $\mathrm{Pr}(y)$ consiste em fixar o valor de λ e estimar $\mathrm{P}\,(\,y - \lambda < Y < y + \lambda\,)$ pela proporção de pontos amostrais pertencentes ao intervalo $(y - \lambda, y + \lambda)$.

Dada uma amostra $(y_1, \ldots, y_n)$ podemos escrever esta estimativa utilizando a função indicadora $\mathrm{I}(\cdot)$ que assume o valor 1 se a proposição entre parênteses for verdadeira e 0, caso contrário.

$$\widehat{\mathrm{Pr}}(y) = \frac{1}{2\,\lambda}\,\frac{1}{n}\sum_{k=1}^{n}\mathrm{I}(y - \lambda < y_k < y + \lambda) \qquad -\infty < y < \infty$$

* Variáveis discretas e variáveis ordinais podem ser incorporadas ao problema (Seções 1.4 e 4.1.3).

Por conveniência de notação, podemos reescrever esta expressão como:

$$\widehat{\Pr}(y) = \frac{1}{\lambda}\frac{1}{n}\sum_{k\in E} K\left(\frac{y-y_k}{\lambda}\right) \qquad -\infty < y < \infty$$

onde $E = \{1,\ldots, n\}$ e

$$K\left(\frac{y-y_k}{\lambda}\right) = \begin{cases} 1/2, & \text{se } y-\lambda < y_k < y+\lambda; \\ 0, & \text{caso contrário.} \end{cases}$$

ou seja, denotando $(y-y_k)/\lambda$ por w, temos:

$$K(w) = \begin{cases} 1/2, & \text{se } |w| < 1; \\ 0, & \text{caso contrário.} \end{cases} \qquad (4-1)$$

que é a função densidade de probabilidade de uma distribuição uniforme em $(-1, 1)$.

A função $K(\cdot)$ é chamada função Núcleo (em inglês, *Kernel*) e λ é denominado parâmetro de alisamento. Este parâmetro pode assumir valores reais positivos e está relacionado com a dispersão de $\mathbf{y}$. Critérios para a sua escolha serão discutidos no final desta seção.

Exemplo 4.1

$n = 10$;

$\lambda = 0,5$; (neste exemplo, escolhido arbitrariamente)

Dados:

$$\begin{aligned} y_1 &= 1,0 \\ y_2 &= y_3 = 1,5 \\ y_4 &= y_5 = 2,0 \\ y_6 &= y_7 = 4,0 \\ y_8 &= y_9 = y_{10} = 4,5. \end{aligned}$$

Para todo y $(-\infty < y < \infty)$ pode-se calcular $\widehat{\Pr}(y)$. Por exemplo, para $y = 2$ temos

$$\widehat{\Pr}(2,0) = \frac{1}{n\,\lambda} \sum_{k \in E} K\left(\frac{2,0 - y_k}{\lambda}\right)$$

$$= \frac{1}{10 \cdot 0,5} \sum_{k \in E} K\left(\frac{2,0 - y_k}{0,5}\right)$$

$$= \frac{1}{10 \cdot 0,5}\,\frac{1}{2}\,\#\{y_k \,|\, 1,5 < y_k < 2,5;\ k \in E\}$$

$$= \frac{1}{10 \cdot 0,5}\,\frac{1}{2}\,\#\{y_4,\, y_5\} = 0,2$$

onde $\#A$ representa o número de elementos do conjunto A.

Os resultados para todos os possíveis valores de y estão na Tabela 4.1.

Valores de y	$\{y_k \,\|\, y - \lambda < y_k < y + \lambda;\ k \in E\}$	$\widehat{\Pr}(y)$
$y \leq 0,5$	$\emptyset$	0,0
$0,5 < y \leq 1,0$	$\{y_1\}$	0,1
$1,0 < y < 1,5$	$\{y_1, y_2, y_3\}$	0,3
$y = 1,5$	$\{y_2, y_3\}$	0,2
$1,5 < y < 2,0$	$\{y_2, y_3, y_4, y_5\}$	0,4
$2,0 \leq y < 2,5$	$\{y_4, y_5\}$	0,2
$2,5 \leq y < 3,0$	$\emptyset$	0,0
$3,0 \leq y \leq 3,5$	$\emptyset$	0,0
$3,5 < y \leq 4,0$	$\{y_6, y_7\}$	0,2
$4,0 < y < 4,5$	$\{y_6, y_7, y_8, y_9, y_{10}\}$	0,5
$4,5 \leq y < 5,0$	$\{y_8, y_9, y_{10}\}$	0,3
$y \geq 5,0$	$\emptyset$	0,0

Tabela 4.1 - Valores estimados para a densidade $\Pr(y)$ pelo método do Núcleo.

A representação gráfica deste exemplo pode ser vista na Figura 4.1.

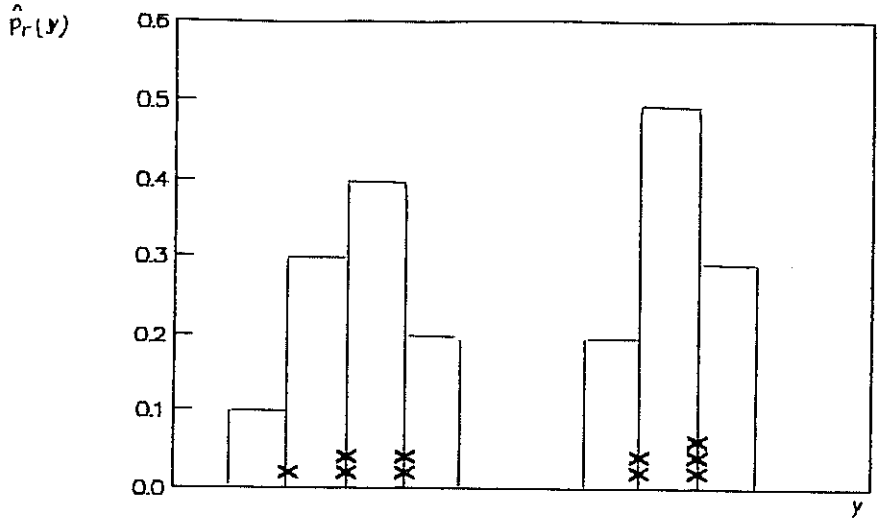


Figura 4.1 - Densidade estimada de y. As observações estão representadas por $\times$.

O estimador de $\Pr(y)$ com a função Núcleo (4−1) apresenta alguns inconvenientes (Silverman (1986) - pág. 12). O mais grave é o fato da densidade estimada não ser contínua, o que pode levar a um observador destreinado a interpretá-la erroneamente. Este problema pode ser contornado adotando outras funções Núcleo. A mais utilizada é a de Parzen (1962) que sugeriu trocar a função Núcleo (4−1) pela função densidade da distribuição normal padrão:

$$K(w) = \frac{1}{\sqrt{2\,\Pi}} \exp\left(-\frac{1}{2}\,w^2\right) \qquad -\infty < w < \infty \qquad (4-2)$$

Uma diferença desta função Núcleo com a definida em $(4-1)$ é que naquela função só contribuíam para a $\widehat{\Pr}(y)$ os valores de y_k que estavam no intervalo $(y-\lambda, y+\lambda)$. Além disso, essa contribuição era constante (igual a $1/(2\,n\,\lambda)$) não importando se y_k estava muito próximo de y ou no limite do intervalo. Já a função definida em (4−2) leva em consideração todos os y_k $(k \in E)$ com os valores próximos a y tendo um peso proporcionalmente maior na estimação da densidade.

Exemplo 4.2

Utilizando os mesmos dados do exemplo anterior e a função Núcleo (4–2) temos,

$$\widehat{\Pr}(y) = \frac{1}{n\,\lambda} \sum_{k \in E} K\left(\frac{y - y_k}{\lambda}\right)$$

$$= \frac{1}{10\,\lambda} \sum_{k \in E} \frac{1}{\sqrt{2\,\Pi}} \exp\left\{-\frac{1}{2}\left(\frac{y - y_k}{\lambda}\right)^2\right\}$$

A escolha do parâmetro de alisamento λ é fundamental. Observe o efeito nas formas das densidades estimadas para $\lambda = 0,1$; $0,5$ e $1,0$ na Figura 4.2.

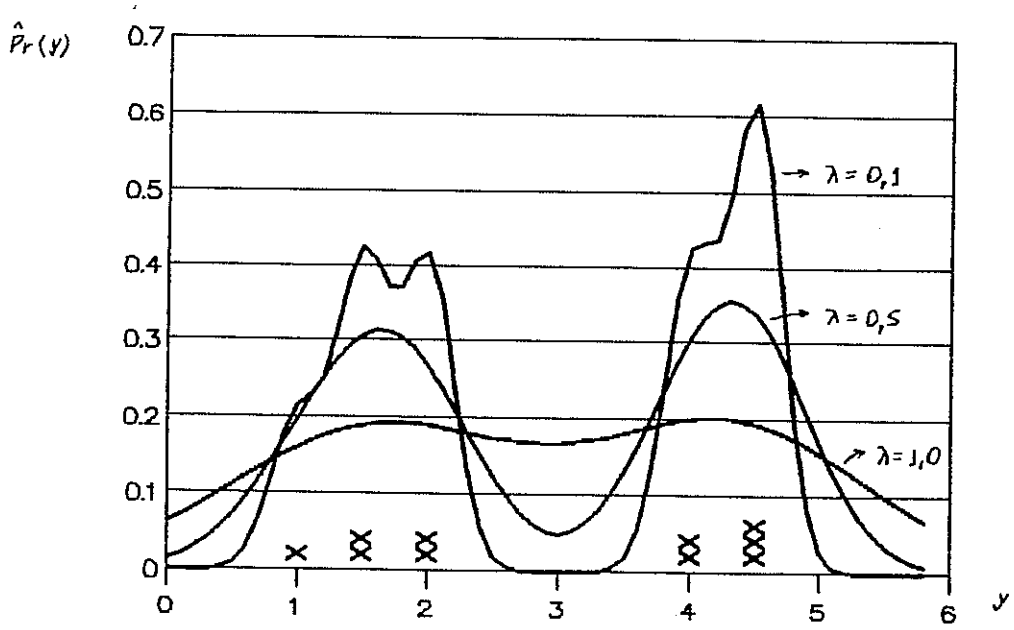


Figura 4.2 - Estimativa da densidade de y para $\lambda = 0,1$; $\lambda = 0,5$ e $\lambda = 1,0$.

Observe que o valor de λ está relacionado com a dispersão da densidade estimada. No caso da Figura 4.2 a escolha de $\lambda = 0,5$ parece ser a mais recomendada. Entretando, a escolha de λ graficamente não é possível nos casos multivariados. Veremos no final desta seção como escolher λ.

Quando $\mathbf{Y}$ é um vetor composto por p variáveis aleatórias contínuas, usualmente transforma-se cada uma das variáveis de forma a reduzi-las a variáveis com variância unitária. A seguir, estima-se a sua densidade por:

$$\widehat{\Pr}(\mathbf{y}) = \frac{1}{n\,\lambda^p} \sum_{k \in E} K\left[\frac{1}{\lambda}\,(\mathbf{y} - \mathbf{y}_k)\right] \qquad \mathbf{y} \in I\!R^p$$

onde, denotando $\frac{1}{\lambda}\,(\mathbf{y} - \mathbf{y}_k)$ por $\mathbf{w}$ temos:

$$K(\mathbf{w}) = \frac{1}{(2\,\Pi)^{p/2}} \exp\left(-\frac{1}{2}\,\mathbf{w}'\,\mathbf{w}\right) \qquad \mathbf{w} \in I\!R^p \qquad (4-3)$$

Como no caso univariado, o parâmetro λ deve ser escolhido criteriosamente. Valores muito grandes para λ geram densidades quase uniformes e valores muito pequenos tendem a gerar gráficos com vários picos e muito dependentes da particular amostra obtida (ver Figura 4.2).

Um critério natural para determinar o "melhor" λ seria através da maximização da seguinte pseudo-função de verossimilhança:

$$\prod_{k \in E} \widehat{\Pr}(\mathbf{y}_k) = \prod_{k \in E} \left\{ \frac{1}{n\,\lambda^p} \sum_{l \in E} K\left[\frac{1}{\lambda}\,(\mathbf{y}_k - \mathbf{y}_l)\right] \right\} \qquad (4-4)$$

onde $K(\cdot)$ é a função Núcleo definida em (4–3).

Ocorre, entretanto, que o valor de λ que maximiza (4–4) tende a zero (Hand (1981) - pág. 28), o que não é razoável nas aplicações práticas.

Inspirados no método da exclusão (Seção 1.2), Habbema, Herman e Den Broek (1974) propuseram que a estimação de $\Pr(\mathbf{y}_k)$ em $(4-4)$ fosse calculada baseando-se na amostra $D = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ excluindo-se o elemento $\{\mathbf{y}_k\}$. Em resumo, ao invés de se estimar por $\widehat{\Pr}(\mathbf{y}_k \,|\, D)$, estima-se por $\widehat{\Pr}(\mathbf{y}_k \,|\, D - \{\mathbf{y}_k\})$. Dessa forma o "melhor" λ é obtido através da maximização de

$$\prod_{k\in E} \widehat{\Pr}(\mathbf{y}_k \,|\, D - \{\mathbf{y}_k\}) = \prod_{k\in E} \left\{ \frac{1}{(n-1)\,\lambda^p} \sum_{\substack{l\in E \\ l\neq k}} K\left[\frac{1}{\lambda}\,(\mathbf{y}_k - \mathbf{y}_l)\right] \right\}$$

$$= \left[\frac{1}{n-1}\frac{1}{(2\,\Pi)^{p/2}\,\lambda^p}\right]^n \prod_{k\in E} \left\{ \sum_{\substack{l\in E \\ l\neq k}} \exp\left[-\frac{1}{2\,\lambda^2}\,(\mathbf{y}_k - \mathbf{y}_l)'\,(\mathbf{y}_k - \mathbf{y}_l)\right] \right\} \qquad (4-5)$$

O valor de λ que maximiza a função acima poderia ser obtida através do método Newton-Raphson (ver Seção 3.3.1). Entretanto, como as derivadas de primeira e segunda ordem não são programáveis de maneira simples, convém utilizar outros métodos de otimização não linear.

Na Figura 4.3 abaixo, apresentamos o gráfico da pseudo-função de verossimilhança (4–5) em função de λ calculado com os dados do Exemplo 4.1. O valor de λ que maximiza essa função é aproximadamente $0,35$.

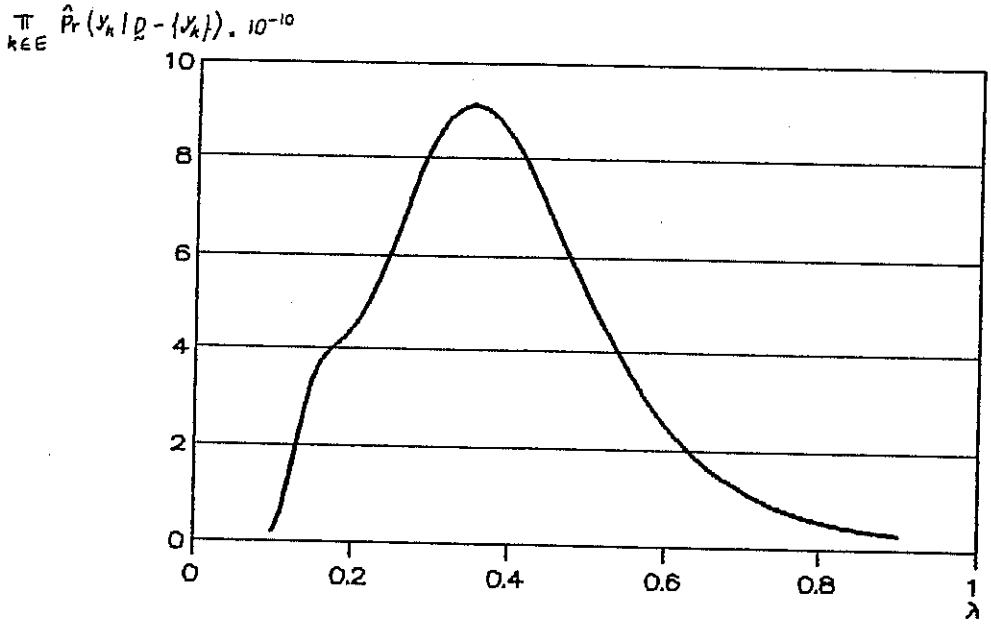


Figura 4.3 - Pseudo-função de verossimilhança em função de λ.

Em referência mais recente, Silverman (1986) - páginas 84 a 88 - apresenta outros critérios para a escolha do parâmetro de alisamento.

4.1.2 Variáveis Qualitativas Nominais

Consideremos o problema de estimar a função de probabilidade de um vetor $\mathbf{Z}$ composto por q variáveis qualitativas $Z_1, \ldots, Z_q$ a partir de uma amostra $D = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$. Como definimos anteriormente, c_j $(j = 1, \ldots, q)$ é o número de categorias diferentes que a variável Z_j pode assumir.

Inicialmente, vamos supor que as variáveis sejam binárias $(c_j = 2; j = 1, \ldots, q)$. A distribuição de probabilidades de $\mathbf{Z}$ é dada pelas probabilidades de cada uma das 2^q possíveis caselas geradas pelas combinações de resultados de $\mathbf{Z}$. Na prática, freqüentemente, o tamanho da amostra é pequeno em relação ao número de caselas impossibilitando que as estimativas usuais sejam confiáveis.

A idéia do método do Núcleo para este caso é estimar as freqüências esperadas de cada casela considerando-se todas as observações da amostra de forma ponderada. Assim, para estimar a freqüência da casela, digamos casela c, atribuímos às observações que pertencem a esta casela o peso máximo: δ^q, onde δ é o parâmetro de alisamento. Já as observações que pertencem as caselas que diferem de c por causa de uma variável apenas, entrariam com o peso $(1-\delta)\delta^{q-1}$, e assim sucessivamente. Estima-se a probabilidade de uma casela dividindo-se a freqüência esperada pelo tamanho da amostra.

Para que este critério de ponderação seja coerente, isto é, dê mais peso às observações próximas da casela cuja probabilidade estamos estimando, é necessário que $0,5 < \delta < 1$.

Por exemplo, seja $q = 2$ e os dados resumidos na Tabela 4.2.

	z_1	
z_2	0	1
0	2	3
1	4	0

Tabela 4.2 - Freqüências observadas.

A estimação da freqüência da primeira casela ($\mathbf{z} = (0,0)'$) é feita somando-se $2\,\delta^2$ referente às duas observações desta casela mais $(3+4)\,(1-\delta)\,\delta^{2-1}$, referente às observações das caselas 2 e 3 ($\mathbf{z} = (1,0)'$ e $(0,1)'$, respectivamente) que diferem da primeira por uma variável apenas e, finalmente, somando-se $0\,(1-\delta)^2$ referente à última casela; o que resulta em $7\,\delta - 5\,\delta^2$. Os valores esperados para todas as caselas considerando $\delta = 0,50$; 0,75 e 1,00 estão na Tabela 4.3.

	z_1	
z_2	0	1
0	2,250	2,250
1	2,250	2,250

$\delta = 0,50$

	z_1	
z_2	0	1
0	2,438	2,312
1	2,812	1,438

$\delta = 0,75$

	z_1	
z_2	0	1
0	2,000	3,000
1	4,000	0,000

$\delta = 1,00$

Tabela 4.3 - Freqüências esperadas.

Observe na tabela acima que a escolha de $\delta = 0;50$ implica numa distribuição uniforme para todas as probabilidades das caselas e $\delta = 1,00$ reproduz a tabela de valores observados.

A escolha do melhor δ é feita maximizando uma função baseada no método da exclusão. Freqüentemente, δ é escolhido pelo algoritmo como sendo um valor próximo a 1. Isto significa que a probabilidade estimada em cada casela é bastante próxima à proporção de observações pertencentes a esta casela.

Aitchison e Aitken (1976) sugerem estender a função Núcleo para variáveis categóricas da seguinte forma:

$$\widehat{\Pr}(\mathbf{z}) = \frac{1}{n} \sum_{k \in E} K(\mathbf{z},\ \mathbf{z}_k)$$

onde $K(\mathbf{z},\ \mathbf{z}_k) = \prod_{j=1}^{q} K_0(z_j - z_{kj})$,

$$\text{com} \qquad K_0(z_j - z_{kj}) = \begin{cases} \delta, & \text{se } z_j - z_{kj} = 0; \\ (1-\delta)/(c_j - 1), & \text{c. c.} \end{cases} \qquad (4-6)$$

Se c_j $(j = 1, \ldots, q)$ for igual a uma constante c' pode-se reescrever esta fórmula de modo sintetizado:

$$\widehat{\Pr}(\mathbf{z}) = \frac{1}{n} \sum_{k \in E} \delta^{q-d} \left(\frac{1-\delta}{c'-1} \right)^d \qquad (4-7)$$

onde d é o número de elementos diferentes de zero do vetor $(\mathbf{z} - \mathbf{z}_k)$.

No caso binário, restringimos o espaço de variação de δ ao intervalo $(0,5; 1)$ para que os "pesos" atuassem de forma coerente. Para o caso geral, a mesma restrição torna-se

$$\delta > \frac{1-\delta}{c_j - 1} \quad \Rightarrow \quad \delta > \frac{1}{c_j} \qquad (j = 1, \ldots, q)$$

ou seja, δ deve ser escolhido no intervalo $A = \bigcap_{j=1}^{q} \left(\frac{1}{c_j}, 1 \right)$.

Para a escolha de δ, novamente o método de exclusão é recomendado uma vez que a maximização com relação a δ de $\prod_{k \in E} \Pr(\, z_k \,|\, D\,)$ leva ao valor $\delta = 1$ que, como vimos no exemplo, não é satisfatório, pois gera as mesmas freqüências observadas.

Aplicando o método da exclusão, como no caso contínuo, temos que escolher δ ($\delta \in A$) que maximize

$$\prod_{k \in E} \Pr(\, z_k \,|\, D - \{z_k\}\,) = \prod_{k \in E} \frac{1}{n} \sum_{\substack{l \in E \\ l \neq k}} K(\mathbf{z}_l,\ \mathbf{z}_k) \qquad (4-8)$$

onde $K(\cdot)$ é a função Núcleo definida em $(4-6)$.

Todos os comentários feitos quando da determinação de λ no caso contínuo valem para a estimação de δ. Inclusive, no que se refere à inviabilidade da maximização pelo método Newton-Raphson.

Apenas para ilustração, calculamos o valor "ótimo" de δ pelo método da exclusão para os dados da Tabela 4.2 e obtivemos $\delta = 0,966$ que gerou os valores esperados apresentados na Tabela 4.4.

	z_1	
z_2	0	1
0	2,0962 (2)	2,8698 (3)
1	3,8018 (4)	0,2322 (0)

Tabela 4.4 - Valores esperados para $\delta = 0,966$. Valores observados entre parênteses.

A utilização da função Núcleo (4–6) para variáveis categóricas e do algoritmo de seleção de δ tem dado resultados satisfatórios. Por exemplo, Aitchison e Aitcken (1976) avaliaram os dados apresentados por Anderson (1976) e verificaram que o método do Núcleo

apresentou desempenho superior ao da Discriminação Logística que por sua vez havia tido um desempenho superior à LDF. Em nossa opinião, este resultado não é surpreendente uma vez que em qualquer caso de reversão (Seção 2.1) o método do Núcleo tende a ter um desempenho superior aos métodos LDF e Discriminação Logística.

4.1.3 Variáveis Qualitativas Ordinais

Uma extensão possível é considerar a estrutura de ordem quando as variáveis são qualitativas ordinais. A função $K_0(z_j - z_{kj})$ da função Núcleo $(4-6)$ deveria, neste caso, assumir valores maiores conforme a diferença entre z_j e z_{kj} for diminuindo. Por exemplo, Hand (1981 - pág. 101) sugere utilizar a seguinte função quando $c_j = 3$:

$$K_0(z_j - z_{kj}) = \begin{cases} \delta_1, & \text{se } |z_j - z_{kj}| = 0; \\ \delta_2, & \text{se } |z_j - z_{kj}| = 1; \\ \delta_3, & \text{se } |z_j - z_{kj}| = 2. \end{cases}$$

onde $\delta_1 > \delta_2 > \delta_3$ e $\delta_1 + \delta_2 + \delta_3 = 1$.

O inconveniente desta função é que ela envolve três parâmetros: δ_1, δ_2 e δ_3 cuja escolha pelo método da exclusão não é trivial.

Para o mesmo caso ($c_j = 3$) a função Núcleo sugerida por Aitchison e Aitken (1976) está na Tabela 4.5 a seguir.

A função sugerida envolve apenas um parâmetro e, de fato, atua de forma coerente para δ no intervalo $\left(\frac{2}{3}, 1\right)$. "Atuar de forma coerente" no caso de variáveis ordinais é dar um peso maior conforme a diferença entre z_j e z_{kj} for menor. Entretanto, esta função traz dois problemas: o espaço de valores válidos para δ é bastante limitado e não existe uma forma única de generalização para $c_j > 3$ a partir da tabela proposta.

z_{kj}	z_j 1	2	3
1	δ^2	$2\,\delta\,(1-\delta)$	$(1-\delta)^2$
2	$(1-\delta^2)/2$	δ^2	$(1-\delta)^2/2$
3	$(1-\delta^2)$	$2\,\delta\,(1-\delta)$	δ^2

Tabela 4.5 - "Pesos" para a estimação da densidade de $\mathbf{Z}$ quando Z_j é uma variável qualitativa ordinal com três categorias. $\left(\frac{2}{3} < \delta < 1\right)$.

Considerando a inadequação das funções apresentadas na literatura pesquisada para o caso de variáveis qualitativas ordinais, sugerimos a utilização de uma função Núcleo baseada na distribuição de Poisson. A função que propomos envolve apenas um parâmetro, não limita ainda mais o espaço de variação de δ, atua de forma "coerente" e, por fim, permite a geração de tabelas de pesos para qualquer valor de c_j.

Ainda para o caso $c_j = 3$, esta função está apresentada na Tabela 4.6, onde S_i $(i = 1,\ldots, 3)$ é uma constante de padronização.

z_{kj}	z_j 1	2	3
1	$\exp(-\delta)/S_1$	$\delta\,\exp(-\delta)/S_1$	$\delta^2\,\exp(-\delta)/S_1$
2	$\delta\,\exp(-\delta)/S_2$	$\exp(-\delta)/S_2$	$\delta\,\exp(-\delta)/S_2$
3	$\delta^2\,\exp(-\delta)/S_3$	$\delta\,\exp(-\delta)/S_3$	$\exp(-\delta)/S_3$

Tabela 4.6 - "Pesos" relacionados com variável qualitativa ordinal Z_j com três categorias. $S_1 = S_3 = \exp(-\delta) + \delta\,\exp(-\delta) + \delta^2\,\exp(-\delta)$ e $S_2 = \exp(-\delta) + 2\,\delta\,\exp(-\delta)$.

Extensões desta função Núcleo para $c_j > 3$ são imediatas:

$$K_0(z_j - z_{kj}) = \frac{\exp(-\delta)\,\delta^{|z_j - z_{kj}|}}{|z_j - z_{kj}|!\,S(z_{kj})}$$

onde $S(z_{kj}) = \sum_{z_j=0}^{c_j-1} \exp(-\delta)\,\delta^{|z_j - z_{kj}|}/|z_j - z_{kj}|!$ é a constante de padronização.

Observe que

$$K_0(w) = \frac{\delta}{w}\,K_0\,(w-1) \qquad (w = 1,\ldots,\,c_j)$$

Como $0 < \delta < 1$ e $1 \leq w \leq c_j$, $\delta/w < 1$ o que implica que

$$K_0(w) < K_0(w-1) \qquad (w = 1,\ldots,\,c_j)$$

Portanto, o valor da função Núcleo quando a diferença entre z_j e z_{kj} for w será sempre inferior ao valor da mesma função se a diferença for $w-1$.

Para que os pesos sejam coerentes a única restrição é que $0 < \delta < 1$, pois, desse modo, fixado z_{kj}, sempre teremos $K_0(0) > K_0(1) > \cdots > K_0(c_j - 1)$ e $\sum_{z_j=0}^{c_j-1} K_0(z_j - z_{kj}) = 1$, que são propriedades bastante interessantes, senão essenciais, de uma função Núcleo para variáveis qualitativas ordinais.

4.1.4 Variáveis Misturadas

Finalmente, estamos aptos a expor o método do Núcleo quando temos uma mistura de variáveis contínuas e qualitativas.

Como antes, $\mathbf{X}$ é um vetor aleatório composto pelo vetor de variáveis contínuas $\mathbf{Y}$ e pelo vetor de variáveis categóricas $\mathbf{Z}$, isto é, $\mathbf{X}' = (\mathbf{Z}'\ \mathbf{Y}')$. A dimensão de $\mathbf{Z}$ é $q \times 1$ e de $\mathbf{Y}$ é $p \times 1$.

Aitchison e Aitken (1976) sugeriram que $\Pr(\mathbf{x})$ fosse estimada da seguinte forma:

$$\widehat{\Pr}(\mathbf{x}) = \frac{1}{n} \sum_{k \in E} K_1 \left[\frac{1}{\lambda} (\mathbf{y} - \mathbf{y}_k) \right] K_2(\mathbf{z}, \mathbf{z}_k) \qquad (4-9)$$

onde $K_1(\cdot)$ é a função Núcleo para variáveis contínuas definida em $(4-3)$ com parâmetro de alisamento λ e $K_2(\cdot)$ é a função Núcleo para variáveis qualitativas definida em $(4-6)$ com parâmetro de alisamento δ.

Convém observar que a fatoração da função Núcleo em $K_1(\cdot)$ e $K_2(\cdot)$ não implica, necessariamente, independência entre $\mathbf{Y}$ e $\mathbf{Z}$. De acordo com Aitchison e Aitken (1976) é possível descrever modelos com dependência complexa em termos de produto de funções Núcleo como em $(4-9)$.

A determinação de λ e δ pode ser feita simultaneamente através da maximização conjunta da pseudo-função de verossimilhança:

$$\prod_{k \in E} \widehat{\Pr}\left(\mathbf{x}_k \,|\, D - \{\mathbf{x}_k\} \right)$$

Entretanto, Vlachonikolis e Marriott (1982) constataram, empiricamente, que a determinação de λ e δ feita separadamente acarreta melhores resultados em relação a taxas de erros. Este procedimento foi adotado em nosso programa AD.

Constatamos, através de simulações, que a escolha do parâmetro de alisamento da função Núcleo para variáveis contínuas é muito influenciada por valores aberrantes. Deste modo, sugerimos que seja feita uma análise crítica dos dados, excluindo-se os eventuais *outliers*, antes da escolha de λ.

Em termos de taxas de erro, acreditamos que este método seja um dos mais eficazes uma vez que consegue ter bom desempenho mesmo na presença de interações (reversão) e necessita para isso, amostras menores que as requeridas pelo método de Krzanowski.

4.2 O Método dos k-Vizinhos Mais Próximos

O método *k-Nearest Neighbor (k-NN)* que traduzimos como "k-Vizinhos Mais Próximos (k-VMP)" foi desenvolvido por Fix e Hodges (1951). É considerado como o primeiro método não paramétrico utilizado em Análise Discriminante.

A idéia deste método é bastante intuitiva: Suponha que, no caso univariado, queremos estimar $\Pr(y)$, Y variável contínua, baseado em uma amostra $(y_1, \ldots, y_n)$. Seja $\mathrm{P}\,(\,a < Y < b\,)$ a probabilidade de Y pertencer ao intervalo (a, b). No método do Núcleo fixamos λ e aproximamos a densidade $\Pr(y)$ por:

$$\Pr(y) \approx \frac{1}{2\,\lambda}\,\mathrm{P}\,(\,y - \lambda < Y < y + \lambda\,)$$

No método dos k-VMP, o λ é tomado em função dos dados, mais explicitamente, ele é igual à distância de y à k-ésima observação mais próxima. Denotando esta distância por $d_{\mathrm{k}}(y)$, a aproximação para a densidade torna-se:

$$\Pr(y) \approx \frac{1}{2\,d_{\mathrm{k}}(y)}\,\mathrm{P}\,\big(\,y - d_{\mathrm{k}}(y) < Y < y + d_{\mathrm{k}}(y)\,\big)$$

Um estimador natural para $\mathrm{P}\,\big(y - d_{\mathrm{k}}(y) < Y < y + d_{\mathrm{k}}(y)\big)$ é, simplesmente, k/n pois no intervalo $\big(\,y - d_{\mathrm{k}}(y), y + d_{\mathrm{k}}(y)\,\big)$ há, por definição, exatamente k observações amostrais. Logo,

$$\widehat{\Pr}(y) = \frac{\mathrm{k}/n}{2\,d_{\mathrm{k}}(y)} \qquad (4-10)$$

O objetivo é obtermos uma expressão para $\Pr(\Pi_i\,|\,y\,)$ a fim de utilizarmos a regra ótima de classificação. Seja k_i o número de observações de Π_i que pertencem ao intervalo $\big(\,y - d_{\mathrm{k}}(y), y + d_{\mathrm{k}}(y)\,\big)$. Como o número de observações amostrais de Π_i é n_i, podemos estimar $\Pr(\,y\,|\,\Pi_i\,)$ de modo análogo à estimação de $\Pr(y)$, o que resulta em:

$$\widehat{\Pr}(y \mid \Pi_i) = \frac{k_i/n_i}{2\, d_k(y)}$$

Finalmente, aplicando o Teorema de Bayes, temos:

$$\widehat{\Pr}(\Pi_i \mid y) = \frac{\widehat{\Pr}(y \mid \Pi_i)\, p_i}{\widehat{\Pr}(y)} = \frac{k_i/n_i}{2\, d_k(y)} \frac{2\, d_k(y)}{k/n} p_i = \frac{k_i}{k} \frac{n}{n_i} p_i \qquad (4-11)$$

Quando a amostragem é conjunta, usualmente p_i é estimado por n_i/n e a regra de classificação resume-se, então, a classificar y na população que for predominante entre as k observações mais próximas de y.

Note que em $(4-11)$, apesar do valor da distância $d_k(y)$ ser irrelevante, a métrica utilizada é de suma importância, uma vez que é ela quem determina quais são as k observações mais próximas a y.

A generalização para o caso multivariado é análoga resultando na mesma expressão (4–11). Dessa forma, a regra de classificação do método dos k-VMP consiste em:

$$\text{Dado } \mathbf{y}, \quad \text{se } \frac{k_i}{k} \frac{n}{n_i} p_i = \max_{h=1,\ldots,g} \frac{k_h}{k} \frac{n}{n_h} p_h \quad \text{então, } \mathbf{y} \in R_i \qquad (4-12)$$

Duas questões merecem alguns comentários: qual valor deve ser dado a k e qual deve ser a medida de distância apropriada.

Para o problema de variáveis contínuas no caso multivariado, Enas e Choi (1986) verificaram, empiricamente, que o k ótimo* para amostras pequenas e duas populações é um valor próximo a $n^{1/4}$ ou $n^{3/8}$ dependendo de fatores como tamanho das amostras, matrizes de dispersão e probabilidades *a priori*.

---

* Por definição, k ótimo é o k, inteiro, que minimiza $\mathrm{E}\left[\left[\Pr(\mathbf{y}) - \widehat{\Pr}(\mathbf{y} \mid \mathrm{k})\right]^2\right]$

Ainda para o caso de variáveis contínuas as medidas de distâncias mais utilizadas são as distâncias euclidiana e a de Mahalanobis. Sendo esta última mais recomendada no caso de populações com matrizes de dispersão muito diferentes. Em comparação com os métodos LDF e Discriminação Logística, Enas e Choi (1986) constataram o melhor desempenho do método dos k-VMP no caso de amostras pequenas ($n < 100$).

Estas mesmas questões são um problema em aberto quando temos variáveis misturadas. Uma solução prática para a escolha de k (Silverman (1986) - pág. 99) é testar diversos valores diferentes e escolher o k que minimizar uma estimativa da taxa de erro real.

Quanto ao problema da definição de uma medida de distância apropriada para variáveis misturadas, pode-se utilizar, por exemplo, a medida de similaridade para este tipo de variáveis sugerida por Gower (1971) e, então, transformá-la em uma medida de distância. Dessa forma, a partir da definição de uma medida de distância apropriada, o método dos k-VMP pode ser aplicado para variáveis misturadas utilizando-se das mesmas regras definidas anteriormente.

Ocorre, porém, que existe uma conexão direta entre o método do Núcleo e o método dos k-VMP. Conforme Silverman (1986) - pág. 97 não há o que escolher entre estes dois métodos, pois o método dos k-VMP é um caso particular do método do Núcleo.

Para elucidar esta questão observe que a estimativa da densidade de X dada em $(4-10)$ pode ser extendida de forma análoga ao caso multivariado resultando em,

$$\widehat{\mathrm{Pr}}(\mathrm{x}) = \frac{\mathrm{k}/n}{c_r \left[d_{\mathrm{k}}(\mathrm{x})\right]^r} \,. \qquad (4-13)$$

onde r é o número de variáveis discriminadoras e c_r é o volume de uma hiper-esfera de raio unitário em r dimensões (isto é, $c_1 = 2$, $c_2 = \Pi$, $c_3 = 4\Pi/3$, etc.).

A mesma expressão (4–13) pode ser escrita utilizando uma função Núcleo e considerando $d_{\mathrm{k}}(\mathbf{x})$ como parâmetro de alisamento da seguinte forma:

$$\widehat{\Pr}(\mathbf{x}) = \frac{1}{n}\frac{1}{\left[d_{\mathrm{k}}(\mathbf{x})\right]^{r}}\sum_{k\in E} K\left(\frac{\mathbf{x}-\mathbf{x}_{\mathrm{k}}}{d_{\mathrm{k}}(\mathbf{x})}\right)$$

$$\text{onde} \quad K(\mathbf{w}) = \begin{cases} 1/c_r, & \text{se } |\mathbf{w}| \leq 1; \\ 0, & \text{caso contrário.} \end{cases}$$

Ou seja, ao fixarmos o valor de k no método dos k-VMP estamos, na realidade, estipulando um valor para λ numa função Núcleo que gera uma densidade (estimada) para $\widehat{\Pr}(x)$ não contínua (ver Exemplo 4.1).

Portanto, não é aconselhável, no estágio atual da teoria do método dos k-VMP, a sua utilização em Análise Discriminante nos problemas com variáveis misturadas. A nossa experiência para este tipo de variáveis tem mostrado que o desempenho do método do Núcleo, com liberdade de escolha do parâmetro de alisamento, é sempre superior ao método dos k-VMP para diversos valores de k e com a medida de distância sugerida por Gower (1971). Provavelmente, o problema esteja na medida de distância utilizada ou na falta de um bom critério para se determinar o "melhor" k. Enfim, enquanto novos estudos não esclarecem estas questões convém utilizar o método do Núcleo cuja teoria já generalizou de forma comprovadamente eficaz as estimativas de densidades de variáveis misturadas.

4.3 Método da Árvore Binária de Classificação

Pode-se dizer que o método da Árvore Binária de Classificação é fruto da evolução dos computadores, pois seria impraticável a sua execução sem o auxílio desses. A principal referência é o livro de Breiman, Friedman, Olshen e Stone (1984): *Classification and Regression Trees*. Trata-se de uma extensão do método AID (*Automatic Interaction Detection*) desenvolvido por Morgan e Sonquist (1963).

A idéia do método da Árvore pode ser percebida acompanhando-se os gráficos da Figura 4.4. Neste exemplo simplificado, Π_1 representa a população dos funcionários aptos a exercer determinada função e Π_2 os não aptos. As variáveis discriminadoras são:

$$X_1 = \text{tempo de serviço na empresa} \qquad (0 \leq X_1 \leq 20);$$

$$X_2 = \text{nota em um teste prático} \qquad (0 \leq X_2 \leq 10).$$

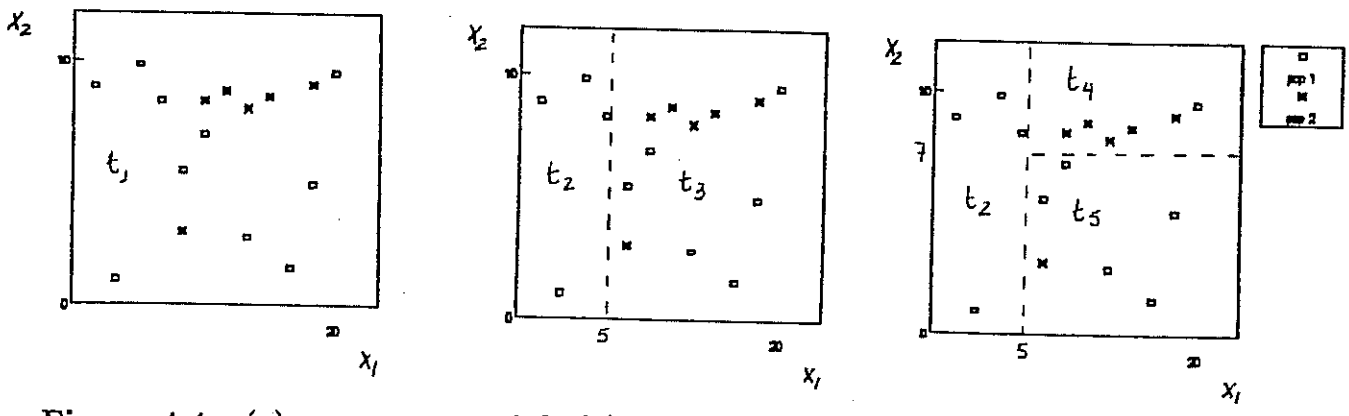


Figura 4.4 - (a) espaço amostral Ω; (b) partição de Ω em t_2 e t_3; (c) partição de t_3 em t_4 e t_5.

A primeira partição (Figura 4.4b) é obtida verificando-se qual é a reta ($x_1 = k_1$,

$0 < k_1 < 20$ ou $x_2 = k_2, 0 < k_2 < 10$) que melhor "separa" as populações em Ω. O resultado foi a partição de $\Omega = t_1$ em t_2 e t_3 onde

$$\begin{aligned} t_2 &= \{\mathbf{x} \in t_1 \,|\, x_1 \leq 5\} \\ \text{e} \qquad t_3 &= \{\mathbf{x} \in t_1 \,|\, x_1 > 5\} \end{aligned}$$

Tentou-se particionar t_2 mas nenhuma partição foi considerada significativa. O mesmo não ocorreu com t_3 (ver Figura 4.4c) que gerou as partições t_4 e t_5:

$$\begin{aligned} t_4 &= \{\mathbf{x} \in t_3 \,|\, x_1 \leq 7\} \\ \text{e} \qquad t_5 &= \{\mathbf{x} \in t_3 \,|\, x_1 > 7\} \end{aligned}$$

Finalmente, novas partições em t_4 e t_5 não foram consideradas significativas.

Observe que a idéia é particionar Ω em retângulos cada vez mais homogêneos ("puros") em relação às populações até que todos os retângulos sejam considerados suficientemente homogêneos para não precisarem mais serem particionados.

O nome método da Árvore provem da conveniente representação que pode ser dada a este método (Figura 4.5), onde círculos representam "nós" intermediários e quadrados representam nós terminais.

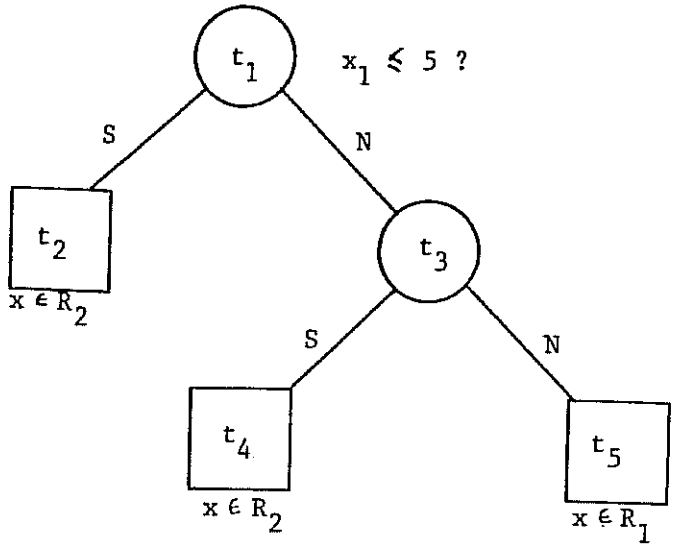


Figura 4.5 - Representação gráfica do método da Árvore.

A classificação de uma nova observação é feita da seguinte maneira: Se $\mathbf{x} \in t_j$ (nó terminal) e a amostra for conjunta * então $\mathbf{x}$ é classificado na população que for predominante em t_j.

Considerando que a amostragem foi conjunta, uma rápida inspeção na Figura 4.4c resulta em:

$$\text{se} \quad \mathbf{x} \in t_5 \quad \text{então,} \quad \mathbf{x} \in R_1$$

$$\text{se} \quad \mathbf{x} \in t_2 \cup t_4 \quad \text{então,} \quad \mathbf{x} \in R_2$$

Assim a construção de uma árvore a partir de uma amostra $\mathbf{D}$ envolve basicamente dois problemas: como particionar e quando parar de particionar.

Na realidade, como veremos adiante, o problema a ser enfrentado não é o "quando parar de particionar" e sim o "quando parar de podar", pois a princípio, as árvores serão

* No caso de amostras separadas a classificação é um pouco diferente e será apresentada no final desta seção.

construídas, intencionalmente, com muitos nós terminais, e então, podadas convenientemente.

Encerramos esta introdução definindo algumas notações e apresentando alguns resultados que serão utilizados neste capítulo:

Para cada nó t sejam $n_i(t)$ o número de observações em t provenientes da população Π_i e $n(t)$ o número total de observações deste nó. Como antes, n_i e n representam o número de observações da população Π_i e o número total de observações, respectivamente.

Algumas estimativas de probabilidades de interpretação imediata são definidas a seguir:

$$\begin{aligned}
\widehat{\Pr}(t\,|\,\Pi_i) &= \widehat{\Pr}(\mathbf{x} \in t\,|\,\mathbf{x} \in \Pi_i) &&= n_i(t)/n_i \\
\widehat{\Pr}(t, \Pi_i) &= \widehat{\Pr}(\mathbf{x} \in t, \mathbf{x} \in \Pi_i) &&= n_i(t)\,p_i/n_i \\
\widehat{\Pr}(t) &= \widehat{\Pr}(\mathbf{x} \in t) &&= \sum_{i=1}^{g} \widehat{\Pr}(t, \Pi_i) \\
\widehat{\Pr}(\Pi_i\,|\,t) &= \widehat{\Pr}(\mathbf{x} \in \Pi_i\,|\,\mathbf{x} \in t) &&= \widehat{\Pr}(\Pi_i, t)/\widehat{\Pr}(t)
\end{aligned}$$

Assim, definimos

$$\widehat{\Pr}(\Pi_i\,|\,t) = \frac{\widehat{\Pr}(t\,|\,\Pi_i)\,\widehat{\Pr}(\Pi_i)}{\sum_{h=1}^{g} \widehat{\Pr}(t|\,\Pi_h)\,\widehat{\Pr}(\Pi_h)} = \frac{n_i(t)\,p_i/n_i}{\sum_{h=1}^{g} n_h(t)\,p_h/n_h}$$

Se a amostragem for conjunta e p_i for desconhecido podemos estimar p_i $(i = 1, \ldots, g)$ por n_i/n e a fórmula acima simplifica-se para:

$$\widehat{\Pr}(\Pi_i\,|\,t) = \frac{n_i(t)}{n(t)}$$

Estas duas últimas fórmulas são essenciais, pois a classificação de uma nova observação baseia-se nelas: Dado $\mathbf{x}$, determino qual o seu correspondente nó terminal t e

$$\text{se} \quad \frac{n_i(t)\,p_i}{n_i} = \max_{h=1,\ldots,g} \frac{n_h(t)\,p_h}{n_h} \qquad \text{então,} \quad \mathbf{x} \in R_i.$$

Observe que há uma certa relação entre esta regra e a dada pelo método dos k-VMP (4−12). No caso de amostras conjuntas, por exemplo, o método dos k-VMP classificava $\mathbf{x}$ na população que fosse predominante entre as k observações mais próximas a $\mathbf{x}$. Já neste método, $\mathbf{x}$ é classificado na população que for predominante no nó terminal t determinado por $\mathbf{x}$.

4.3.1 Regra de Partição

Seja t um nó intermediário qualquer. O primeiro passo para particionar t consiste em determinar o conjunto S formado por todas as possíveis partições s em t. Uma notação mais precisa seria $S(t)$.

O objetivo é determinar qual a partição $s \in S$ que melhor separa as populações segundo um determinado critério. A partição escolhida será denotada por s^* e será chamada de partição ótima.

Cada partição s dependerá apenas de uma variável X_j ($j = 1,\ldots, r$). Suponhamos que X_j seja uma variável quantitativa como no exemplo da seção anterior. Dessa forma, as possíveis partições em t com relação a X_j são especificadas por:

$$t_E = \{\, \mathbf{x} \in t \,|\, x_j \leq c_{jl} \,\} \qquad \text{e} \qquad t_D = \{\, \mathbf{x} \in t \,|\, x_j > c_{jl} \,\}$$

onde c_{jl} ($j = 1,\ldots, r$; $l = 1,\ldots, n-1$) é o ponto médio entre dois pontos amostrais de X_j obtidos após terem sido ordenados.

Se X_j for uma variável qualitativa com os seus resultados representados por elementos do conjunto $\{0, 1, \ldots, c_j - 1\}$, então as possíveis partições são da forma:

$$t_E = \{\mathbf{x} \in t \,|\, x_j \in C\} \qquad \text{e} \qquad t_D = \{\mathbf{x} \in t \,|\, x_j \notin C\}$$

onde C é um subconjunto com k elementos $(k < j)$ de $\{0, 1, \ldots, c_j - 1\}$.

Consideremos o caso em que $\mathbf{X} = (X_1, X_2)'$, X_1 variável aleatória quantitativa e X_2 categórica. Se numa amostra for observado os seguintes valores para X_1: $\{30, 4, 12, 2\}$, então as possíveis partições são:

$$s_1 \quad , \quad s_2 \quad \text{e} \quad s_3$$

onde s_1 é a partição em t que gera

$$t_E = \{\mathbf{x} \in t \,|\, x_1 \leq 3\} \qquad \text{e} \qquad t_D = \{\mathbf{x} \in t \,|\, x_1 > 3\}$$

s_2 é a partição que gera,

$$t_E = \{\mathbf{x} \in t \,|\, x_1 \leq 8\} \qquad \text{e} \qquad t_D = \{\mathbf{x} \in t \,|\, x_1 > 8\}$$

e s_3 gera,

$$t_E = \{\mathbf{x} \in t \,|\, x_1 \leq 21\} \qquad \text{e} \qquad t_D = \{\mathbf{x} \in t \,|\, x_1 > 21\}$$

Se X_2 for uma variável categórica assumindo valores em $\{0, 1, 2\}$ então as possíveis partições são:

$$s_4 \quad , \quad s_5 \quad \text{e} \quad s_6$$

onde s_4 é a partição em t que gera,

$$t_E = \{\, \mathbf{x} \in t \,|\, \mathbf{x} \in \{0\} \,\} \qquad \text{e} \qquad t_D = \{\, \mathbf{x} \in t \,|\, \mathbf{x} \in \{1,\ 2\} \,\}$$

s_5 gera,

$$t_E = \{\, \mathbf{x} \in t \,|\, \mathbf{x} \in \{1\} \,\} \qquad \text{e} \qquad t_D = \{\, \mathbf{x} \in t \,|\, \mathbf{x} \in \{0,\ 2\} \,\}$$

e s_6,

$$t_E = \{\, \mathbf{x} \in t \,|\, \mathbf{x} \in \{2\} \,\} \qquad \text{e} \qquad t_D = \{\, \mathbf{x} \in t \,|\, \mathbf{x} \in \{0,\ 1\} \,\}$$

O conjunto S será formado por $\{\, s_1,\ s_2,\ s_3,\ s_4,\ s_5,\ s_6 \,\}$. Uma vez determinada a partição ótima $s^* \in S$ (veremos adiante como fazer isso), t é particionada em t_E e t_D de acordo com s^* (Figura 4.6) e o processo recomeça considerando as possíveis partições em t_E, depois em t_D e assim sucessivamente até que cada nó terminal seja "puro", isto é, suas observações provenham de uma única população ou tenha poucas observações *.

---

* O número máximo de observações restantes em cada nó terminal não "puro" é um valor pré-fixado (usualmente, em torno de 5). Eventualmente, o nó pode não ser "puro" mas ter todas as observações idênticas; neste caso não há como particionar o nó.

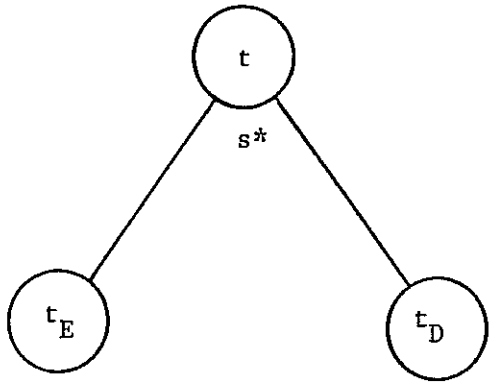


Figura 4.6 - $t = t_E \cup t_D$

Vários critérios foram sugeridos para a definição da partição ótima (s^*) para o nó t. Entretanto, empiricamente, foi constatado (Breiman *et al* (1984) - pág. 38) que a escolha de um ou outro critério influencia pouco na estrutura da árvore. Dessa forma, restringiremos nossa atenção para o critério *Gini*, escolhido por ser o mais viável computacionalmente e por ter interpretações bastante intuitivas.

O critério *Gini* baseia-se na seguinte medida de impureza:

$$i(t) = 1 - \sum_{i=1}^{g} \widehat{\Pr}^2(\Pi_i \,|\, t) \qquad (4-14)$$

Note que esta função assume seu valor máximo quando o nó t for "impuro", ou seja, quando $\widehat{\Pr}(\Pi_i \,|\, t) = 1/g$ $(i = 1, \ldots, g)$ e assume seu valor mínimo quando o nó t for "puro", isto é, $\widehat{\Pr}(\Pi_i \,|\, t) = 1$ para algum i e $\widehat{\Pr}(\Pi_h \,|\, t) = 0$, para todo $h \neq i$. É neste sentido que dizemos que $i(t)$ mede a impureza do nó t.

Interpretações interessantes para $i(t)$ são as seguintes:

Redefina as observações no nó t para 1 se a observação provier de Π_i, e para 0, caso contrário. Se a amostragem for conjuta a variância amostral destas ob-

servações será $\widehat{\Pr}(\Pi_i \,|\, t)\,\left(1-\widehat{\Pr}(\Pi_i \,|\, t)\right)$. Se somarmos para todas as populações teremos a medida de impureza $i(t)$, isto é, $\sum_{i=1}^{g} \widehat{\Pr}(\Pi_i \,|\, t)\,\left(1-\widehat{\Pr}(\Pi_i \,|\, t)\right) = i(t)$. Assim, valores pequenos para $i(t)$ indicam que a somatória das variâncias neste nó é pequena e, portanto, indicam a existência de uma certa homogeneidade nas observações com relação à população.

Outra interpretação baseia-se em regras de classificação: retire aleatoriamente uma observação do nó t e classifique-a na população Π_h com probabilidade $\widehat{\Pr}(\Pi_h \,|\, t)$. A probabilidade estimada de que a observação provenha na verdade de Π_i é $\widehat{\Pr}(\Pi_i \,|\, t)$. Dessa forma a probabilidade estimada de classificação incorreta é:

$$\sum_{i=1}^{g}\left\{\widehat{\Pr}(\Pi_i \,|\, t)\,\left[\sum_{\substack{h=1\\ h\neq i}}^{g}\widehat{\Pr}(\Pi_h \,|\, t)\right]\right\} = \sum_{i=1}^{g}\left\{\widehat{\Pr}(\Pi_i \,|\, t)\,\left[1-\widehat{\Pr}(\Pi_i \,|\, t)\right]\right\}$$

$$= 1-\sum_{i=1}^{g}\widehat{\Pr}^{2}(\Pi_i \,|\, t) \quad = \quad i(t)$$

Utilizando o critério *Gini* para medir a impureza em cada nó t, podemos definir a impureza da árvore T como sendo uma somatória das impurezas dos nós terminais ponderadas pelas suas respectivas probabilidades de ocorrência:

$$\mathrm{I(T)} = \sum_{t\in \mathrm{T}'} i(t)\widehat{\Pr}(t)$$

onde T' é o conjunto dos nós terminais de T.

Suponha agora que particionamos um dado nó terminal t da árvore T, resultando nos novos nós terminais t_E e t_D.

A impureza desta nova árvore ($\mathrm{T_1}$) é dada por:

$$\mathrm{I}(\mathrm{T}_1) = \left[\sum_{t \in \mathrm{T}' - \{t\}} i(t)\, \widehat{\mathrm{Pr}}(t) \right] + i(t_E)\, \widehat{\mathrm{Pr}}(t_E) + i(t_D)\, \widehat{\mathrm{Pr}}(t_D)$$

O decréscimo * provocado na impureza da árvore original é,

$$\mathrm{I}(\mathrm{T}) - \mathrm{I}(\mathrm{T}_1) = i(t)\, \widehat{\mathrm{Pr}}(t) - i(t_E)\, \widehat{\mathrm{Pr}}(t_E) - i(t_D)\, \widehat{\mathrm{Pr}}(t_D) \qquad (4-15)$$

Considere, agora, as proporções q_E e q_D relacionadas com as partições t_E e t_D de t:

$$q_E = \frac{\widehat{\mathrm{Pr}}(t_E)}{\widehat{\mathrm{Pr}}(t)} \qquad \text{e} \qquad q_D = \frac{\widehat{\mathrm{Pr}}(t_D)}{\widehat{\mathrm{Pr}}(t)} \qquad (4-16)$$

No caso de amostras conjuntas estas proporções simplificam-se para

$$q_E = \frac{n(t_E)}{n(t)} \qquad \text{e} \qquad q_D = \frac{n(t_D)}{n(t)}$$

Substituindo $(4-16)$ em $(4-15)$ temos:

$$\mathrm{I}(\mathrm{T}) - \mathrm{I}(\mathrm{T}_1) = [i(t) - i(t_E)\, q_E - i(t_D)\, q_D]\ \widehat{\mathrm{Pr}}(t)$$

Observe que determinar a partição s em t que maximiza o decréscimo na árvore T é equivalente a determinar a partição que maximiza $\Delta\, i(s,t)$ dado a seguir:

$$\Delta\, i(s,t) = i(t) - i(t_E)\, q_E - i(t_D)\, q_D$$

uma vez que $\widehat{\mathrm{Pr}}(t)$ independe da partição escolhida.

Assim, definimos: A melhor partição s^* de um nó t em t_E e t_D é a que maximiza

* A demonstração de que, realmente, se trata de um decréscimo pode se vista em Breiman *et al* (1984) - páginas 126 e 127.

$$\Delta\, i(s,t) = i(t) - i(t_E)\, q_E - i(t_D)\, q_D \qquad (4-17)$$

ou seja, a partição ótima é a que maximiza o decréscimo na impureza da árvore original.

Uma regra de parada bastante intuitiva mas não recomendável seria fixar uma constante β e declarar que o nó é terminal se

$$\Delta\, i(s^*,t)\, \widehat{\Pr}(t) < \beta$$

Este procedimento é insatisfatório uma vez que, não raramente, a partição ótima num nó t provoca um decréscimo insignificante mas as partições geradas a partir deste nó: t_E e t_D provocam grande decréscimo na impureza total da árvore. Isto acontece de maneira sensível quando ocorre uma "reversão" conforme ilustra a Figura 4.7a. A primeira partição implica em $\Delta\, i(s^*,t_1)$ insignificante (Figura 4.7b). No entanto novas partições geram nós terminais "puros" (Figura 4.7c).

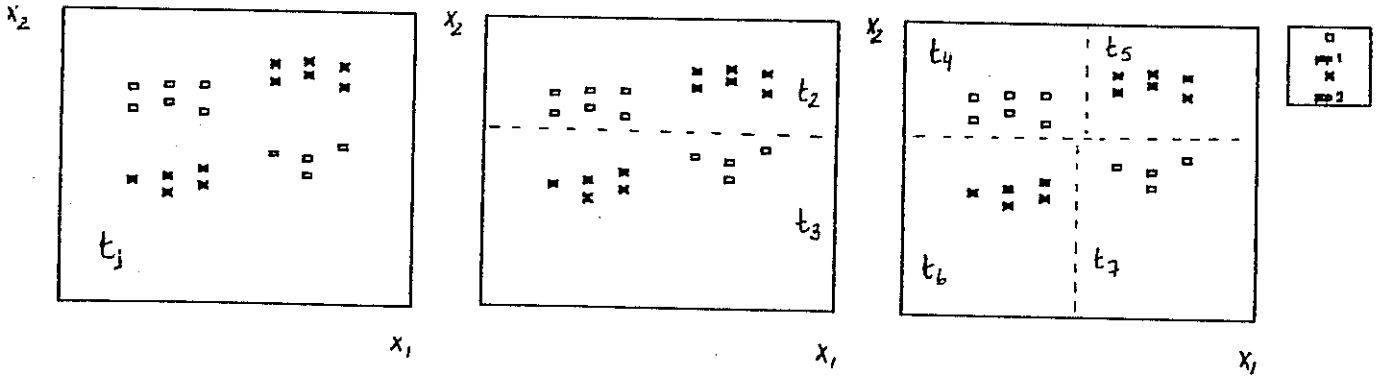


Figura 4.7 - Dados com "reversão". (a) espaço amostral $\Omega = t_1$; (b) partição de t_1 em t_2 e t_3; (c) partição de t_2 em t_4 e t_5 e de t_3 em t_6 e t_7 gerando nós terminais "puros".

4.3.2 O Tamanho Ótimo da Árvore

Na seção anterior vimos como particionar o espaço amostral Ω, denominado nó inicial t_1, até obtermos uma árvore que não possa mais ser particionada, que vamos denotar por T_{max}.

Em geral, T_{max} não gera uma boa regra de classificação apesar da taxa de erro aparente ser muito baixa. Isto decorre do fato desta árvore ser muito dependente da particular amostra utilizada.

O objetivo desta seção é apresentar uma forma de construir uma seqüência de árvores T_{max}, $T_1, \ldots, t_1$, onde T_1 é obtido através de convenientes "podas" de T_{max}, T_2 é obtido podando-se T_1, e assim sucessivamente até que reste apenas o nó inicial t_1. A seguir deve-se escolher entre as árvores dessa seqüência a árvore que minimizar uma estimativa da taxa de erro real. Esta árvore será denotada por T^*.

Em outras palavras o que ocorre é que as partições efetuadas a partir de t_1 geram árvores com taxas de erro aparente cada vez menores. Entretanto, para estas mesmas árvores, a taxa de erro real diminui até uma árvore ótima (T^*) voltando a crescer a partir daí. Ao contrário da taxa de erro aparente, a minimização da taxa de erro da amostra teste consegue determinar, com relativa precisão, qual é a árvore T^*.

O primeiro passo é determinar a seqüência de árvores: T_{max}, $T_1, \ldots, t_1$ candidatas a T^* utilizando-se de critérios baseados na taxa de erro aparente. O segundo passso é escolher a árvore (T^*) que resultar na menor estimativa para a taxa de erro real. Usualmente, utiliza-se a taxa de erro da amostra teste.

Antes, porém, vamos definir o conceito de sub-árvore: Uma sub-árvore T_t de uma árvore T consiste no nó t e todos as ramificações de t em T. As Figuras 4.8a e b ilustram o caso.

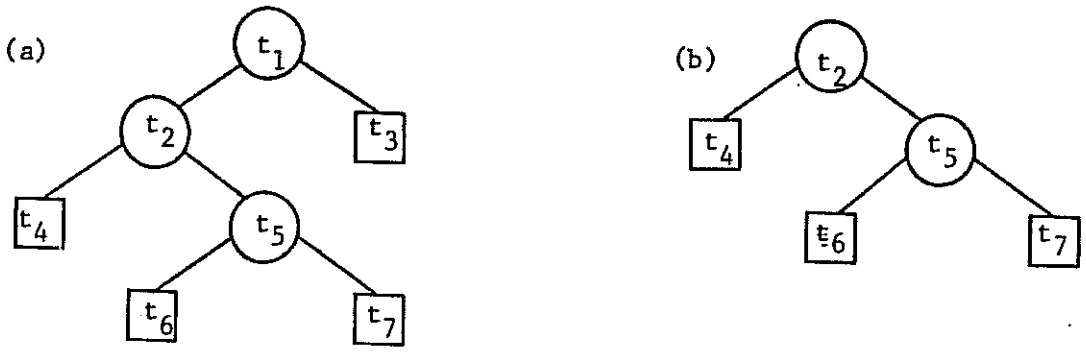


Figura 4.8 - (a) árvore T; (b) sub-árvore T_{t_2}.

Para o primeiro passo, precisamos determinar as expressões para a taxa de erro aparente de um nó (t), de uma árvore (T) e de uma subárvore (T_t). Lembrando que a regra de classificação para um nó t é feita classificando-se todas as observações deste nó na população cuja probabilidade *a posteriori* estimada $\widehat{\Pr}(\Pi_i \mid t)$ for maior $(i = 1, \ldots, g)$ deduz-se que a taxa de erro aparente no nó t é dada por:

$$e_{ap}(t) = 1 - \max_{i=1,\ldots,g} \widehat{\Pr}(\Pi_i \mid t)$$

Conseqüentemente, a taxa de erro aparente de uma árvore T será:

$$e_{ap}(T) = \sum_{t \in T'} e_{ap}(t)\, \widehat{\Pr}(t)$$

onde T' é o conjunto de nós terminais de T.

Finalmente, definimos a taxa de erro aparente devido à sub-árvore T_t de maneira análoga:

$$e_{ap}(T_t) = \sum_{t' \in T'_t} e_{ap}(t')\, \widehat{\Pr}(t')$$

onde T'_t é o conjunto dos nós terminais de T_t.

Partindo-se de T_{max} o objetivo é definir como as "podas" devem ser feitas até se chegar ao nó inicial t_1. Este critério de "poda" está baseado numa medida que pondera a taxa de erro aparente $(e_{ap}(T))$ e o número de nós terminais da árvore T, que será denotado por $n(T)$.

Para qualquer nó, sub-árvore ou árvore T esta medida é dada por

$$R_\alpha(T) = e_{ap}(T) + \alpha\, n(T) \qquad (4-18)$$

onde $\alpha \geq 0, \alpha \in I\!R$ é denominado parâmetro de complexidade, pois, diz-se que uma sub-árvore é tanto mais complexa quanto mais nós ela tiver.

O critério para a escolha das árvores "candidatas" a T^* consiste em, dado α, escolher a sub-árvore T de T_{max} que minimizar $R_\alpha(T)$. Esta sub-árvore será denotada por $T(\alpha)$. Observe que $T(0) = T_{max}$ e para α suficientemente grande $T(\alpha) = t_1$.

A determinação, para cada α, de $T(\alpha)$, considerando todas as possíveis sub-árvores é inviável computacionalmente. Pode-se provar (Breiman *et al* (1984) - páginas 67 e 68) que o procedimento descrito abaixo também leva a determinação de um conjunto de árvores que inclui T^*, porém, com muito mais rapidez.

Neste procedimento o objetivo é construir uma seqüência de árvores $T_0 = T_{max}$, $T_1, T_2, \ldots, T_N = t_1$ candidatas a T^*. O procedimento consiste em:

(i) Seja $i = 0$ e $T_0 = T_{max}$;

(ii) Calcule $\alpha_{i+1}(t) = [e_{ap}(t) - e_{ap}(T_t)] \,/\, [n(T_t) - 1]$ para cada nó não terminal t de T_i;

(iii) Pode * o nó que resultar no menor valor para $\alpha_{i+1}(t)$. Denote a árvore resultante de $T_{i+1}(\alpha)$;

---

* podar um nó t significa retirar da árvore todas as partições feitas a partir de t transformando-o num nó terminal.

(iv) Incremente o índice i em uma unidade e retorne ao passo (ii) até que T_i coincida com o nó inicial t_1.

A idéia do procedimento é a seguinte: Para cada nó não terminal t é possível, a partir de $(4-18)$, calcular a sua medida de complexidade $R_\alpha(t)$ e sua correspondente sub-árvore T_t cuja notação é $R_\alpha(T_t)$:

$$\begin{aligned} R_\alpha(t) &= e_{ap}(t) + \alpha \\ \text{e} \qquad R_\alpha(T_t) &= e_{ap}(T_t) + \alpha\, n(T_t) \end{aligned}$$

onde $n(T_t)$ é o número de nós terminais da sub-árvore T_t.

Quando estas duas medidas forem iguais significa que as partições efetuadas a partir de t são desprezíveis, isto é, a árvore com nó terminal t é menos complexa do que a que contém a sub-árvore T_t e deve ser preferida, apesar do incremento na taxa de erro aparente. Essa igualdade ocorre quando:

$$R_\alpha(t) = R_\alpha(T_t) \quad \Rightarrow \quad \alpha = \frac{e_{ap}(t) - e_{ap}(T_t)}{n(T_t) - 1}$$

Assim, para cada nó não terminal t podemos calcular qual deve ser o valor de α para que a igualdade ocorra.

A seguir devemos podar a sub-árvore T_t correspondente ao menor valor de α. Observe que ao fazermos isto estaremos podando a sub-árvore que mantém a "ligação mais fraca" com a árvore. Isto é, a sub-árvore a ser podada é aquela que dá o menor peso para o número de nós terminais a fim de que a igualdade das medidas de complexidade ocorra.

Uma vez obtidas as candidatas a T^* : T_1, $T_2, \ldots,$ t_1 escolhe-se a árvore que minimizar uma estimativa da taxa de erro real. Por motivos óbvios a taxa de erro aparente não pode ser usada pois escolheria sempre T_{max}. Se amostras suficientemente grandes estiverem disponíveis a melhor solução é utilizar a taxa de erro da amostra teste. Caso

contrário, pode-se utilizar a taxa de erro da validação cruzada a partir de técnicas específicas (Breiman *et al* (1984) - páginas 75 a 78).

4.3.3 Exemplo

Uma loja quer se utilizar da mala-direta para vender seus produtos. Por questão de economia esta loja quer enviar catálogos apenas para compradores em potencial. Através da análise de "malas diretas" enviadas anteriormente, sabe-se que existem três tipos de populações:

Π_1 : clientes que compram pelo catálogo;

Π_2 : clientes que solicitam mais informações;

Π_3 : clientes que não compram.

O objetivo é determinar qual o perfil dos clientes da população Π_1, com base, exclusivamente, nas seguintes variáveis:

$$X_1 = \text{renda do cliente;}$$

$$X_2 = \begin{cases} 1, & \text{nunca comprou pelo catálogo;} \\ 0, & \text{já comprou pelo catálogo.} \end{cases}$$

$$X_3 = \begin{cases} 1, & \text{não tem filhos;} \\ 0, & \text{tem filhos.} \end{cases}$$

Para isto, dados foram coletados de pesquisas anteriores e estão na Tabela 4.8.

k	x_1	x_2	x_3	Π
1	0,5	0	1	Π_2
2	4,0	0	1	Π_2
3	20,0	1	1	Π_1
4	3,0	1	1	Π_2
5	5,0	1	0	Π_1
6	1,0	0	0	Π_2
7	10,0	0	0	Π_3
8	8,0	0	0	Π_3
9	7,0	1	0	Π_1
10	15,0	1	1	Π_1
11	6,0	0	1	Π_2
12	8,0	1	1	Π_2

Tabela 4.8 - Dados hipotéticos.

Inicialmente temos $t_1 = \Omega$. Para o nó inicial (raiz) a impureza foi calculada conforme (4–14). Admitindo que a amostragem é conjunta:

$$i(t_1) = 1 - \sum_{i=1}^{3} \widehat{\Pr}^2 (\Pi_i \,|\, t_1) = 1 - \left[\left(\frac{4}{12}\right)^2 + \left(\frac{6}{12}\right)^2 + \left(\frac{2}{12}\right)^2 \right] = 0,611$$

A primeira partição $s \in S$ deve dividir t_1 em t_2 e t_3 nas proporções q_E e q_D respectivamente de tal forma que $\Delta\, i(s, t_1) = i(t_1) - q_E\, i(t_2) - q_D\, i(t_3)$ seja máxima.

Por exemplo, a partição

$$s' : \; t_2 = \{ \mathbf{x} \in t_1 \,|\, x_2 = 0 \}, t_3 = \{ \mathbf{x} \in t_1 \,|\, x_2 = 1 \} \quad \text{resulta em,}$$

$$\begin{aligned} t_2 &= \{1,2,6,7,8,11\} \;\Rightarrow\; q_2 = 6/12 \\ t_3 &= \{3,4,5,9,10,12\} \;\Rightarrow\; q_3 = 6/12 \end{aligned}$$

$$\left.\begin{array}{lcl} \Pr(\Pi_1 \,|\, t_2) & = & 0 \\ \Pr(\Pi_2 \,|\, t_2) & = & 4/6 \quad (k = 1,2,6 \text{ e } 11) \\ \Pr(\Pi_3 \,|\, t_2) & = & 2/6 \quad (k = 7 \text{ e } 8) \end{array}\right\} \; i(t_2) = 1 - \left[\left(\frac{4}{6}\right)^2 + \left(\frac{2}{6}\right)^2\right] = 0,444$$

$$\left.\begin{array}{lcl} \Pr(\Pi_1 \,|\, t_3) & = & 4/6 \quad (k = 3,5,9 \text{ e } 10) \\ \Pr(\Pi_2 \,|\, t_3) & = & 2/6 \quad (k = 4 \text{ e } 12) \\ \Pr(\Pi_3 \,|\, t_3) & = & 0 \end{array}\right\} \; i(t_3) = 1 - \left[\left(\frac{4}{6}\right)^2 + \left(\frac{2}{6}\right)^2\right] = 0,444$$

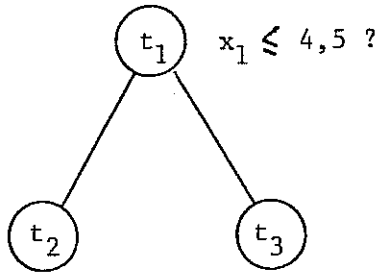
logo, $\Delta\, i(s', t_1) = 0,611 - 0,444 \cdot 6/12 - 0,444 \cdot 6/12 = 0,167$

O decréscimo na impureza $\Delta\, i(s, t_1)$ foi calculado para todas as partições $s \in S$ (os resultados estão na Tabela 4.9).

Partição	candidatos a t_2	q_2	q_3	$i(t_2)$	$i(t_3)$	$\Delta\, i(s, t_1)$
s_1	$\{x_2 = 0\}$	0,500	0,500	0,444	0,444	0,167
s_2	$\{x_3 = 0\}$	0,417	0,583	0,640	0,408	0,106
s_3	$\{x_1 \leq 0,75\}$	0,083	0,917	0,000	0,628	0,035
s_4	$\{x_1 \leq 2,00\}$	0,167	0,833	0,000	0,640	0,078
s_5	$\{x_1 \leq 3,50\}$	0,250	0,750	0,000	0,642	0,130
s_6	$\{x_1 \leq 4,50\}$	0,333	0,667	0,000	0,625	0,194
s_7	$\{x_1 \leq 5,50\}$	0,417	0,583	0,320	0,653	0,097
s_8	$\{x_1 \leq 6,50\}$	0,500	0,500	0,278	0,611	0,167
s_9	$\{x_1 \leq 7,50\}$	0,583	0,417	0,408	0,640	0,106
s_{10}	$\{x_1 \leq 8,50\}$	0,750	0,250	0,494	0,444	0,130
s_{11}	$\{x_1 \leq 12,50\}$	0,833	0,167	0,560	0,000	0,144
s_{12}	$\{x_1 \leq 17,50\}$	0,917	0,083	0,596	0,000	0,066

Tabela 4.9 - Partições possíveis em t_1. Para cada partição, $t_3 = t_1 - (t_1 \cap t_2)$.

O maior decréscimo em $\Delta_i(s, t_1)$ ocorre na partição s_6, dessa forma temos:



O conjunto t_2 é puro: todas as observações pertencem à população Π_2, logo t_2 é um nó terminal. O conjunto t_3 formado pelas observações $3, 5, 7$ a 12 não é puro e pode ser particionado em t_4 e t_5.

A função de impureza $i(t_3)$ é dada por

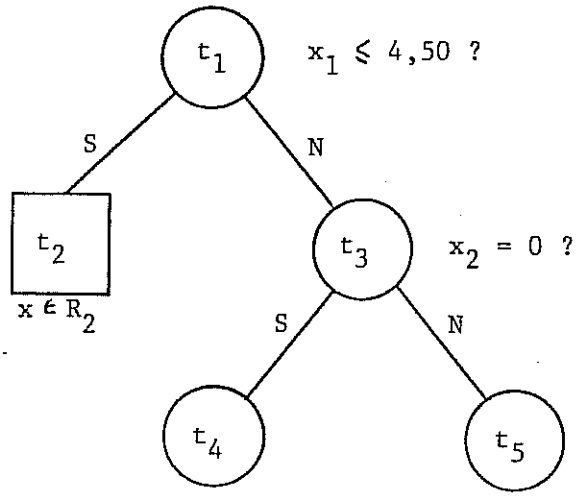
$$i(t_3) = 1 - \sum_{i=1}^{3} \Pr^2(\,\Pi_i \,|\, t_3\,) = 1 - \left(\frac{4}{8}\right)^2 - \left(\frac{2}{8}\right)^2 - \left(\frac{2}{8}\right)^2 = 0,625$$

Todas as partições possíveis em t_3 e seus respectivos resultados estão listados na Tabela 4.10.

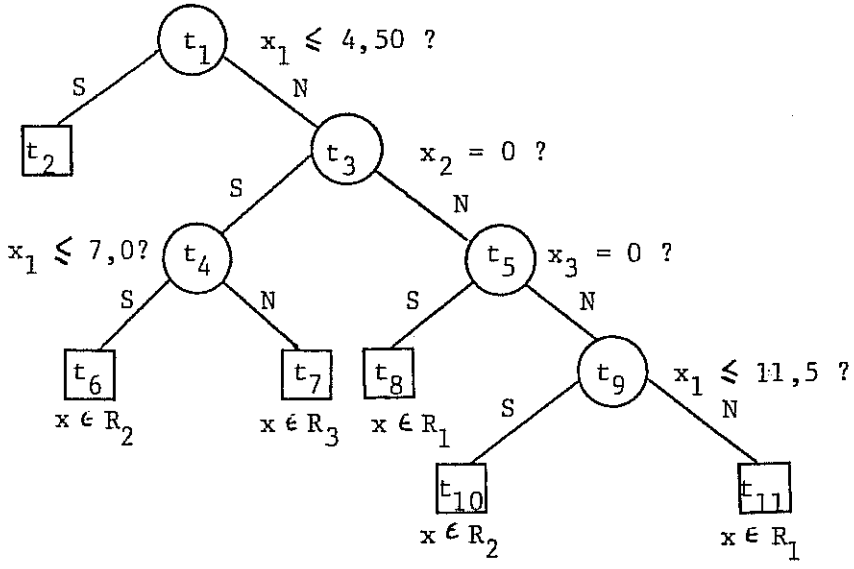
Partição	candidatos a t_4	q_4	q_5	$i(t_4)$	$i(t_5)$	$\Delta_i(s, t_3)$
s_{13}	$\{x_2 = 0\}$	0,375	0,625	0,444	0,320	0,258
s_{14}	$\{x_3 = 0\}$	0,500	0,500	0,500	0,500	0,125
s_{15}	$\{x_1 < 5,5\}$	0,125	0,875	0,000	0,653	0,054
s_{16}	$\{x_1 < 6,5\}$	0,250	0,750	0,500	0,611	0,042
s_{17}	$\{x_1 < 7,5\}$	0,375	0,625	0,444	0,640	0,058
s_{18}	$\{x_1 < 8,5\}$	0,625	0,375	0,640	0,444	0,058
s_{19}	$\{x_1 < 12,5\}$	0,750	0,250	0,667	0,000	0,125
s_{20}	$\{x_1 < 17,5\}$	0,833	0,167	0,640	0,000	0,225

Tabela 4.10 - Partições possíveis em t_3. Para cada partição, $t_5 = t_3 - (t_3 \cap t_4)$.

A melhor partição foi s_{13}. A árvore torna-se



Continuando as partições até que o erro aparente seja nulo, obtivemos a seguinte árvore T_{max}:

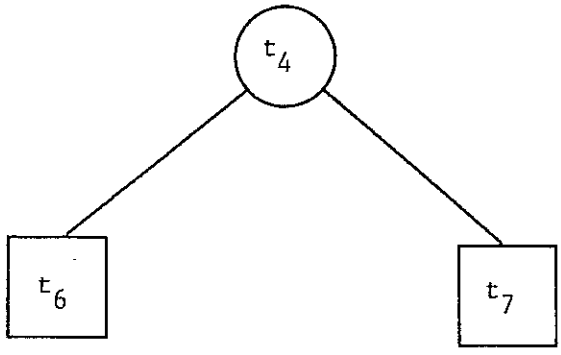


O próximo passo consiste em obter uma seqüência de árvores $T_{max}, T_1, \ldots, \{t_1\}$ e escolher uma delas através, por exemplo, da minimização da taxa de erro da amostra teste.

A determinação de T_1 a partir de T_{max} é feita excluindo-se as partições de t (nó intermediário) correspondente ao valor mínimo de

$$\alpha_1(t) = \frac{e_{ap}(t) - e_{ap}(T_t)}{n(T_t) - 1}$$

Vamos calcular, com detalhes, a função $\alpha_1(t_4)$. O nó t_4 tem apenas uma partição:



Logo, $e_{ap}(T_{t_4}) = e_{ap}(t_6) + e_{ap}(t_7) = 0$, pois t_6 e t_7 são puros. Claramente, $n(T_{t_4}) = 2$. Assim,

$$\alpha_1(t_4) = \frac{\frac{1}{3} \cdot \frac{3}{12} - 0}{2 - 1} = 0,083$$

Os valores das demais funções $\alpha_1(\cdot)$ foram calculados analogamente, resultando em:

$$\begin{aligned} \alpha_1(t_1) &= 0,100 \\ \alpha_1(t_3) &= 0,083 \\ \alpha_1(t_5) &= 0,042 \\ \alpha_1(t_9) &= 0,083 \end{aligned}$$

O valor mínimo de α_1 é obtido com o nó t_5. Logo, a árvore T_2 tem estrutura dada pela Figura 4.9.

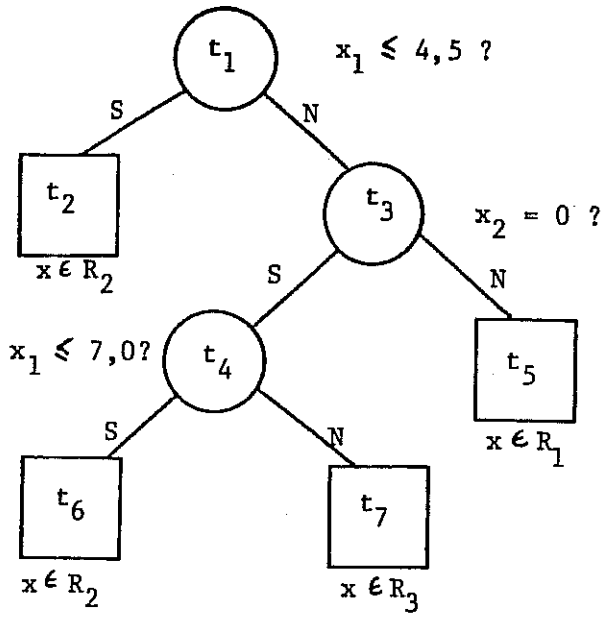


Figura 4.9 - Árvore T_2.

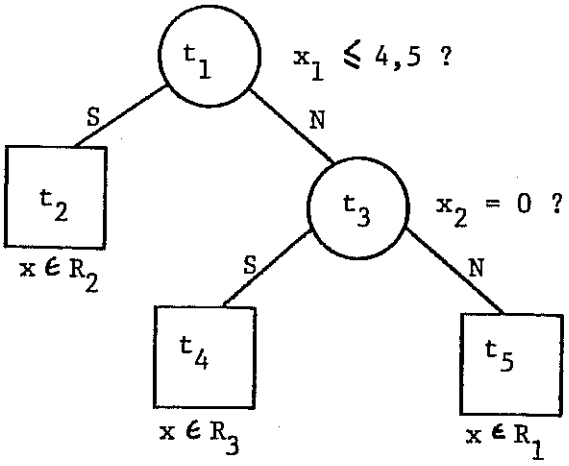
Repetindo o procedimento para todos os nós intermediários de T_2 obtemos,

$$\alpha_2(t_1) = 0,100$$

$$\alpha_2(t_3) = 0,125$$

$$\alpha_2(t_4) = 0,083$$

Assim, a próxima poda a ser feita é em t_4. A árvore T_3 fica



Finalmente, para os nós intermediários de T_3, t_1 e t_3, obtivemos:

$$\alpha_3(t_1) = 0,208$$

$$\alpha_3(t_3) = 0,250$$

Como a poda deve ser feita em t_1, o procedimento pára e as árvores selecionadas são: T_1, T_2, T_3 e t_1.

Usualmente a escolha da melhor árvore é feita através de uma amostra teste: escolhe-se a árvore que minimizar e_{at}. Neste caso, devido ao pequeno tamanho da amostra escolhemos a melhor árvore como sendo T_2 (Figura 4.9) de maneira puramente descritiva.

Conclui-se que a população Π_1 será atingida se forem enviados catálogos para clientes com renda superior a 4,5 salários mínimos, que já compraram pelo catálogo e que não tenham filhos ou para cliente com renda superior a 11,5 salários mínimos, que já compraram pelo catálogo e que tenham filhos.

Capítulo 5

COMPARAÇÕES DOS MÉTODOS E UM EXEMPLO DE APLICAÇÃO

Neste capítulo procuramos comparar os diversos métodos descritos anteriormente analisando alguns trabalhos publicados nesta área e um exemplo real.

Na Seção 5.1 apresentaremos, de forma concisa, os trabalhos mais importantes encontrados na literatura referentes à comparação de métodos com variáveis misturadas. Ainda nesta seção, inclímos algumas conclusões e comentários gerais. Já na Seção 5.2 é apresentado um exemplo de aplicação da Análise Discriminante denominado, na área financeira, de *credit scoring*.

5.1 Comparação dos Métodos

Acreditamos que não existe um método que seja sempre melhor que os demais. O desempenho de cada método está intimamente relacionado à estrutura de dados de cada problema. Os poucos trabalhos envolvendo comparações em Análise Discriminante com variáveis misturadas são muitas vezes restritos a dois métodos apenas. Um resumo de alguns trabalhos é apresentado a seguir.

Titterington *et al* (1981) compararam, através de um exemplo prático, os métodos Núcleo (para variáveis categóricas), LDF, QDF e Discriminação Logística e concluíram que o desempenho varia muito mais efetuando-se uma seleção de variáveis do que, especificamente, com o método utilizado. Já Vlachonikolis e Marriott (1982) e Knoke (1982) concluíram, a partir de simulações, que a inclusão de termos apropriados (como por exemplo, interações) no modelo dos métodos LDF ou Discriminação Logística podem transformá-los em métodos com desempenhos comparáveis ao de Krzanowki porém, com muito menos parâmetros a serem estimados.

Alguns trabalhos basearam-se na comparação de determinado método com a LDF. Entre eles temos:

(i) Murphy e Moran (1986) advogam que o desempenho do método do Núcleo é sempre superior à LDF na presença de variáveis com distribuição normal pouco correlacionadas. Este resultado é de certa forma surpreendente;

(ii) Krzanowski (1977) comparou o método que neste trabalho leva o seu nome com o método LDF. Verificou, através de simulações, que o método de Krzanowski sempre tem desempenho pelo menos igual à LDF, mostrando-se bastante superior nos casos de reversão (Figura 2.1);

(iii) Press e Wilson (1978) apresentaram exemplos em que a Discriminação Logística teve desempenho superior à LDF.

As conclusões da maioria destes trabalhos foram baseados em simulações onde se sabia, *a priori*, quais as interações envolvidas ou os parâmetros de cada distribuição. Na prática, estas interações não são facilmente detectáveis e os parâmetros necessitam ser estimados. Para uma estimação confiável deste parâmetros é necessário levarmos em conta, entre outros fatores, o tamanho da amostra e a presença ou não de *outliers*.

A partir destas e de outras considerações, elaboramos a Tabela 5.1, onde de forma sucinta estão as nossas opiniões acerca da utilização dos métodos contemplados neste trabalho. Procuramos incluir apenas informações gerais passíveis de comparação.

Método	Requer amostras grandes ?	Interações	A regra de classificação é simples ?	Tempo de processamento dispendido	Outros fatores que prejudicam o desempenho do método
LDF	não	- podem ser incorporadas ao modelo	sim	baixo	- variáveis com distribuição não normal; - matrizes de dispersão diferentes em cada população;
QDF	não	- podem ser incorporadas ao modelo	não	baixo	- variáveis com distribuição não normal;
LOGÍSTICO	não	- podem ser incorporadas ao modelo	sim	médio	- suposição (3-1) falsa: - configurações de dados que levam à divergência do algoritmo:
KRZANOWSKI	em termos	- permite entre variáveis categóricas e contínuas; - podem ser incorporadas entre variáveis contínuas;	em termos	baixo	- distribuição dos dados condicionada em cada casela não normal ou com matriz de dispersão não constante:
NÚCLEO	não	- permite	não	elevado	- variáveis discriminadoras correlacionadas;
K-VMP	não	- permite	não	elevado	- medida de distância adotada não tem se mostrado eficaz;
ÁRVORE BINÁRIA	sim	- permite	sim	muito elevado	- amostras pequenas

Tabela 5.1 - Características dos métodos.

5.2 Um Exemplo de Aplicação: *Credit Scoring*

Um problema importante na área financeira é prever se o solicitante de um empréstimo irá ou não honrar esse compromisso. Quando essa previsão é feita com base em uma função que pondera as variáveis cadastrais e financeiras do solicitante, o problema é denominado *credit scoring*.

Esta é uma das aplicações da Análise Discriminante mais difíceis de serem implementadas, pois a amostra disponível, geralmente, é composta por supostos bons pagadores, isto é, clientes que foram considerados pelo gerente de crédito como bons pagadores e, portanto, merecedores de crédito. A população de maus clientes para o modelo *credit scoring* é criada, basicamente, a partir dos erros do gerente, ou seja, clientes que obtiveram o crédito e não o honraram. A informação contida nas propostas que foram indeferidas de imediato pelo gerente raramente são registradas e, portanto, torna-se impossível obter informações desta população. A partir destas considerações recomendamos que esse tipo de modelo seja utilizado para auxiliar a decisão gerencial e não substituí-la.

Um termo comumente utilizado nesta área é "escore do cliente". Em geral, esse escore é a probabilidade *a posteriori** do cliente ser um bom pagador dado os valores de suas variáveis discriminadoras. Valores altos para o escore indicam que o cliente é de baixo risco, e portanto, deve receber o crédito.

A utilização do *credit scoring* no processo decisório varia conforme a empresa e a confiabilidade do modelo. Na Figura 5.1 apresentamos um fluxo simplificado de sua utilização.

---

* ou uma função desta probabilidade; por exemplo: o valor da função discriminante de Fisher calculada em $\mathbf{x}$.

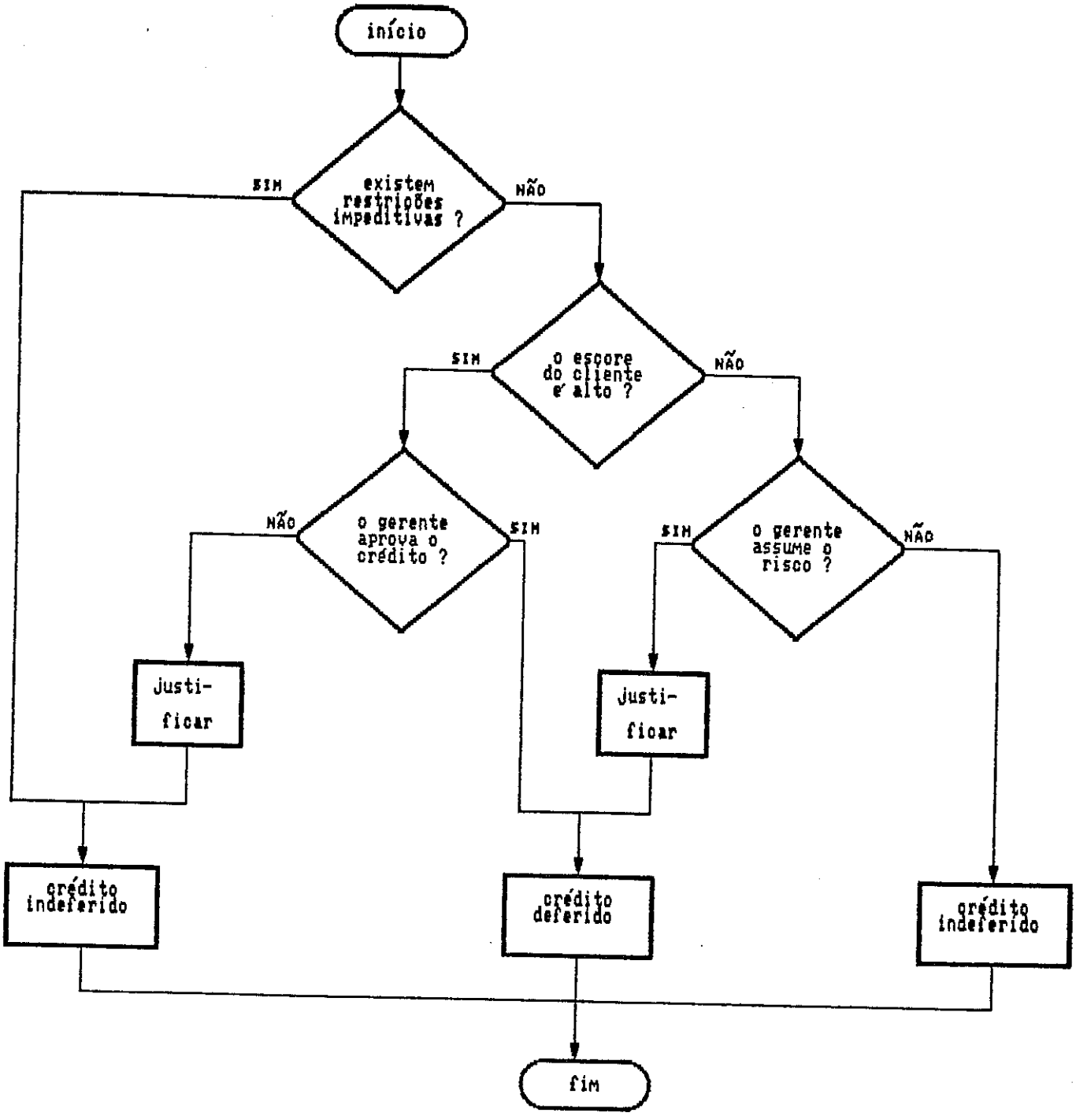


Figura 5.1 - Exemplo de utilização do *credit scoring* no processo de concessão de crédito.

O ponto de corte, que separa escores considerados altos (bons pagadores - pop 1) de escores considerados baixos (maus pagadores - pop2), é escolhido considerando-se aspectos econômico-administrativos. A Figura 5.2 ilustra dois casos. O ponto de corte A é escolhido, por exemplo, quando há forte concorrência no setor e a administração não quer perder sua clientela de bons pagadores mesmo que a qualidade de sua clientela, em geral, piore. Já o ponto de corte B é utilizado em épocas de contenção de créditos ou para se atingir a elite de uma população de bons pagadores.

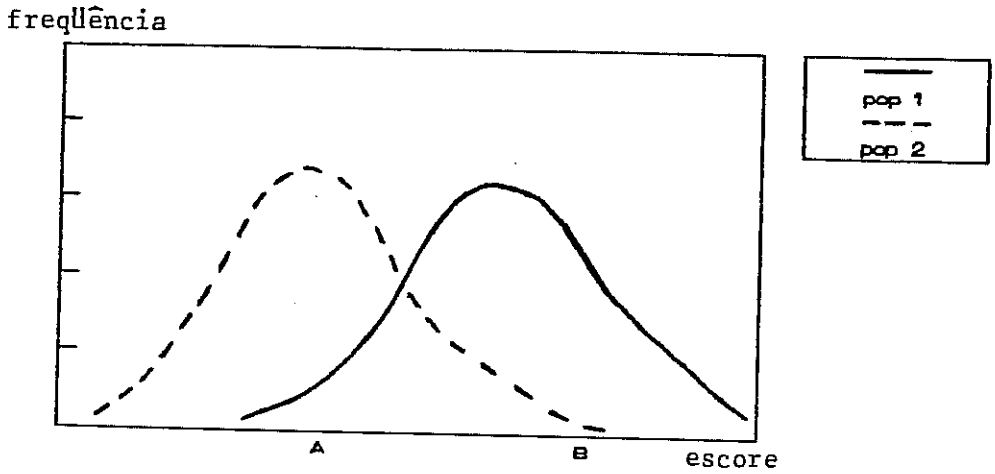


Figura 5.2 - Distribuição dos escores.

Uma forma conservadora de incluir o *credit scoring* num processo de decisão é definir uma região de dúvida onde os escores não são nem "muito altos" nem "muito baixos". Nesta região, análises mais específicas seriam realizadas, prevalecendo as conclusões destas. Um exemplo para o caso da Figura 5.2 seria:

escore $< A$ $\rightarrow$ escore baixo $\rightarrow$ crédito indeferido automaticamente;

$A \leq$ escore $\leq B$ $\rightarrow$ escore médio $\rightarrow$ crédito sujeito a outras análises;

escore $> B$ $\rightarrow$ escore alto $\rightarrow$ crédito deferido automaticamente.

Outro problema de grande importância refere-se à escolha das variáveis a serem uti-

lizadas no *credit scoring*. Apenas como ilustração relacionamos algumas variáveis que, geralmente, possuem poder discriminatório em problemas ligados a áreas financeiras:

pessoas físicas - número de dependentes, renda, tempo no emprego e número de cartões de crédito.

pessoas jurídicas - porte da empresa, setor de atividade, índices contábeis, tradição, restrições e experiência no ramo.

Possíveis variáveis discriminadoras em problemas ligados a pessoas jurídicas são citados em Silva (1988).

Nas seções seguintes veremos um exemplo de aplicação em *credit scoring*.

5.2.1 Descrição dos Dados e Critérios Utilizados

Os dados consistem de observações extraídas de 249 clientes, sendo 199 da população Π_1, os que honraram os créditos e 50 da população Π_2, os que não honraram.

O significado de cada variável foi omitido a pedido do agente financeiro que forneceu os dados. A classificação das variáveis consideradas é dado a seguir:

X_1 – qualitativa nominal (binária)

X_2 – qualitativa nominal (binária)

X_3 – qualitativa ordinal (4 categorias diferentes)

X_4 – quantitativa contínua

X_5 – quantitativa contínua

X_6 – quantitativa contínua

X_7 – quantitativa contínua

Um estudo descritivo de cada variável é apresentado no Apêndice A.

Os dados foram obtidos de amostragens separadas. Sabe-se que, historicamente, a taxa de inadimplência é da ordem de 10% neste tipo de operação de crédito. Deste modo utilizamos $p_1 = 0,90$ e $p_2 = 0,10$.

Em Finanças, o custo de classificação incorreta definido na Seção 1.2 assume grande importância. Neste exemplo trabalhamos com duas possíveis situações:

situação A : $c(1\,|\,2) = c(2\,|\,1)$ - custos iguais*

situação B : $c(1\,|\,2) = 9\,c(2\,|\,1)$

Para a situação B, o "custo" de se conceder o crédito a um mau pagador é nove vezes o "custo " de se negar a um bom pagador.

Os dados foram submetidos a cada um dos métodos vistos anteriormente, com exceção do método da Árvore Binária de Classificação cujo *software* não nos foi possível desenvolver devido a sua complexidade.

Para adequar os dados ao uso dos métodos LDF, QDF e Discriminação Logística, efetuou-se uma reparametrização na variável X_3 da seguinte forma:

X_3	$X_3^{(1)}$	$X_3^{(2)}$
0 ou 1	1	0
2	0	1
3	0	0

* Gera a regra ótima de classificação definida na Seção 1.3.1.

As categorias respresentadas por 0 e 1 de X_3 foram consideradas conjuntamente em virtude de não haver observações de Π_2 na categoria representada por 0. Caso não fossem consideradas dessa forma não seria possível calcular a inversa da estimativa da matriz de variâncias e covariâncias de Π_2 utilizada no método QDF, e na Discriminação Logística o algoritmo iterativo não convergiria. Com o intuito de uniformizar os dados utilizados o procedimento acima foi utilizado, também, para o método LDF.

Para o método do Núcleo foram considerados dois procedimentos: o primeiro trata a variável X_3 como qualitativa nominal utilizando a função Núcleo definida em $(4-6)$ e o segundo utiliza uma função Núcleo específica para variáveis qualitativas ordinais sugerida na Seção 4.1.3.

Dessa forma, os métodos considerados foram:

LDF	–	Função Discriminante Linear
QDF	–	Função Discriminante Quadrática
KRZANOWSKI	–	Método de Krzanowski
LOGÍSTICO	–	Discriminação Logística
NÚCLEO 1	–	Método do Núcleo (X_3 como qualitativa nominal)
NÚCLEO 2	–	Método do Núcleo (X_3 como qualitativa ordinal)
K-VMP 3	–	Método dos k-Vizinhos Mais Próximos (k= 3)
K-VMP 5	–	Método dos k-Vizinhos Mais Próximos (k= 5)
K-VMP 7	–	Método dos k-Vizinhos Mais Próximos (k= 7)

5.2.2 Resultados

Os resultados obtidos são apresentados nas Tabelas 5.2 e 5.3 para as situações A e B, respectivamente.

É interessante verificar que a regra de classificação é de pouca praticidade no caso de custos iguais (Tabela 5.2). Observe que todos os métodos classificaram as observações predominantemente na população Π_1, ou seja, o modelo concede o crédito a quase todos os clientes. Este fato deve-se, principalmente, à grande diferença entre as probabilidades *a priori*.

Método	Observações de Π_2 classificadas em Π_1	Observações de Π_1 classificadas em Π_2	e_{ap}
LDF	50 (100,00%)	0 (0,00%)	10,00%
QDF	39 (78,00%)	6 (3,02%)	10,51%
KRZANOWSKI	49 (98,00%)	0 (0,00%)	9,80%
LOGÍSTICO	50 (100,00%)	0 (0,00%)	10,00%
NÚCLEO 1	43 (86,00%)	9 (4,52%)	12,67%
NÚCLEO 2	23 (46,00%)	5 (2,51%)	6,86%
K-VMP 3	58 (96,00%)	2 (1,01%)	10,50%
K-VMP 5	50 (100,00%)	1 (0,50%)	10,45%
K-VMP 7	50 (100,00%)	2 (1,01%)	10,90%

Tabela 5.2 - Situação A: Custos iguais.

Os resultados diferiram de maneira acentuada na situação B (Tabela 5.3). Os métodos LDF, KRZANOWSKI e LOGÍSTICO tiveram desempenhos similares gerando taxas de erro aparente em torno de 33% com leve tendência a classificar em Π_2. Esta tendência é observada, também, no método QDF, porém, com uma taxa de erro aparente significativamente superior. Os métodos NÚCLEO 1 e NÚCLEO 2 apresentaram uma certa

tendência a classificar predominantemente em Π_1. Nota-se um desempenho bastante superior do método NÚCLEO 2. Já o método dos k-VMP apresentou fraco desempenho.

Método	Observações de Π_2 classificadas em Π_1	Observações de Π_1 classificadas em Π_2	e_{ap}
LDF	13 (26,00%)	67 (33,67%)	32,90%
QDF	7 (14,00%)	79 (39,70%)	37,13%
KRZANOWSKI	12 (24,00%)	65 (32,66%)	31,80%
LOGÍSTICO	13 (26,00%)	70 (35,18%)	34,26%
NÚCLEO 1	23 (46,00%)	58 (29,15%)	30,83%
NÚCLEO 2	15 (30,00%)	16 (8,04%)	10,24%
K-VMP 3	23 (46,00%)	83 (41,71%)	42,14%
K-VMP 5	38 (76,00%)	48 (24,12%)	29,31%
K-VMP 7	25 (50,00%)	75 (37,69%)	38,92%

Tabela 5.3 - Situação B: $c(1|2) = 9\,c(2|1)$.

5.2.3 Comentários e Conclusões

O desempenho relativamente fraco dos métodos apresentados deve-se principalmente ao problema que está sendo analisado. Convém recordar que os dados compõem-se de clientes que foram considerados pelo gerente como bons pagadores e tiveram o seu crédito aprovado. Assim, a discriminação nesta sub-população não é uma tarefa simples. Além disso, as variáveis utilizadas eram as disponíveis que não podem ser consideradas, em absoluto, as melhores em termos de discriminação.

(i) Função Discriminante Linear (LDF)

O método LDF apresentou um desempenho comparável aos outros métodos (com exceção do método NÚCLEO 2). Entretanto, há estudos mais complexos (ver Seção 5.1) em que o método LDF apresenta desempenho nitidamente inferior a outros métodos específicos para variáveis misturadas.

No nosso exemplo, a regra gerada foi:

$$\begin{array}{ll} \text{se} & -0,242\,x_1 + 1,046\,x_2 + 1,385\,x_3^{(1)} + 0,759\,x_3^{(2)} - 0,008\,x_4 + \\ & +0,035\,x_5 - 0,381\,x_6 + 0,249\,x_7 \;>\; \alpha_0 \\ & \\ & \text{então} \qquad \mathbf{x} \in R_1 \\ & \text{c. c.} \qquad \mathbf{x} \in R_2 \end{array}$$

onde $\alpha_0 = +0,745$ para a situação A e $+1,453$ para a situação B.

(ii) Função Discriminante Quadrática (QDF)

O desempenho do método QDF diferiu significativamente do método LDF devido às diferenças entre as matrizes de dispersão de Π_1 e Π_2.

Pode-se observar no Apêndice C que a distribuição dos escores para este método "separa" melhor as populações do que o método anterior. Observe que esta conclusão não poderia ser tirada analisando-se, exclusivamente, as Tabelas 5.2 e 5.3.

(iii) Método de Krzanowski

O problema apresenta três variáveis categóricas: duas binárias (X_1 e X_2) e uma ordinal com quatro categorias (X_3) o que resulta em 16 ($2 \times 2 \times 4$) caselas possíveis. A freqüência amostral por casela é dada na Tabela 5.4.

X_1	X_2	X_3	casela m	n_{1m}	n_{2m}
0	0	0	1	1 (0,50%)	0 (0,00%)
1	0	0	2	1 (0,50%)	0 (0,00%)
0	1	0	3	6 (3,02%)	0 (0,00%)
1	1	0	4	3 (1,51%)	0 (0,00%)
0	0	1	5	1 (0,50%)	0 (0,00%)
1	0	1	6	2 (1,01%)	2 (4,00%)
0	1	1	7	21 (10,55%)	1 (2,00%)
1	1	1	8	23 (11,56%)	1 (2,00%)
0	0	2	9	7 (3,52%)	3 (6,00%)
1	0	2	10	8 (4,02%)	7 (14,00%)
0	1	2	11	28 (14,07%)	3 (6,00%)
1	1	2	12	30 (15,07%)	4 (8,00%)
0	0	3	13	12 (6,03%)	4 (8,00%)
1	0	3	14	12 (6,03%)	8 (16,00%)
0	1	3	15	16 (8,04%)	6 (12,00%)
1	1	3	16	28 (14,07%)	11 (22,00%)
total				199	50

Tabela 5.4 - Freqüência por casela (n_{im} é o número de elementos da população Π_i na casela m).

Observe que há um número elevado de caselas vazias ou com poucas observações. Isto implica que as estimativas produzidas pelo método de Krzanowski não são muito confiáveis o que resulta numa estimativa pouco precisa da taxa de erro real.

(iv) Discriminação Logística

A regra de classificação para este método se assemelha a regra gerada pela LDF, como conseqüência os métodos tiveram desempenhos semelhantes. Este resultado é freqüente na literatura e uma possível justificativa pode ser vista na Seção 3.3.2.

A regra de classificação é dada por:

$$\begin{array}{ll} \text{se} & -0,333\,x_1 + 0,986\,x_2 + 1,800\,x_3^{(1)} + 0,705\,x_3^{(2)} - 0,018\,x_4 + \\ & +0,048\,x_5 - 0,436\,x_6 + 0,258\,x_7 > \alpha_0 \\ \\ \text{então} & \mathbf{x} \in R_1 \\ \text{c. c.} & \mathbf{x} \in R_2 \end{array}$$

onde $\alpha_0 = +0,799$ para a situação A e $+1,398$ para a situação B.

(v) Método do Núcleo

Os parâmetros de alisamento estimados foram:

		Π_1	Π_2
NÚCLEO 1	categóricas	0,98	0,86
	contínuas	0,20	0,28
NÚCLEO 2	categóricas	0,98	0,52
	contínuas	0,20	0,31

Este resultado vai ao encontro do comentário de Vlachonikolis e Marriot (1988) onde, em seus exemplos, também a parte contínua precisava de mais "alisamento" que a parte discreta. Convém recordar que quanto menor o parâmetro de alisamento maior é a contribuição dos "vizinhos" (ver Tabela 4.3).

Observamos que os parâmetros de alisamento para os dois métodos diferem significativamente somente na estimação de δ na população Π_2. Isto se deve ao fato de que para esta população os valores das variáveis categóricas estão mais concentradas do que em Π_1 (ver Análise Descritiva no Apêndice B).

O desempenho do método NÚCLEO 1 apresentou ligeira tendência a classificar em Π_1 resultando numa taxa de erro aparente comparável aos métodos paramétricos e Discrimi-

nação Logística. Já o método NÚCLEO 2 apresentou desempenho significativamente superior em virtude deste método conseguir extrair mais informações de X_3 do que o método anterior.

Outro fator que contribuiu para o melhor desempenho do método NÚCLEO 2 é que a variável X_3 tem alto poder discriminatório.

(vi) Método dos k-Vizinhos Mais Próximos (k-VMP)

O método dos k-VMP apresentou desempenho geral insatisfatório. Uma possível idéia para se melhorar o desempenho deste método neste exemplo em particular, seria definir uma medida de distância que considerasse a estrutura de ordem em X_3 aos moldes do que foi sugerido para o método do Núcleo (Seção 4.1.3).

Outros valores para k foram testados e os resultados sempre se mostraram inferiores à LDF.

A partir deste exemplo podemos tirar algumas conclusões que acreditamos ser válidas para problemas relacionados com *credit scoring* em geral. São elas:

(i) Neste tipo de problema as probabilidades *a priori* são, geralmente, bastante diferentes. A utilização de custos de classificação incorreta é uma solução elegante, entretanto, poucos são os casos em que se pode avaliá-los com precisão. Deste modo, sugerimos que a escolha da regra seja feita considerando vários pontos de corte e selecionando aquele que, administrativamente, mais convier à empresa.

(ii) O fraco desempenho apresentado pelo método de Krzanowski neste exemplo deve-se ao pequeno tamanho da amostra em relação ao número de caselas geradas. Os resultados obtidos não refletem o elevado potencial deste método nos casos em que seja possível estimar, com confiança, todos os parâmetros envolvidos.

(iii) Os resultados obtidos pelos métodos LDF, QDF e LOGÍSTICO poderiam ter sido melhores caso fossem efetuadas técnicas de seleção de variáveis. Os outros métodos são menos sensíveis à exclusão de variáveis pouco informativas.

(iv) A utilização do *credit scoring* na prática está mais relacionada com sitemas de apoio à decisão do que, propriamente, com sistemas em que o único objetivo é classificar uma observação. Freqüentemente, é de interesse que o método utilizado gere informações a respeito da classificação sugerida, informando, por exemplo, quais são as variáveis "mais discriminadoras". Dessa forma, os métodos Núcleo e k-VMP são pouco recomendáveis na prática, uma vez que geram regras pouco informativas.

APÊNDICE

APÊNDICE A - Detalhes sobre a Aplicação do Método de Newton-Raphson em Discriminação Logística

Para se determinar o valor de $\boldsymbol{\alpha}$ que maximiza $\mathrm{L}(\mathbf{D},\boldsymbol{\alpha})$, dado em (3–6), através do método de Newton-Raphson é necessário calcular os valores de suas derivadas até segunda ordem:

$$\mathrm{L}(\mathbf{D},\boldsymbol{\alpha}) = \sum_{i=1}^{g} \sum_{k\in E_i} \ln \Pr(\,\Pi_i \,|\, \mathbf{x}_k\,)$$

$$\frac{\partial \mathrm{L}(\mathbf{D},\boldsymbol{\alpha})}{\partial \alpha_{ij}} = \sum_{k\in E_i} x_{kj} - \sum_{k\in E_i} x_{kj} \Pr(\,\Pi_i \,|\, x_k\,) \qquad (i = 1,\ldots, g-1;\, j = 0,\ldots, r)$$

$$\frac{\partial^2 L(\mathbf{D},\boldsymbol{\alpha})}{\partial \alpha_{hm} \partial \alpha_{ij}} = \begin{cases} \sum\limits_{k\in E_i} \Pr(\,\Pi_i \,|\, x_k\,) \Pr(\,\Pi_h \,|\, x_k\,)\; x_{kj}\; x_{km}, & \text{se } h \neq i \\[2ex] -\sum\limits_{k\in E_i} \Pr(\,\Pi_i \mid x_k\,)\,[\,1 - \Pr(\,\Pi_i \,|\, x_k\,)\,]\, x_{kj}\; x_{km}, & \text{se } h = i \end{cases}$$

para $h,\, i = 1,\ldots, g-1;\;\; m,\, j = 0,\ldots, r$ e, como antes $E = \{1,\ldots, n\}$.

Para sermos mais precisos, $S(\hat{\boldsymbol{\alpha}}_n)$ é um vetor de dimensão $(g-1)(r+1)$ contendo o seguinte valor na l-ésima linha:

$$\left.\frac{\partial \mathrm{L}(\mathbf{D},\boldsymbol{\alpha})}{\partial \alpha_{ij}}\right|_{\hat{\boldsymbol{\alpha}}_n} \qquad (i = 1,\ldots, g-1;\, j = 0,\ldots, r)$$

onde $l = (i-1)(r+1) + j + 1$.

E $I(\hat{\boldsymbol{\alpha}}_n)$ é a matriz quadrada de dimensão $(g-1)(r+1)$ contendo o seguinte valor em sua l-ésima linha e c-ésima coluna:

$$\left.\frac{\partial^2 \mathrm{L}(\mathbf{D},\boldsymbol{\alpha})}{\partial \alpha_{hm} \partial \alpha_{ij}}\right|_{\hat{\boldsymbol{\alpha}}_n} \qquad (h,\, i = 1,\ldots, g-1;\, m,\, j = 0,\ldots, r)$$

onde $l = (i-1)(r+1) + j + 1$ e $c = (h-1)(r+1) + m + 1$.

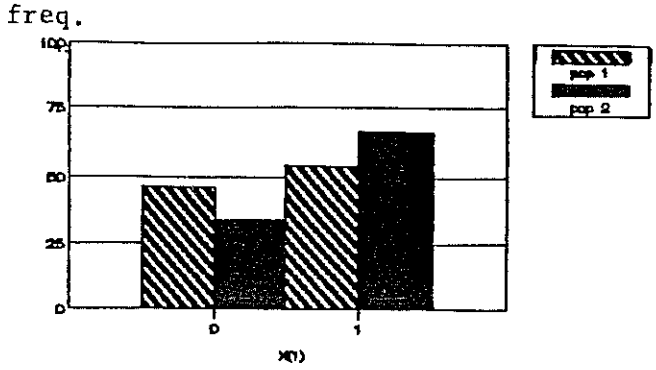
Dado o valor inicial de $\hat{\boldsymbol{\alpha}}_0$ obtem-se os demais através da regra de formação $(3-9)$.

APÊNDICE B - Estudo Descritivo das Variáveis que Compõem o Exemplo da Seção 5.2

VARIÁVEL X_1

tipo: binária (0-1)

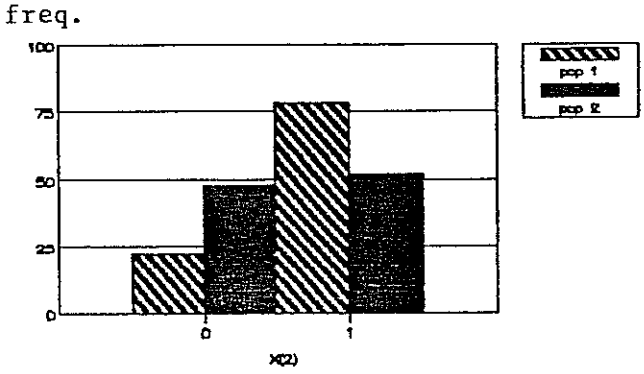
X_1	Π_1	Π_2
0	92(46%)	17(34%)
1	107(54%)	33(66%)



VARIÁVEL X_2

tipo: binária (0-1)

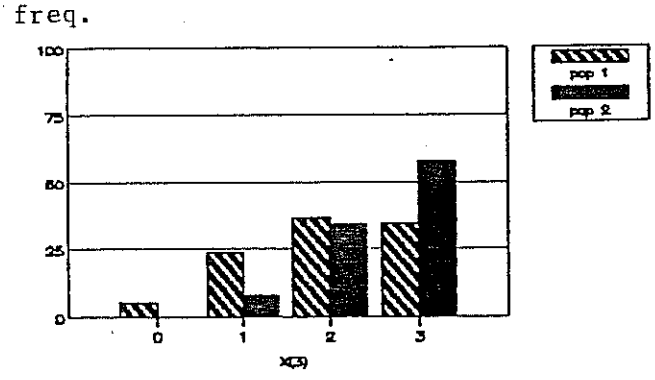
X_2	Π_1	Π_2
0	44(22%)	24(48%)
1	155(78%)	26(52%)



VARIÁVEL X_3

tipo: qualitativa ordinal

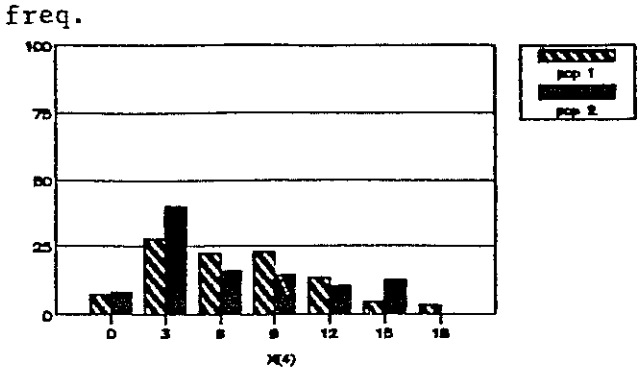
X_3	Π_1	Π_2
0	11(5%)	0(0%)
1	47(24%)	4(8%)
2	73(37%)	17(34%)
3	68(34%)	29(58%)



VARIÁVEL X_4

tipo: contínua

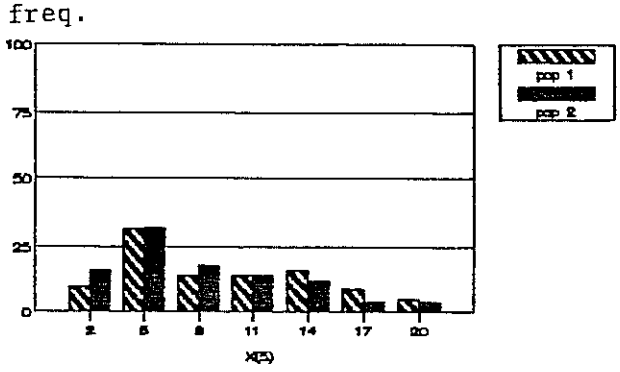
X_4	Π_1	Π_2
média	7,00	6,24
mínimo	0,00	1,00
máximo	19,00	15,00
desvio-padrão	4,21	4,46



VARIÁVEL X_5

tipo: contínua

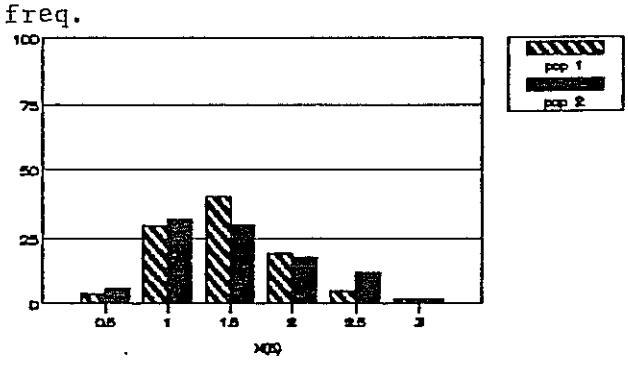
X_5	Π_1	Π_2
média	9,53	8,22
mínimo	2,00	2,00
máximo	20,00	19,00
desvio-padrão	4,92	4,54



VARIÁVEL X_6

tipo: contínua

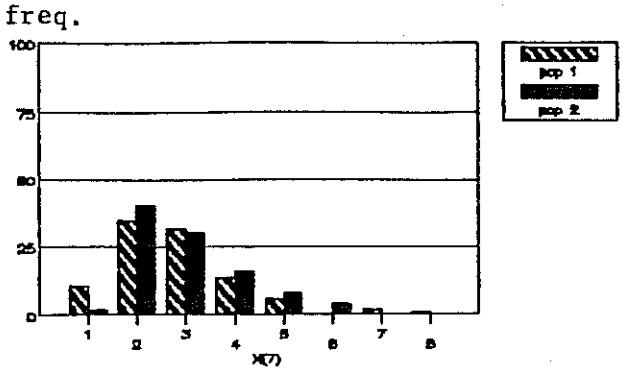
X_6	Π_1	Π_2
média	1,49	1,54
mínimo	0,29	0,42
máximo	3,21	2,79
desvio-padrão	0,49	0,55



VARIÁVEL X_7

tipo: contínua

X_7	Π_1	Π_2
média	2,83	2,93
mínimo	0,38	0,97
máximo	8,23	5,72
desvio-padrão	1,30	1,13



APÊNDICE C - Distribuição dos Escores para os Métodos Abordados no Exemplo da Seção 5.2

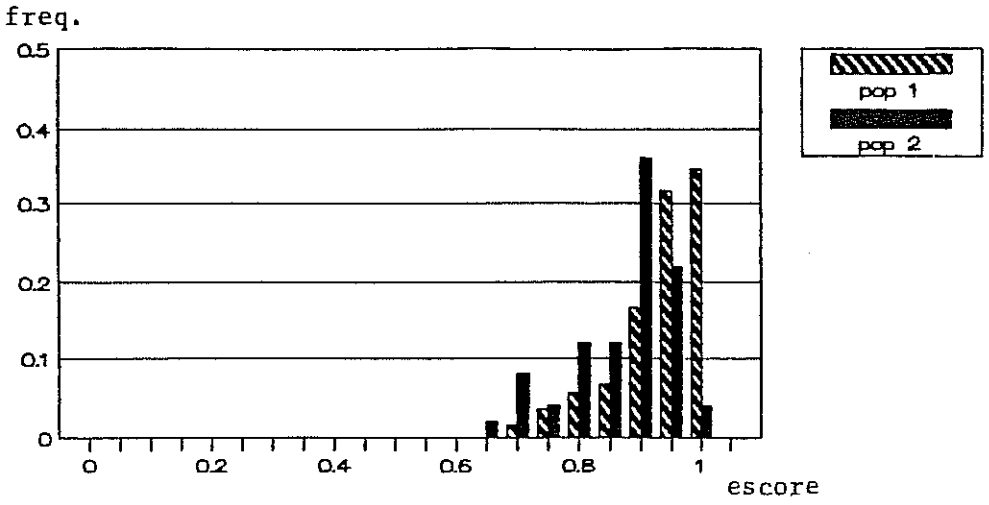


Figura C.1 - Método LDF.

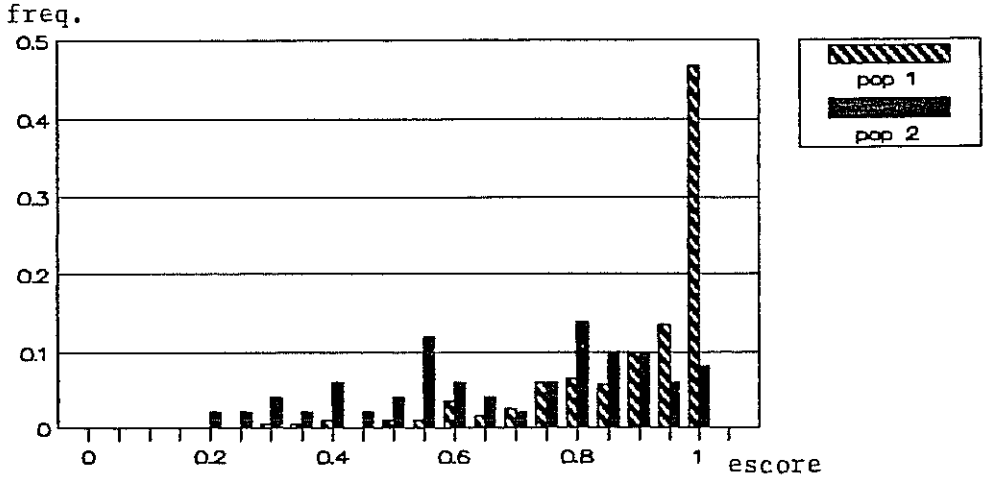


Figura C.2 - Método QDF.

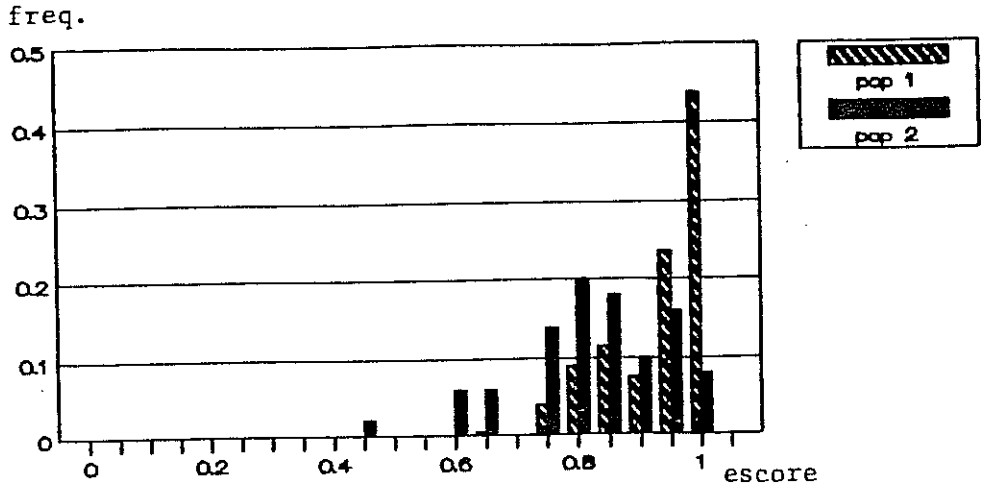


Figura C.3 - Método de Krzanowski.

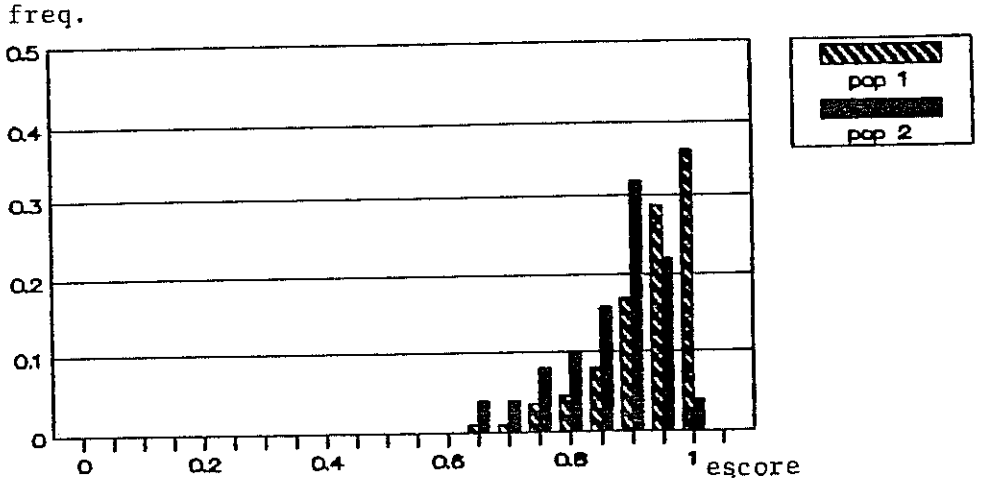


Figura C.4 - Discriminação Logística.

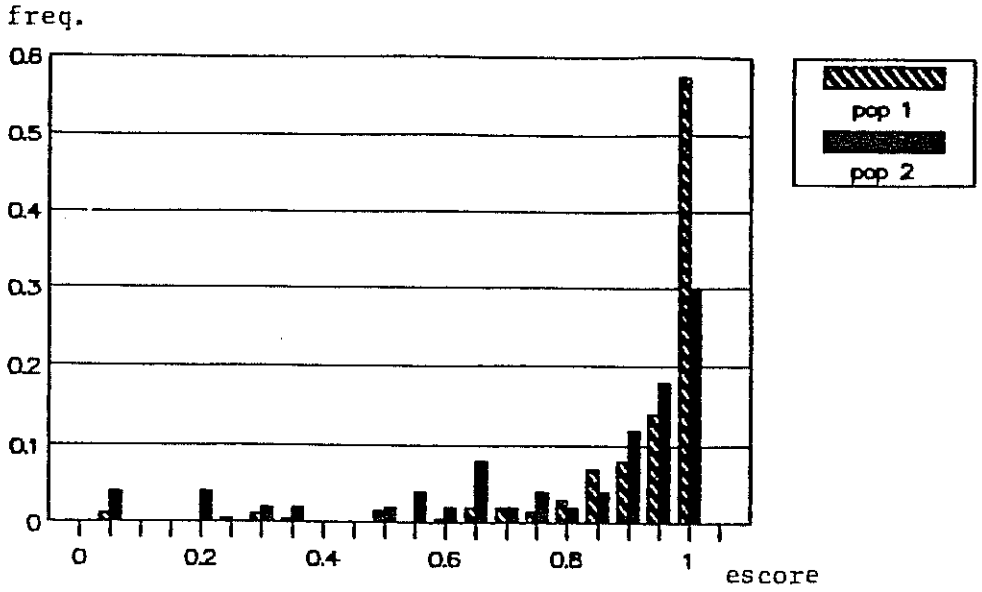


Figura C.5 - Método do Núcleo (Variável X_3 tratada como qualitativa nominal).

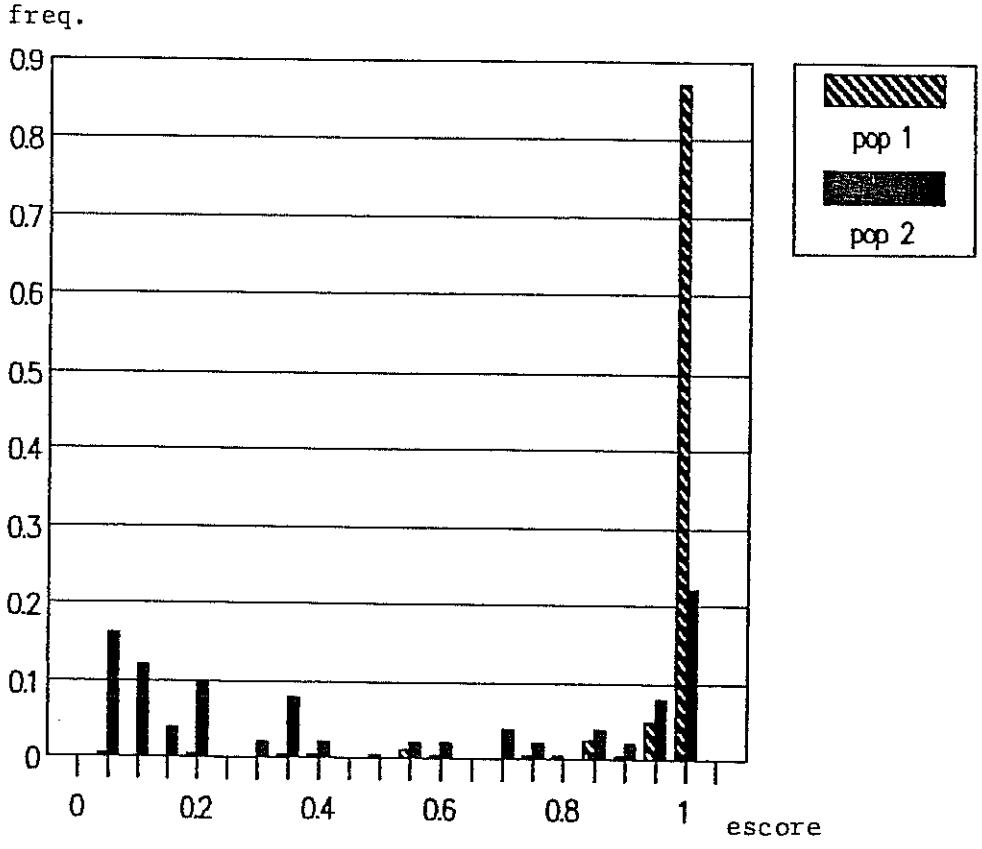


Figura C.6 - Método do Núcleo (Variável X_3 tratada como qualitativa ordinal).

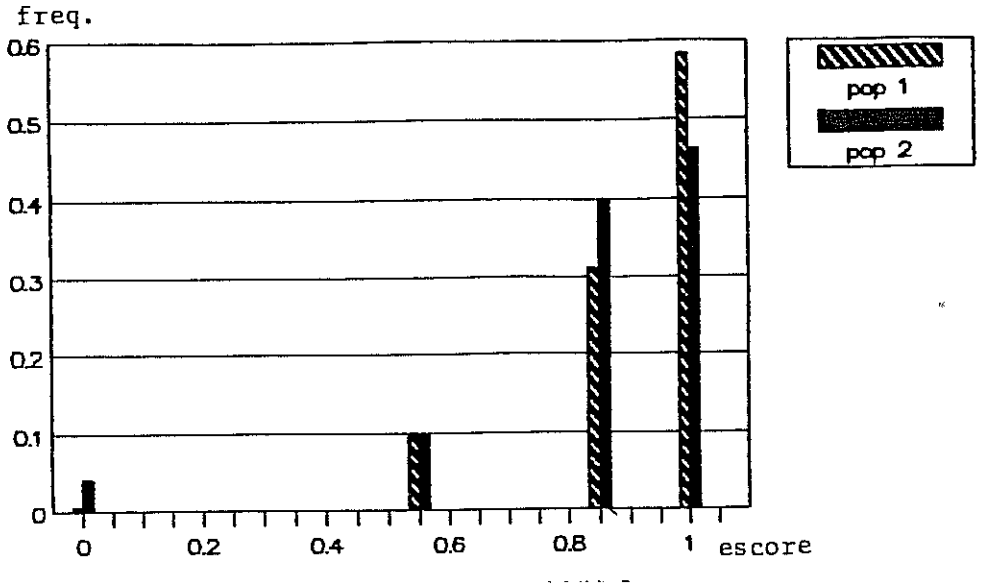


Figura C.7 - Método dos k-VMP (k=3).

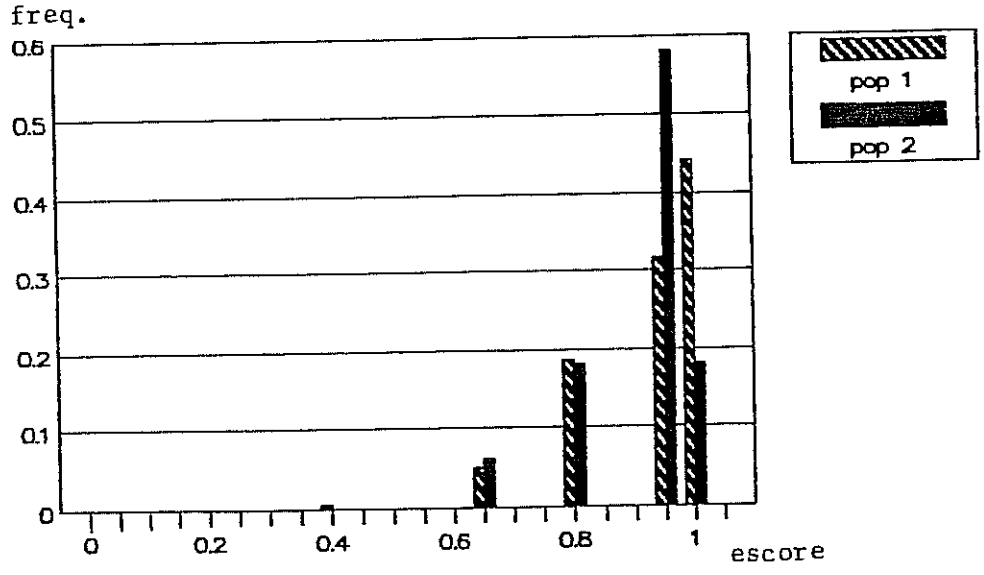


Figura C.8 - Método dos k-VMP (k=5).

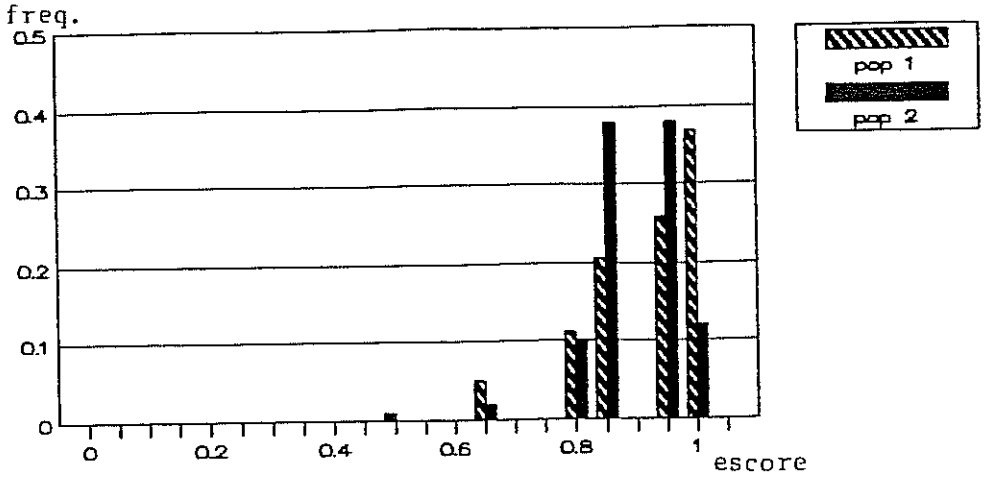


Figura C.9 - Método dos k-VMP (k=7).

APÊNDICE D - Listagem do Programa AD

```
{escrito em TURBO PASCAL}

PROGRAM ad;

  USES CRT;
  { declaracao das variaveis }
CONST n_r  =  10;
      n_n  = 200;
      n_no = 100;

TYPE  mr   = ARRAY [1..n_r] OF REAL;
      mrr  = ARRAY [1..n_r,1..n_r] OF REAL;
      mn   = ARRAY [1..n_n] OF REAL;
      mnr  = ARRAY [1..n_n,1..n_r] OF REAL;
      mn_i = ARRAY [1..n_n] OF INTEGER;
      morr = ARRAY [0..n_r,0..n_r] OF REAL;

VAR   p,q,r,n,n1,n2,no,g                    :INTEGER;
      pp1,pp2                               :REAL;
      pop,pop_est                           :mn_i;
      dados                                 :mnr;
      maxi2,mini2                           :mr;
      pr_est              :ARRAY[1..n_n+n_no] OF REAL;
      pop_est2            :ARRAY[1..n_no] OF INTEGER;
      c                   :ARRAY[1..n_r] OF INTEGER;
      conf                :ARRAY[1..2,1..2] OF INTEGER;
      opcao,exit                            :CHAR;
      leu_dados,separadas                   :BOOLEAN;
      sa                                    :TEXT;


{-------------------------------------------------------------

                    FUNCOES E PROCEDIMENTOS AUXILIARES

-------------------------------------------------------------}

PROCEDURE tela_menu;
```

```
BEGIN
  WRITELN('[ 1 ]  Informacoes                          ');
  WRITELN('[ 2 ]  Leitura de dados                     ');
  WRITELN('[ 3 ]  Funcao Discriminante Linear          ');
  WRITELN('[ 4 ]  Funcao Discriminante Quadratica');
  WRITELN('[ 5 ]  Location model (Krzanowski)          ');
  WRITELN('[ 6 ]  Discriminacao Logística              ');
  WRITELN('[ 7 ]  Núcleo (Kernel)                      ');
  WRITELN('[ 8 ]  k-VMP  (K-nn)                        ');
  WRITELN('[ 9 ]  Saída                                ');
END;

{-------------------------------------------------------------}

PROCEDURE informacoes_1;

BEGIN
  CLRSCR;
WRITELN('AD e um "software" estatistico que engloba 6
tecnicas estatisticas');
WRITELN('diferentes para classificar observacoes em
populacoes pre-definidas.');
WRITELN('Estas tecnicas, chamadas de Analise Discriminante
podem ser aplica-');
WRITELN('das em problemas onde ha mistura de variaveis
continuas e categori-');
WRITELN('cas. Para a execucao do programa e necessario que o
usuario crie um');
WRITELN('arquivo de dados no formato padrao ASC II com a
seguinte estrutura:');
WRITELN('        x(1)   x(2)   ...   x(q)   x(q+1)   ...
x(r)   pop');
WRITELN('  1
1 ');
WRITELN('  2          n1 observacoes da populacao 1
1 ');
WRITELN('  :
: ');
WRITELN('  n1
1 ');
WRITELN(' n1+1
2 ');
WRITELN(' n1+2        n2 observacoes da populacao 2
2 ');
WRITELN('  :
: ');
WRITELN(' n1+n2
2 ');
WRITELN('  1
0 ');
WRITELN('  :    e, opcionalmente, n0 observacoes a serem
0 ');
WRITELN('  n0                classificadas
0 ');
```

```
WRITELN('Cada observacao ocupa uma linha. As variaveis sao
dispostas nas co-');
WRITELN('lunas sendo que as variaveis categoricas devem ser
colocadas primei-');
WRITELN('ro. A ultima coluna e reservada para indicar a
populacao a que per-');
WRITELN('tence a populacao.  A separacao dos dados pode ser
feita por branco,');
WRITELN('virgula ou ponto e virgula.');
WRITELN('TECLE ENTER PARA PROSSEGUIR');
          opcao:=READKEY;
          CLRSCR;
        END;

{-----------------------------------------------------------}

        PROCEDURE informacoes_2;

        BEGIN
          CLRSCR;
          WRITELN;
          WRITELN('arquivo de dados nao informado - entrar
na opcao 1');
          WRITELN;
          WRITE('TECLE ENTER PARA PROSSEGUIR');
          opcao:=READKEY;
          CLRSCR;
        END;

{-----------------------------------------------------------}

        FUNCTION ndif(f : INTEGER) : INTEGER;

        VAR nconj,a,vndif,i            :INTEGER;
            conjunto                   :mn;
            achou                      :BOOLEAN;

        BEGIN
          vndif:=1;
          nconj:=1;
          conjunto[1]:=dados[1,f];
          FOR a:=2 TO n DO
            BEGIN
              achou:=FALSE;
              i:=1;
              WHILE (i<=nconj) AND (NOT achou) DO
                IF conjunto[i]=dados[a,f]
                THEN achou:=true
                ELSE i:=i+1;
              IF NOT achou
              THEN
                BEGIN
                  vndif:=vndif+1;
                  conjunto[nconj+1]:=dados[a,f];
```

```
                nconj:=nconj+1
              END;
          END;
        ndif:=vndif;
      END;

{-----------------------------------------------------------}

      PROCEDURE le_dados;

      VAR a,b,i,j,tcoluna,dlinha,
          dcoluna,fim,status,sup     :INTEGER;
          minid,maxid                :REAL;
          np                       :ARRAY[0..2] OF INTEGER;
          adados                     :TEXT;
          tnum,saida                 :STRING[20];
          texto                      :STRING[255];
          resp                       :CHAR;

      BEGIN
        FILLCHAR(dados,SIZEOF(dados),0);
        FILLCHAR(pop,SIZEOF(pop),0);
        WRITE('saida na tela ? (S/N) ');
        resp:=READKEY;
        WRITELN(resp);
        IF (resp='s') OR (resp='S')
        THEN saida:='CON'
        ELSE
          BEGIN
            WRITE('saida: ');
            READLN(saida);
          END;
        WRITE('nome do arquivo de dados = ');
        READLN(texto);
        WRITE('amostragem separada ? (S/N) ');
        resp:=READKEY;
        WRITELN(resp);
        IF (resp='s') OR (resp='S')
        THEN separadas:=true
        ELSE separadas:=false;
        WRITE('numero de variaveis qualitativas = ');
        READ(q);
        ASSIGN(sa,saida);
        REWRITE(sa);
        tnum:=' ';
        n:=0;
        r:=0;
        ASSIGN(adados,texto);
        RESET(adados);
        dlinha:=1;
        WHILE (NOT EOF(adados)) DO
          BEGIN
            texto:='';
            tcoluna:=1;
```

```
                dcoluna:=1;
                READLN(adados,texto);
                fim:=LENGTH(texto);
                WHILE tcoluna <= fim DO
                  IF (texto[tcoluna] IN ['0'..'9','.','-'])
                  THEN
                    BEGIN
                      tnum:='';
                      tnum:=texto[tcoluna];
                      INC(tcoluna);
                      a:=1;
              WHILE (texto[tcoluna] IN ['0'..'9','.','-']) AND
(tcoluna <= fim) DO
                        BEGIN
                          tnum:=tnum+texto[tcoluna];
                          INC(tcoluna);
                          INC(a);
                        END;
                      tnum[0]:=chr(a);
                      val(tnum,dados[dlinha,dcoluna],status);
                      INC(dcoluna);
                      tnum:='';
                    END
                  ELSE
                    BEGIN
                      tnum:='';
                      INC(tcoluna);
                    END;
                dlinha:=dlinha+1;
              END;
            close(adados);
            DEC(dlinha);
            DEC(dcoluna);
            r:=dcoluna-1;
            FOR a:=1 TO dlinha DO
              b:=trunc(dados[a,r+1]);
            np[0]:=0;
            np[1]:=0;
            np[2]:=0;
            FOR a:=1 TO dlinha DO
              BEGIN
                b:=trunc(dados[a,r+1]);
                pop[a]:=b;
                np[b]:=np[b]+1;
              END;
            sup:=0;
            no:=np[0];
            n1:=np[1];
            n2:=np[2];
            n:=n1+n2;
            WRITELN(sa);
            WRITELN(sa,' ANALISE DOS DADOS ');
            WRITELN(sa);
            WRITELN(sa,'parte dos dados lidos:');
```

```
          WRITELN(sa);
          WRITELN(sa,'-------------------------------------
-----------------');
          FOR i:=1 TO 5 DO
            BEGIN
              WRITE(sa,i:1);
              FOR j:=1 TO r DO
                WRITE(sa,' ',dados[i,j]:4:1,' ');
              WRITE(sa,' ',pop[i]);
              WRITELN(sa);
            END;
          WRITELN(sa,' ... ');
          FOR i:=n+no-3 TO n+no DO
            BEGIN
              WRITE(sa,i:1);
              FOR j:=1 TO r DO
              WRITE(sa,' ',dados[i,j]:4:1,' ');
              WRITE(sa,' ',pop[i]);
              WRITELN(sa);
            END;
          WRITELN(sa,'-------------------------------------
-----------------');
          WRITELN(sa);
          WRITELN(sa,'n = ',n);
          WRITELN(sa,'r = ',r);
          WRITELN(sa,'q = ',q);
          WRITELN(sa);
          FOR b:=1 TO q DO
            BEGIN
              c[b]:=ndif(b);
              WRITELN(sa,'c[',b,'] = ',c[b]);
            END;
          FOR b:=1 TO r DO
            BEGIN
              maxid:=dados[1,b];
              minid:=maxid;
              FOR a:=2 TO dlinha DO
                BEGIN
                  IF maxid < dados[a,b]
                  THEN maxid:=dados[a,b];
                  IF minid > dados[a,b]
                  THEN minid:=dados[a,b]
                END;
              maxi2[b]:=maxid;
              mini2[b]:=minid;
        WRITELN(sa,'mini[',b,'] =',mini2[b]:6:2,'
maxi[',b,'] = ',maxi2[b]:6:2);
            END;
          i:=1;
          WRITELN(sa);
          WRITELN(sa,'no = ',no);
          WRITELN(sa,'n1 = ',n1);
          WRITELN(sa,'n2 = ',n2);
          p:=r-q;
```

```
          WRITELN(sa);
          IF separadas
          THEN
            BEGIN
              WRITE('pp1 = ');
              READLN(pp1);
              WRITE('pp2 = ');
              READLN(pp2);
            END
          ELSE
            BEGIN
              pp1:=n1/n;
              WRITELN(sa,'pp1 = ',pp1:6:4);
              pp2:=n2/n;
              WRITELN(sa,'pp2 = ',pp2:6:4);
            END;
          WRITELN(sa);
        END;

{-----------------------------------------------------------}

        PROCEDURE resultados;

        VAR i,j                          :INTEGER;
            e,a,b,a1,b1                  :REAL;

        BEGIN
          WRITELN(sa);
          WRITELN(sa);
          WRITELN(sa,'          Pop. Atual
Populacao Prevista');
          WRITELN(sa,'          ----------          --------
----------');
          WRITE(sa,'                                  ');
          FOR i:=1 TO 2 DO
            WRITE(sa,i,'                ');
          WRITELN(sa);
          WRITELN(sa);
          FOR i:=1 TO 2 DO
            BEGIN
              WRITE(sa,'                ',i,'
');
              FOR j:=1 TO 2 DO
              WRITE(sa,conf[i,j],'                ');
              WRITELN(sa);
            END;
          e:=conf[1,2]*pp1/n1+conf[2,1]*pp2/n2;
          WRITELN(sa);
          WRITELN(sa);
          WRITELN(sa);
          WRITELN(sa);
          WRITELN(sa);
          WRITELN(sa,'                                  Taxa de
erro aparente');
```

```
          WRITELN(sa,'                                          ----------
-----------');
          WRITELN(sa);
          WRITELN(sa,'                                                E =
',e:1:5);
          WRITELN(sa);
          WRITELN(sa);
          WRITELN(sa);
          IF no<> 0
          THEN
            BEGIN
              WRITELN(sa,'            CLASSIFICACAO DAS
OBSERVACOES REQUERIDAS');
              WRITELN(sa);
              WRITELN(sa,'                obs.:
pop. estimada');
              FOR j:=1 TO no DO
              WRITELN(sa,'                ',j,'
',pop_est2[j]);
            END;
        END;

{-------------------------------------------------------------}

        FUNCTION pot(a,b : REAL) : REAL;

        VAR      x                      :REAL;

        BEGIN
          IF a > 0
          THEN x:=EXP(b*LN(a))
          ELSE IF a = 0
               THEN x:=0
               ELSE IF FRAC(b/2)=0
                    THEN x:=EXP(b*LN(-a))
                    ELSE x:=-EXP(b*LN(-a));
          pot:=x;
        END;

{-------------------------------------------------------------}

        FUNCTION deter(VAR m : mrr; VAR v : INTEGER) : REAL;

        { descricao : calcula o determinante de m, matriz v
x v }

        VAR   i,j,k                      :INTEGER;
              det,cte                    :REAL;
              m_aux                      :mr;

        BEGIN
          det:=1;
          FOR k:=1 TO v-1 DO
            BEGIN
```

```
            IF m[k,k]=0
            THEN
              BEGIN
                deter:=0;
                k:=v;
              END
            ELSE
              FOR i:=k+1 TO v DO
                BEGIN
                  IF m[i,k]<>0
                  THEN
                    BEGIN
                      cte:=m[k,k]/m[i,k];
                      det:=det/cte;
                      FOR j:=1 TO v DO
                        m[i,j]:=m[i,j]*cte-m[k,j];
                    END;
                END;
            det:=det*m[k,k];
          END;
        det:=det*m[v,v];
        deter:=det;
      END;

{-----------------------------------------------------------}

      FUNCTION k_q(z,zk : mr; delta : REAL) : REAL;

      VAR   j                                  :INTEGER;
            prod                               :REAL;

      BEGIN
        prod:=1;
        FOR j:=1 TO q DO
          IF z[j] = zk[j]
          THEN prod:=prod*delta
          ELSE prod:=prod*(1-delta)/(c[j]-1);
        k_q:=prod;
      END;

{-----------------------------------------------------------}

      FUNCTION k_p(y,yk : mr; lambda : REAL) : REAL;

      VAR   j                                  :INTEGER;
            sum,aux                            :REAL;

      BEGIN
        sum:=0;
        FOR j:=1 TO p DO
          sum:=sum+pot((y[j]-yk[j]),2);
        aux:=-p/2*LN(2*PI*lambda)-sum/2/lambda;
        IF aux>80
        THEN aux:=8
```

```
          ELSE IF aux<-80
               THEN aux:=-80;
          k_p:=EXP(aux);
        END;

{-------------------------------------------------------------}

        FUNCTION f_q(delta :

        REAL; inf,sup : INTEGER) : REAL;

        VAR   k,j,l                                  :INTEGER;
              sum,lnprod                             :REAL;
              zl,zk                                  :mr;

        BEGIN
          lnprod:=0;
          FOR k:=inf TO sup DO
            BEGIN
              sum:=0;
              FOR j:=1 TO q DO
                zk[j]:=dados[k,j];
              FOR l:=inf TO sup DO
                IF l <> k
                THEN
                  BEGIN
                    FOR j:=1 TO q DO
                      zl[j]:=dados[l,j];
                    sum:=sum+k_q(zl,zk,delta);
                  END;
              lnprod:=lnprod+LN(sum/(sup-inf));
            END;
          f_q:=lnprod;
        END;

{-------------------------------------------------------------}

        FUNCTION f_p(lambda : REAL; inf,sup : INTEGER) :
REAL;

        VAR   k,j,l                                     :INTEGER;
              sum,lnprod                             :REAL;
              yl,yk                                  :mr;

        BEGIN
          lnprod:=0;
          FOR k:=inf TO sup DO
            BEGIN
              sum:=0;
              FOR j:=q+1 TO r DO
                yk[j-q]:=dados[k,j];
              FOR l:=inf TO sup DO
                IF l <> k
                THEN
```

```
                    BEGIN
                      FOR j:=q+1 TO r DO
                        yl[j-q]:=dados[l,j];
                      sum:=sum+k_p(yk,yl,lambda);
                    END;
                lnprod:=lnprod+LN(sum/(sup-inf));
              END;
            f_p:=lnprod;
          END;

{-----------------------------------------------------------}

          PROCEDURE ache_max(VAR k_a,k_c: REAL; cat: BOOLEAN;
inf,sup: INTEGER);

          VAR   k_b,epsilon,f1,f2,f3     :REAL;

          BEGIN
            WRITELN(sa,'a = ',k_a:4:2,'  c = ',k_c:4:2);
            epsilon:=0.01;
            k_b:=(k_a+k_c)/2;
            IF cat
            THEN
              BEGIN
                f1:=f_q(k_b-epsilon,inf,sup);
                f2:=f_q(k_b,inf,sup);
                f3:=f_q(k_b+epsilon,inf,sup);
                WRITELN(sa,'f1 = ',f1:6:4,'  f2 = ',f2:6:4,'
f3 = ',f3:6:4);
                                IF (f1 < f2) AND (f2 < f3)
                THEN k_a:=k_b
                ELSE IF (f1 > f2) AND (f2 > f3)
                     THEN k_c:=k_b
                     ELSE IF (f1 < f2) AND (f2 > f3)
                          THEN
                            BEGIN
                              k_a:=k_b;
                              k_c:=k_b;
                            END;
              END
            ELSE
              BEGIN
                f1:=f_p((k_b-epsilon),inf,sup);
                f2:=f_p(k_b,inf,sup);
                f3:=f_p((k_b+epsilon),inf,sup);
                WRITELN(sa,'f1 = ',f1:6:4,'  f2 = ',f2:6:4,'
f3 = ',f3:6:4);
                IF (f1 < f2) AND (f2 < f3)
                THEN   k_a:=k_b
                ELSE IF (f1 > f2) AND (f2 > f3)
                     THEN k_c:=k_b
                     ELSE IF (f1 < f2) AND (f2 > f3)
                          THEN
                            BEGIN
```

```
                                k_a:=k_b;
                                k_c:=k_b;
                              END;
            END;
        END;

{-------------------------------------------------------------}

        PROCEDURE dispersao(dados : mnr; l1,l2,c1,c2 :
integer; VAR cov : mrr);

{ descricao: calcula a matriz de variancias e covariancias
             denominada cov (de dimensao c2-c1+1) que e'
             estimada a partir da amostra representada pela
             matriz de dados da linha l1 ate a linha l2 e da
             coluna c1 ate a coluna c2 }

        VAR   i,j,k                                :INTEGER;
              ex                                   :mr;

        BEGIN
          FILLCHAR (ex,SIZEOF(ex),0);
          FILLCHAR (cov,SIZEOF(cov),0);
          FOR j:=c1 TO c2 DO
            BEGIN
              FOR k:=l1 TO l2 DO
                ex[j-c1+1]:=ex[j-c1+1]+dados[k,j];
              ex[j-c1+1]:=ex[j-c1+1]/(l2-l1+1);
            END;
          FOR i:=c1 TO c2 DO
            FOR j:=c1 TO c2 DO
              FOR k:=l1 TO l2 DO
                cov[i-c1+1,j-c1+1]:=cov[i-c1+1,j-c1+1]+
                (dados[k,i]-ex[i-c1+1])*(dados[k,j]-ex[j-
c1+1]);
          FOR i:=1 TO (c2-c1+1) DO
            FOR j:=1 TO (c2-c1+1) DO
              cov[i,j]:=cov[i,j]/(l2-l1);
        END;

{-------------------------------------------------------------}

        PROCEDURE inversa(cov : mrr; r : integer; VAR invcov
: mrr);

{ descricao: calcula a matriz inversa de cov e armazena
             na matriz invcov }

        VAR   mat                                  :morr;
              co,id,li,i,j,nvar                    :INTEGER;
              de,be                                :REAL;

        BEGIN
```

```
          FOR i:=1 TO r DO
            FOR j:=1 TO r DO
              mat[i-1,j-1]:=cov[i,j];
          FOR i:=1 TO r DO
            BEGIN
              mat[i,r]:=1;
              mat[r,i]:=1;
            END;
          nvar:=r-1;
          FOR li:=0 TO nvar DO
            BEGIN
              de:=mat[li,li];
              IF de=0
              THEN
                WRITELN('MULTICOLINEARIDADE !!! - PROGRAMA
ABORTADO');
              FOR co:=0 TO nvar+1 DO
                mat[li,co]:=mat[li,co]/de;
              FOR id:=0 TO nvar+1 DO
                BEGIN
                  IF id<>li
                  THEN
                    BEGIN
                      be:=mat[id,li];
                      FOR co:=0 TO nvar+1 DO
                          IF mat[li,co]>exp(50)
                          THEN mat[id,co]:=mat[id,co]-
exp(50)*be
                          ELSE mat[id,co]:=mat[id,co]-
mat[li,co]*be;
                      mat[id,li]:=-be/de;
                    END;
                END;
              mat[li,li]:=1/de;
            END;
          FOR li:=0 TO nvar DO
            FOR co:=0 TO nvar DO
              invcov[li+1,co+1]:=mat[li,co];
        END;

{-------------------------------------------------------------}

        PROCEDURE inversa_log(cov : morr; v:integer; VAR
invcov : morr);


        VAR   mat                                   :morr;
              co,id,li,i,j,nvar                     :INTEGER;
              de,be                                 :REAL;

        BEGIN
          FOR i:=1 TO v DO
            FOR j:=1 TO v DO
              mat[i-1,j-1]:=cov[i,j];
```

```
          FOR i:=1 TO v DO
            BEGIN
              mat[i,v]:=1;
              mat[v,i]:=1;
            END;
          nvar:=v-1;
          FOR li:=0 TO nvar DO
            BEGIN
              de:=mat[li,li];
              IF de=0
              THEN
                WRITELN('MULTICOLINEARIDADE !!! - PROGRAMA
ABORTADO');
              FOR co:=0 TO nvar+1 DO
                mat[li,co]:=mat[li,co]/de;
              FOR id:=0 TO nvar+1 DO
                BEGIN
                  IF id<>li
                  THEN
                    BEGIN
                      be:=mat[id,li];
                      FOR co:=0 TO nvar+1 DO
                          IF mat[li,co]>exp(50)
                          THEN mat[id,co]:=mat[id,co]-
exp(50)*be
                          ELSE mat[id,co]:=mat[id,co]-
mat[li,co]*be;
                      mat[id,li]:=-be/de;
                    END;
                END;
              mat[li,li]:=1/de;
            END;
         FOR li:=0 TO nvar DO
           FOR co:=0 TO nvar DO
             invcov[li+1,co+1]:=mat[li,co];
        END;

{-------------------------------------------------------------}

        PROCEDURE bolha(VAR tab : mn; VAR popu : mn_i; knn :
INTEGER);

{ descricao: ordena o vetor popu de acordo com o vetor tab }

        VAR   k,j,ind,auxi                        :INTEGER;
              aux,mini                            :REAL;

        BEGIN
          FOR k:=1 TO knn+1 DO
            BEGIN
              mini:=tab[k];
              ind:=k;
              FOR j:=k+1 TO n DO
                BEGIN
```

```
                IF tab[j] < mini
                THEN BEGIN
                       mini:=tab[j];
                       ind:=j;
                     END;
              END;
            aux:=tab[k];
            tab[k]:=mini;
            tab[ind]:=aux;
            auxi:=popu[k];
            popu[k]:=popu[ind];
            popu[ind]:=auxi;
          END;
        END;

{-----------------------------------------------------------}

                    PROCEDIMENTOS PRINCIPAIS

{-----------------------------------------------------------}

        PROCEDURE ldf;

        VAR     i,j,k                          :INTEGER;
                sum,constante                  :REAL;
                media1,media2,soma,dif,m_rx1,
                alfa                           :mr;
                spooled,invs                   :mrr;
                m_n            :ARRAY[1..n_n+n_no] OF REAL;

        BEGIN
          FILLCHAR(media1,SIZEOF(media1),0);
          FILLCHAR(media2,SIZEOF(media2),0);
          WRITELN(sa,'                FUNCAO DISCRIMINANTE
LINEAR');
          FOR i:=1 TO n1 DO
            FOR j:=1 TO r DO
              media1[j]:=media1[j]+dados[i,j];
          FOR i:=n1+1 TO n DO
            FOR j:=1 TO r DO
              media2[j]:=media2[j]+dados[i,j];
          FOR j:=1 TO r DO
            BEGIN
              media1[j]:=media1[j]/n1;
              media2[j]:=media2[j]/n2;
              soma[j]:=media1[j]+media2[j];
              dif[j]:=media1[j]-media2[j];
            END;
          dispersao(dados,1,n,1,r,spooled);
          inversa(spooled,r,invs);
          FOR i:=1 TO r DO
            BEGIN
              sum:=0;
              FOR j:=1 TO r DO
```

```
                sum:=sum+invs[i,j]*dif[j];
              alfa[i]:=sum;
            END;
          sum:=0;
          FOR k:=1 TO r DO
            sum:=sum+soma[k]*alfa[k];
          sum:=-sum/2+ln(pp1/pp2);
          constante:=sum;
          FOR i:=1 TO n+no DO
            BEGIN
              sum:=0;
              FOR k:=1 TO r DO
                sum:=sum+dados[i,k]*alfa[k];
            m_n[i]:=sum;
            END;
          FOR k:=1 TO n DO
            BEGIN

pr_est[k]:=exp(m_n[k]+constante)/(1+exp(m_n[k]+constante));
              WRITELN(sa,k,'    Pr[pop 1/x] =
',pr_est[k]:1:5);
              IF pr_est[k]-0.5 > 0
              THEN
                pop_est[k]:=1
              ELSE
                pop_est[k]:=2;

conf[pop[k],pop_est[k]]:=conf[pop[k],pop_est[k]]+1;
              WRITELN('pop[',k,'] = ',pop[k],'
pop_est[',k,'] = ',pop_est[k]);
            END;
          FOR k:=n+1 TO n+no DO
            BEGIN

pr_est[k]:=exp(m_n[k]+constante)/(1+exp(m_n[k]+constante));
              WRITELN(sa,pr_est[k]:1:8,'        ',pop[k]);
              IF pr_est[k]-0.5 > 0
              THEN
                pop_est2[k-n]:=1
              ELSE
                pop_est2[k-n]:=2;
            END;
          WRITELN(sa);
          WRITELN(sa);
          WRITELN(sa,'    parametros estimados:');
          WRITELN(sa);
          WRITELN(sa,'constante = ',constante);
          FOR i:=1 TO r DO
            WRITELN(sa,'    alfa',i,' = ',alfa[i]);
          WRITELN(sa);
          WRITELN(sa,'    regra de classificacao:');
          WRITELN(sa,' se alfa`x > -alfa0');
          WRITELN(sa,' entao x e classificado na populacao
1');
```

```
          WRITELN(sa,' cc x e classificado na populacao 2');
          WRITELN(sa);
          writeln(sa);
        END;

{-----------------------------------------------------------}

        PROCEDURE qdf;

        VAR     i,j,k,l                          :INTEGER;
                dets1,dets2,alfa1,alfa2,sum      :REAL;
                media1,media2,x1,x2,
                x1t,x2t,m_rx1,m_rx2              :mr;
                s1,s2,invs1,invs2                :mrr;

        BEGIN
          FILLCHAR(media1,SIZEOF(media1),0);
          FILLCHAR(media2,SIZEOF(media2),0);
          WRITELN(sa,'                    FUNCAO DISCRIMINANTE
QUADRATICA');

          FOR i:=1 TO n1 DO
            FOR j:=1 TO r DO
               media1[j]:=media1[j]+dados[i,j];
          FOR i:=n1+1 TO n DO
            FOR j:=1 TO r DO
               media2[j]:=media2[j]+dados[i,j];
          FOR j:=1 TO r DO
            BEGIN
              media1[j]:=media1[j]/n1;
              media2[j]:=media2[j]/n2;
            END;
          dispersao(dados,n1+1,n,1,r,s2);
          inversa(s2,r,invs2);
          dets2:=ABS(deter(s2,r));
          dispersao(dados,1,n1,1,r,s1);
          inversa(s1,r,invs1);
          dets1:=ABS(deter(s1,r));
          FOR k:=1 TO n DO
            BEGIN
              FOR j:=1 TO r DO
                BEGIN
                  x1[j]:=dados[k,j]-media1[j];
                  x2[j]:=dados[k,j]-media2[j];
                END;
              FOR j:=1 TO r DO
                BEGIN
                  alfa1:=0;
                  alfa2:=0;
                  FOR l:=1 TO r DO
                    BEGIN
                      alfa1:=alfa1+x1[l]*invs1[l,j];
                      alfa2:=alfa2+x2[l]*invs2[l,j];
                    END;
```

```
                m_rx1[j]:=alfa1;
                m_rx2[j]:=alfa2;
              END;
            alfa1:=0;
            alfa2:=0;
            FOR l:=1 TO r DO
              BEGIN
                alfa1:=alfa1+m_rx1[l]*x1[l];
                alfa2:=alfa2+m_rx2[l]*x2[l];
              END;
            alfa1:=EXP(alfa1*(-0.5))*pp1/SQRT(ABS(dets1));
            alfa2:=EXP(alfa2*(-0.5))*pp2/SQRT(ABS(dets2));
            pr_est[k]:=alfa1/(alfa1+alfa2);
            IF pr_est[k]-0.5 > 0
            THEN
              pop_est[k]:=1
            ELSE
              pop_est[k]:=2;
            WRITELN(sa,pr_est[k]:1:8);

conf[pop[k],pop_est[k]]:=conf[pop[k],pop_est[k]]+1;
            WRITELN('pop[',k,'] = ',pop[k],'
pop_est[',k,'] = ',pop_est[k]);
          END;
        FOR k:=n+1 TO n+no DO
          BEGIN
            FOR j:=1 TO r DO
              BEGIN
                x1[j]:=dados[k,j]-media1[j];
                x2[j]:=dados[k,j]-media2[j];
              END;
            FOR j:=1 TO r DO
              BEGIN
                alfa1:=0;
                alfa2:=0;
                FOR l:=1 TO r DO
                  BEGIN
                    alfa1:=alfa1+x1[l]*invs1[l,j];
                    alfa2:=alfa2+x2[l]*invs2[l,j];
                  END;
                m_rx1[j]:=alfa1;
                m_rx2[j]:=alfa2;
              END;
            alfa1:=0;
            alfa2:=0;
            FOR l:=1 TO r DO
              BEGIN
                alfa1:=alfa1+m_rx1[l]*x1[l];
                alfa2:=alfa2+m_rx2[l]*x2[l];
              END;
            alfa1:=EXP(alfa1*(-0.5))*pp1/SQRT(ABS(dets1));
            alfa2:=EXP(alfa2*(-0.5))*pp2/SQRT(ABS(dets2));
            pr_est[k]:=alfa1/(alfa1+alfa2);
            IF pr_est[k]-0.5 > 0
```

```
            THEN
              pop_est2[k-n]:=1
            ELSE
              pop_est2[k-n]:=2;
          END;
        END;
{-------------------------------------------------------------}

        PROCEDURE lm;

        CONST  _c_maxi = 64;

        VAR     i,j,k,l,h,t,c_max                :INTEGER;
                sum,prod,maxi,dets,m_1x1         :REAL;
                prob_est                :ARRAY[1..2] OF REAL;
                spooled,invs                     :mrr;
                m           :ARRAY[1..n_n+n_no] OF INTEGER;
                dy                               :mnr;
                som,my      :ARRAY[1..c_maxi,1..n_r] OF REAL;
                n_m         :ARRAY[1..c_maxi] OF INTEGER;
                pr_m        :ARRAY[1..c_maxi] OF REAL;
                y1,m_px1                         :mr;
                pp                      :ARRAY[1..2] OF REAL;

        BEGIN
          FILLCHAR(som,SIZEOF(som),0);
          FILLCHAR(n_m,SIZEOF(som),0);
          FILLCHAR(spooled,SIZEOF(spooled),0);
          WRITELN(sa,'            MODELO DE KRZANOWSKI -
location model');
          FOR k:=1 TO n+no DO
            BEGIN
              sum:=0;
              FOR j:=2 TO q DO
                BEGIN
                  prod:=1;
                  FOR h:=1 TO j-1 DO
                    prod:=prod*c[h];
                  sum:=sum+prod*dados[k,j];
                END;
              m[k]:=1+TRUNC(dados[k,1]+sum);
            END;
          c_max:=1;
          FOR j:=1 TO q DO
            c_max:=c_max*c[j];
          FOR k:=1 TO n1 DO
            BEGIN
              i:=m[k];
              FOR j:=q+1 TO r DO
                som[i,j]:=som[i,j]+dados[k,j];
              INC(n_m[i]);
            END;
          FOR k:=n1+1 TO n DO
            BEGIN
```

```
            i:=m[k]+c_max;
            FOR j:=q+1 TO r DO
              som[i,j]:=som[i,j]+dados[k,j];
            n_m[i]:=n_m[i]+1;
          END;
        WRITELN(sa);
        WRITELN(sa,'   tabela de frequencias por casela');
        WRITELN(sa);
        WRITELN(sa,'casela       freq pop 1      freq pop
2');
        FOR i:=1 TO c_max DO
          WRITELN(sa,i:10,n_m[i]:10,n_m[i+c_max]:10);
        WRITELN(sa);
        FOR l:=1 TO c_max DO
          pr_m[l]:=n_m[l]/n1;
        FOR l:=1 TO c_max DO
          pr_m[c_max+l]:=n_m[c_max+l]/n2;
        FOR l:=1 TO 2*c_max DO
          IF n_m[l] <> 0
          THEN
            FOR j:=q+1 TO r DO
              my[l,j]:=som[l,j]/n_m[l];
        {inserir futuramente impressao das matriz de
medias}
        FOR k:=1 TO n DO
          BEGIN
            l:=c_max*(pop[k]-1)+m[k];
            FOR j:=q+1 TO r DO
              dy[k,j-q]:=dados[k,j]-my[l,j];
            FOR i:=1 TO p DO
              FOR j:=1 TO p DO

spooled[i,j]:=spooled[i,j]+dy[k,i]*dy[k,j];
          END;
        FOR i:=1 TO p DO
          FOR j:=1 TO p DO
            spooled[i,j]:=spooled[i,j]/(n-2*c_max);
        dets:=deter(spooled,p);
        inversa(spooled,p,invs);
        pp[1]:=pp1;
        pp[2]:=pp2;
        FOR k:=1 TO n DO
          BEGIN
            maxi:=0;
            FOR i:=1 TO 2 DO
              BEGIN
                l:=c_max*(i-1)+m[k];
                FOR j:=q+1 TO r DO
                  y1[j-q]:=dados[k,j]-my[l,j];
                FOR j:=1 TO p DO
                  BEGIN
                    sum:=0;
                    FOR t:=1 TO p DO
                      sum:=sum+y1[t]*invs[t,j];
```

```
                    m_px1[j]:=sum;
                  END;
                sum:=0;
                FOR t:=1 TO p DO
                  sum:=sum+m_px1[t]*y1[t];
                m_1x1:=sum;
                prob_est[i]:=pr_m[1]*pp[i]*EXP(-m_1x1/2);
              END;

pr_est[k]:=prob_est[1]/(prob_est[1]+prob_est[2]);
            IF pr_est[k]-0.5 > 0
            THEN pop_est[k]:=1
            ELSE pop_est[k]:=2;

conf[pop[k],pop_est[k]]:=conf[pop[k],pop_est[k]]+1;
            WRITELN(sa,k,'    Pr[pop 1/x] =
',pr_est[k]:1:5);
          END;
        FOR k:=n+1 TO n+no DO
          BEGIN
            maxi:=0;
            FOR i:=1 TO 2 DO
              BEGIN
                l:=c_max*(i-1)+m[k];
                FOR j:=q+1 TO r DO
                  y1[j-q]:=dados[k,j]-my[l,j];
                FOR j:=1 TO p DO
                  BEGIN
                    sum:=0;
                    FOR t:=1 TO p DO
                      sum:=sum+y1[t]*invs[t,j];
                    m_px1[j]:=sum;
                  END;
                sum:=0;
                FOR t:=1 TO p DO
                  sum:=sum+m_px1[t]*y1[t];
                m_1x1:=sum;
                prob_est[i]:=pr_m[1]*pp[i]*EXP(-m_1x1/2);
              END;

pr_est[k]:=prob_est[1]/(prob_est[1]+prob_est[2]);
            IF pr_est[k]-0.5 > 0
            THEN pop_est2[k-n]:=1
            ELSE pop_est2[k-n]:=2;
            WRITELN(sa,k,'    Pr[pop 1/x] =
',pr_est[k]:1:5);
          END;
      END;

{-----------------------------------------------------------}

      PROCEDURE logistic;

      VAR   b,h,m,i,j,k,cont,lin,col,l,inf,
```

```
              sup,n_it,n_it_max,ind_l,v        :INTEGER;
              maxi,sum,sum2,epsilon            :REAL;
              m_rxl,score                      :mr;
              pr_i_k       :ARRAY[1..n_n,1..2] OF REAL;
              mh,invmh                         :morr;
              param        :ARRAY[1..n_r,1..2] OF REAL;
              d           :ARRAY[1..n_n+n_no,1..n_r] OF REAL;
              prob_est              :ARRAY[1..2] OF REAL;

        BEGIN
          WRITELN(sa,'
DISCRIMINACAO LOGISTICA');
          WRITELN(sa,'
========================');
          WRITELN(sa);
          WRITELN(sa);
          v:=r+1;
          n_it_max:=8;
          epsilon:=1E-20;
          FILLCHAR(param,SIZEOF(param),0);
          FOR k:=1 TO n+no DO
            BEGIN
              FOR j:=r+1 DOWNTO 2 DO
                d[k,j]:=dados[k,j-1];
              d[k,1]:=1;
            END;
          maxi:=epsilon+1;
          n_it:=1;
          WHILE (n_it<n_it_max) AND (maxi>epsilon) DO
            BEGIN
              FOR j:=1 TO v DO
                param[j,1]:=param[j,2];
              { calculo de pr[ POP(i) | X(k) ]  }
              FOR k:=1 TO n DO
                BEGIN
                  sum:=0;
                  FOR cont:=1 TO r+1 DO
                    sum:=sum+d[k,cont]*param[cont,1];
                  pr_i_k[k,1]:=EXP(sum)/(EXP(sum)+1);
                  pr_i_k[k,2]:=1-pr_i_k[k,1];
                END;
              { calculo da funcao "score" }
              FOR j:=0 TO r DO
                BEGIN
                  ind_l:=j+1;
                  sum2:=0;
                  FOR k:=1 TO n DO
                    sum2:=sum2+d[k,j+1]*pr_i_k[k,1];
                  sum:=0;
                  FOR k:=1 TO n1 DO
                    sum:=sum+d[k,j+1];
                  score[ind_l]:=sum-sum2;
                END;
              { calculo da matriz hessiana }
```

```
                FOR j:=0 TO r DO
                  FOR m:=0 TO r DO
                    BEGIN
                      sum:=0;
                      FOR k:=1 TO n DO
                        sum:=sum-pr_i_k[k,1]*(1-
pr_i_k[k,1])*d[k,j+1]*d[k,m+1];
                      mh[j+1,m+1]:=sum;
                    END;
                { algoritmo de Newton-Raphson }
                WRITELN(sa);
                inversa_log(mh,v,invmh);
                FOR i:=1 TO v DO
                  BEGIN
                    sum:=0;
                    FOR cont:=1 TO v DO
                      sum:=sum+invmh[i,cont]*score[cont];
                    m_rx1[i]:=sum;
                  END;
                maxi:=0;
                FOR cont:=1 TO v DO
                  BEGIN
                    param[cont,2]:=param[cont,1]-m_rx1[cont];
                    maxi:=maxi+sqr(param[cont,2]-
param[cont,1]);
                  END;
                WRITELN(sa,'iteracao no. ',n_it,'  sqr',maxi);
                n_it:=n_it+1;
              END;
            { correcao para amostragem feita separadamente }
            IF separadas
            THEN
              param[1,2]:=param[1,2]+ln(pp1*n2/pp2/n1);
            WRITELN(sa,'PARAMETROS ESTIMADOS');
            FOR i:=1 TO 2 DO
              BEGIN
                WRITELN(sa,'pop:',i);
                inf:=(i-1)*(r+1)+1;
                sup:=i*(r+1);
                FOR cont:=inf TO sup DO
                  WRITELN(sa,param[cont,2]);
                WRITELN(sa);
              END;
            { classificacao para cada observacao amostral }
            FOR k:=1 TO n DO
              BEGIN
                maxi:=-1000;
                FOR i:=1 TO 2 DO
                  BEGIN
                    inf:=(i-1)*(r+1)+1;
                    sup:=i*(r+1);
                    sum:=0;
                    FOR cont:=1 TO r+1 DO
                      sum:=sum+d[k,cont]*param[inf+cont-1,2];
```

```
                  prob_est[i]:=EXP(sum);
                END;

pr_est[k]:=prob_est[1]/(prob_est[1]+prob_est[2]);
              IF pr_est[k]-0.5 > 0
              THEN pop_est[k]:=1
              ELSE pop_est[k]:=2;
              WRITELN(sa,k,'    Pr[pop 1/x] =
',pr_est[k]:1:5);

conf[pop[k],pop_est[k]]:=conf[pop[k],pop_est[k]]+1;
              WRITELN('pop[',k,'] = ',pop[k],'
pop_est[',k,'] = ',pop_est[k]);
            END;
          FOR k:=n+1 TO n+no DO
            BEGIN
              maxi:=-1000;
              FOR i:=1 TO 2 DO
                BEGIN
                  inf:=(i-1)*(r+1)+1;
                  sup:=i*(r+1);
                  sum:=0;
                  FOR cont:=1 TO r+1 DO
                    sum:=sum+d[k,cont]*param[inf+cont-1,2];
                  prob_est[i]:=EXP(sum);
                END;

pr_est[k]:=prob_est[1]/(prob_est[1]+prob_est[2]);
              IF pr_est[k]-0.5 > 0
              THEN pop_est[k]:=1
              ELSE pop_est[k]:=2;
            END;
        END;

{-------------------------------------------------------------}

        PROCEDURE kernel;

        VAR   min_c,i,j,k,l,inf,sup              :INTEGER;
              a,b,delta,lambda,s1,s2,dp,kp,kq    :REAL;
              m1_qx1,m1_px1,m2_qx1,m2_px1        :mr;
              param                 :ARRAY[1..2,1..2] OF REAL;
              pr                    :ARRAY[1..2] OF REAL;
              cat                                :BOOLEAN;

        BEGIN
          WRITELN(sa,'                                  O METODO
DO NUCLEO');
          WRITELN(sa,'
==================');
          WRITELN(sa);
          WRITELN(sa);
          min_c:=c[1];
          FOR j:=2 TO q DO
```

```
            IF min_c > c[j]
            THEN min_c:=c[j];
          inf:=1;
          sup:=n1;
          FOR i:=1 TO 2 DO
            BEGIN
              FOR j:=q+1 TO r DO
                BEGIN
                  s1:=0;
                  s2:=0;
                  FOR k:=inf TO sup DO
                    BEGIN
                      s1:=s1+dados[k,j];
                      s2:=s2+dados[k,j]*dados[k,j];
                    END;
                  dp:=SQRT(s2/(sup-inf+1)-SQR(s1/(sup-
inf+1)));
                  FOR k:=inf TO sup DO
                    dados[k,j]:=dados[k,j]/dp;
                END;
              inf:=n1+1;
              sup:=n;
            END;
          inf:=1;
          sup:=n1;
          FOR i:=1 TO 2 DO
            BEGIN
              WRITELN(sa,'i = ',i);
              cat:=TRUE;
              a:=1/min_c;
              b:=1;
              WHILE ABS(a-b) > 0.01 DO
                ache_max(a,b,cat,inf,sup);
              param[i,1]:=(a+b)/2;
              cat:=FALSE;
              a:=0;
              b:=1;
              WHILE ABS(a-b) > 0.01 DO
                ache_max(a,b,cat,inf,sup);
              param[i,2]:=(a+b)/2;
              inf:=n1+1;
              sup:=n;
            END;
          WRITELN(sa);
          WRITELN(sa, 'PARAMETROS ESTIMADOS');
          WRITELN(sa);
          WRITELN(sa,'pop   lambda   delta');
          FOR i:=1 TO 2 DO
            WRITELN(sa,' ',i,'     ',param[i,2]:4:2,'
',param[i,1]:4:2);
          WRITELN(sa);
          FOR k:=1 TO n DO
            BEGIN
              FILLCHAR(pr,SIZEOF(pr),0);
```

```
                FOR j:=1 TO q DO
                  m1_qx1[j]:=dados[k,j];
                FOR j:=q+1 TO r DO
                  m1_px1[(j-q)]:=dados[k,j];
                inf:=1;
                sup:=n1;
                FOR i:=1 TO 2 DO
                  BEGIN
                    delta:=param[i,1];
                    lambda:=param[i,2];
                    a:=0;
                    FOR l:=inf TO sup DO
                      BEGIN
                        IF l <> k
                        THEN
                          BEGIN
                            FOR j:=1 TO q DO
                              m2_qx1[j]:=dados[l,j];
                            FOR j:=q+1 TO r DO
                              m2_px1[(j-q)]:=dados[l,j];
                            kq:=k_q(m1_qx1,m2_qx1,delta);
                            kp:=k_p(m1_px1,m2_px1,lambda);
                            pr[i]:=pr[i]+kq*kp;
                          END
                        ELSE a:=1;
                      END;
                    pr[i]:=pr[i]/(sup-inf+1-a);
                    inf:=n1+1;
                    sup:=n;
                  END;
                pr_est[k]:=pr[1]*pp1/(pr[1]*pp1+pr[2]*pp2);
                IF pr_est[k]-0.5 > 0
                THEN pop_est[k]:=1
                ELSE pop_est[k]:=2;
                WRITELN(sa,k,'    Pr[pop 1/x] =
',pr_est[k]:1:5);
                WRITELN('pop[',k,'] = ',pop[k],'
pop_est[',k,'] = ',pop_est[k]);

conf[pop[k],pop_est[k]]:=conf[pop[k],pop_est[k]]+1;
              END;
            FOR k:=n+1 TO n+no DO
              BEGIN
                FILLCHAR(pr,SIZEOF(pr),0);
                FOR j:=1 TO q DO
                  m1_qx1[j]:=dados[k,j];
                FOR j:=q+1 TO r DO
                  m1_px1[(j-q)]:=dados[k,j];
                inf:=1;
                sup:=n1;
                FOR i:=1 TO 2 DO
                  BEGIN
                    delta:=param[i,1];
                    lambda:=param[i,2];
```

```
                a:=0;
                FOR l:=inf TO sup DO
                  BEGIN
                    IF l <> k
                    THEN
                      BEGIN
                        FOR j:=1 TO q DO
                          m2_qxl[j]:=dados[l,j];
                        FOR j:=q+1 TO r DO
                          m2_pxl[(j-q)]:=dados[l,j];
                        kq:=k_q(m1_qxl,m2_qxl,delta);
                        kp:=k_p(m1_pxl,m2_pxl,lambda);
                        pr[i]:=pr[i]+kq*kp;
                      END
                    ELSE a:=1;
                  END;
                pr[i]:=pr[i]/(sup-inf+1-a);
                inf:=nl+1;
                sup:=n;
              END;
            pr_est[k]:=pr[1]*pp1/(pr[1]*pp1+pr[2]*pp2);
            IF pr_est[k]-0.5 > 0
            THEN pop_est[k]:=1
            ELSE pop_est[k]:=2;
          END;
        FOR j:=q+1 TO r DO
          FOR k:=1 TO n DO
            dados[k,j]:=dados[k,j]*dp;
      END;

{-------------------------------------------------------------}

      PROCEDURE knn;

      VAR   i,j,k,l,aux,lin,linha,
            col,knn,inf,sup                    :INTEGER;
            vol,soma,rlc,mini,e                :REAL;
            rd                  :ARRAY[1..n_r] OF REAL;
            d_aux                              :mnr;
            ki                      :ARRAY[1..2] OF REAL;
            dist                               :mn;
            popu                               :mn_i;
            knn_ok                             :BOOLEAN;

      BEGIN
        WRITELN(sa,'                    METODO DOS K-
VIZINHOS MAIS PROXIMOS');
        WRITELN(sa,'
==================================');
        WRITELN(sa);
        WRITELN(sa);
        knn_ok:=FALSE;
        WHILE NOT knn_ok DO
          BEGIN
```

```
                WRITELN('DIGITE O VALOR PARA O K, ISTO E,
QUANTAS OBSERVACOES');
                WRITELN('PROXIMAS A K SERAO CONSIDERADAS :');
                READLN(knn);
                IF ((knn MOD 2) = 0)
                THEN
                  WRITELN('K PRECISA SER IMPAR')
                ELSE
                  knn_ok:=TRUE;
              END;
            FOR j:=1 TO q DO
              rd[j]:=dados[1,j];
            FOR j:=q+1 TO r DO
              rd[j]:=maxi2[j]-mini2[j];
            (* classificacao para cada observacao amostral *)
            FOR k:=1 TO n DO
              BEGIN
                FOR lin:=1 TO n DO
                  (* calculo da similaridade entre x[lin] e
x[k]  *)
                  BEGIN
                    soma:=0;
                    FOR j:=1 TO q DO
                      IF dados[lin,j]=dados[k,j]
                      THEN soma:=soma+1;
                    FOR j:=q+1 TO r DO
                      soma:=soma+1-ABS((dados[lin,j]-
dados[k,j])/rd[j]);
                    rlc:=soma/r;
                    dist[lin]:=2*(1-rlc);
                  END;
                FILLCHAR(ki,SIZEOF(ki),0);
                FOR lin:=1 TO n DO
                 popu[lin]:=pop[lin];
                bolha(dist,popu,knn);
                FOR lin:=2 TO knn+1 DO
                  ki[popu[lin]]:=ki[popu[lin]]+1;

pr_est[k]:=ki[1]*pp1/n1/(ki[1]*pp1/n1+ki[2]*pp2/n2);
                IF pr_est[k]-0.5 > 0
                THEN pop_est[k]:=1
                ELSE pop_est[k]:=2;
                WRITELN(sa,k,'    Pr[pop 1/x] =
',pr_est[k]:1:5);

conf[pop[k],pop_est[k]]:=conf[pop[k],pop_est[k]]+1;
                WRITELN('pop[',k,'] = ',pop[k],'
pop_est[',k,'] = ',pop_est[k]);
              END;
            FOR k:=n+1 TO n+no DO
              BEGIN
                FOR lin:=1 TO n DO
                  (* calculo da similaridade entre x[lin] e
x[k]  *)
```

```
                BEGIN
                  soma:=0;
                  FOR j:=1 TO q DO
                    IF dados[lin,j]=dados[k,j]
                    THEN soma:=soma+1;
                  FOR j:=q+1 TO r DO
                    soma:=soma+1-ABS((dados[lin,j]-
dados[k,j])/rd[j]);
                  rlc:=soma/r;
                  dist[lin]:=2*(1-rlc);
                END;
              FOR lin:=1 TO 2 DO
                ki[lin]:=0;
              FOR lin:=1 TO n DO
               popu[lin]:=pop[lin];
              bolha(dist,popu,knn);
              FOR lin:=2 TO knn+1 DO
                ki[popu[lin]]:=(ki[popu[lin]])+1;
              pr_est[k]:=ki[1]*n*pp1/knn/n1;
              IF pr_est[k]-0.5 > 0
              THEN pop_est[k]:=1
              ELSE pop_est[k]:=2;
            END;
        END;

{-----------------------------------------------------------}

                    PROGRAMA PRINCIPAL

{-----------------------------------------------------------}

        BEGIN
          leu_dados:=FALSE;
          opcao:='1';
          WHILE opcao <> '9' DO
            BEGIN
              FILLCHAR(conf,SIZEOF(conf),0);
              FILLCHAR(pr_est,SIZEOF(pr_est),0);
              CLRSCR;
              tela_menu;
              WRITE  ('                    Escolha a opcao
desejada - ');
              opcao:=READKEY;
              WRITE(opcao);
              CLRSCR;
              CASE opcao OF
                '1': informacoes_1;
                '2': BEGIN
                       le_dados;
                       leu_dados:=TRUE;
                       WRITE('PRESSIONE ENTER PARA
PROSSEGUIR');
                       exit:=READKEY;
                     END;
```

```
                '3': BEGIN
                       IF leu_dados
                       THEN ldf
                       ELSE informacoes_2;
                     END;
                '4': BEGIN
                       IF leu_dados
                       THEN qdf
                       ELSE informacoes_2;
                     END;
                '5': BEGIN
                       IF leu_dados
                       THEN lm
                       ELSE informacoes_2;
                     END;
                '6': BEGIN
                       IF leu_dados
                       THEN logistic
                       ELSE informacoes_2;
                     END;
                '7': BEGIN
                       IF leu_dados
                       THEN kernel
                       ELSE informacoes_2;
                     END;
                '8': BEGIN
                       IF leu_dados
                       THEN knn
                       ELSE informacoes_2;
                     END;
                '9': BEGIN
                       WRITE('deseja realmente sair ?
(s/n)');
                       exit:=READKEY;
                       IF (exit <> 's') AND (exit <> 'S')
                       THEN opcao:='1';
                     END;
              END;
            IF (opcao IN ['3'..'8']) AND leu_dados
            THEN
              BEGIN
                resultados;
                WRITE('PRESSIONE ENTER PARA PROSSEGUIR');
                exit:=READKEY;
              END;
          END;
        IF leu_dados
        THEN CLOSE(sa);
        WRITELN('FIM DO AD');
      END.
```

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Aitchison, J. e Aitken, C. G. G. (1976) Multivariate Binary Discrimination by the Kernel Method. *Biometrika*, **63**: 413–420.

[2] Albert, A. e Anderson, J. A. (1984) On the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*. **71**: 1–10.

[3] Albert, A. e Lesaffre, E. (1986) Multiple Group Logistic Discrimination. *Comp. and Maths. with Appls.*, **vol.12A, nº2**: 209–224.

[4] Anderson, J. A. e Blair, V. (1982) Penalized Maximum Likelihood Estimation in Logistic Regression and Discrimination. *Biometrika*, **69**: 123–136.

[5] Anderson, J. A. e Richardson, S. C. (1979) Logistic Discrimination and Bias Correction in Maximum Likelihood Estimation. *Technometrics*, **21**: 71–78.

[6] Anderson, J. A. (1972) Separate Sample Logistic Discrimination. *Biometrika*, **59**: 19–35.

[7] Anderson, J. A. (1982) *Logistic Discrimination in Handbook of Statistics 2.* P. R. Krishnaiah and L. N. Kanal North-Holland Publishing Company: 161–191.

[8] Bishop, Y. M. M.; Fienberg, S. E. e Holland, P. W. (1975) *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, MIT Press.

[9] Breiman, L.; Friedman, J. H.; Olshen, R. A. e Stone, C. J. (1984) *Classification and Regression Trees.* Internacional, Belmont.

[10] Bull, S. B. e Donner, A. (1987) The Efficiency of Multinomial Logistic Regression Compared with Multiple Group Discriminant Analysis. *JASA*, **82**: 1118–1122.

[11] Chang, P. C. e Afifi, A. A. (1974) Classification Based on Dichotomous and Continuous Variables. *JASA*, **69**: 336–339.

[12] Choi, S. C. (1986) Discrimination and Classification: Overview. *Comp. and Maths. with Appls.*, **vol.12A, nº2**: 173–177.

[13] Cochran, W. G. e Hopkins, C. E. (1961) Some Classification Problems with Multivariate Qualitative Data. *Biometrics*, **7**: 10–32.

[14] Copas, J. B. (1983) Plotting p Against x. *Appl. Statist.*, **32**: 25–31.

[15] Cox, D. R. (1966) Some Procedures Associated with the Logistic Qualitative Response Curve. *In Research Papers in Statistics: Fetschrift for J. Neyman ed. F. N. David.* Wiley, New York: 55–71.

[16] Cox, D. R. (1970) *The Analysis of Binary Data.* Methuen & Co. LTd., London.

[17] Daudin, J. J. (1986) Selection of Variables in Mixed-Variable Discriminant Analysis. *Biometrics*, **42**: 473–481.

[18] Day, N. E. e Kerridge, D. F. (1967) A General Maximum Likelihood Discriminant. *Biometrics*, **23**: 313–323.

[19] Dixon, W. J. e Brown, M. C. (1979) *Biomedical Computer Programs P-Series (BMDP).* University of California Press, Berkeley.

[20] Efron, B. (1975) The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis. *JASA*, **70**: 892–898.

[21] Efron, B. (1986) How Biased is the Apparent Error Rate of a Prediction Rule? *JASA*, **81**: 461–470.

[22] Enas, G. G. e Choi, S. C. (1986) Choice of the Smoothing Parameter and Efficiency of k-Nearest Neighbor Classification. *Comp. and Maths. with Appls.*, **vol.12A, nº2**: 235–244.

[23] Farewell, V. T. (1979) Some Results on the Estimation of Logistic Models Based on Retrospective Data. *Biometrika*, **66**: 27–32.

[24] Fisher, R. A. (1936) The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, **7**: 179–188.

[25] Fix, E. e Hodges, J. (1951) Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties. *Technical report*. Randolph Field, Texas: USAF School of Aviation Medicine.

[26] Gill, P. E. e Murray, W. (1972) Quasi-Newton Methods for Unconstrained Optimization. *J. Inst. Math. Appl.*, **9**: 91–108.

[27] Goldstein, M. e Dillon, W. R. (1978) *Discrete Discriminant Analysis*. Wiley.

[28] Gower, J. C. (1971) A General Coefficient o f Similarity and Some of its Properties. *Biometrics*, **27**: 857–872.

[29] Habbema, J. D. F., Hermans, J. e van den Broek, K. (1974) A Stepwise Discriminant Analysis Program Using Density Estimation. *Compstat. Ed. G. Bruckman*. Physica-Verlag, Vienna: 101–110.

[30] Haberman, S. J. (1974) *The Analysis of Frequency Data*. Chicago Press.

[31] Hand, D. J. (1981) *Discrimination and Classification*. Wiley.

[32] Hills, M. (1967) Discrimination and Allocation with Discrete Data. *Applied Statistics*, **16**: 237–250.

[33] Hosmer, D. W. e Lemeshow, S. (1980) Goodness of Fit Tests for the Multiple Logistic Regression Model. *Commun. Statist. Theor. Meth.*, **A9(10)**: 1043–1069.

[34] Johnson, R. A. e Wichern, D. W. (1982) *Applied Multivariate Statistical Analysis*. Prentice-Hall.

[35] Kay, R. e Little, S. (1987) Transformations of the Explanatory Variables in the Logistic Regression Model for Binary Data. *Biometrika*, **74**: 495–501.

[36] Knoke, D e Burke, P. J. (1980) *Log-linear models.* Series: Quantitative Applications in the Social Sciences. Sage University Papers, Beverly Hills, London.

[37] Knoke, J. D. (1982) Discriminant Analysis with Discrete and Continuous Variables. *Biometrics*, **38**: 191–200.

[38] Krzanowski, W. J. (1975) Discrimination and Classification Using Both Binary and Continuous Variables. *JASA*, **70**: 782–790.

[39] Krzanowski, W. J. (1977) The Performance of Fisher's Linear Discriminant Function under Non-Optimal Conditions. *Technometrics*, **19**: 191–200.

[40] Krzanowski, W. J. (1980) Mixtures of Continuous and Categorical Variables in Discriminant Analysis. *Biometrics*, **36**: 493-499.

[41] Krzanowski, W. J. (1982) Mixtures of Continuous and Categorical Variables in Discriminant Analysis: A Hypothesis-Testing Approach. *Biometrics*, **38**: 991–1002.

[42] Krzanowski, W. J. (1983-A) Distance between Populations Using Mixed Continuous and Categorical Variables. *Biometrika*, **70**: 235–243.

[43] Krzanowski, W. J. (1983-B) Stepwise Location Model Choice in Mixed-Variable Discrimination. *Appl. Statist.*, **32**, nº**3**: 260–266.

[44] Krzanowski, W. J. (1986) Multiple Discriminant Analysis in the Presence of Mixed Continuous and Categorical Data. *Comp. and Math. with Appls.*, **vol.12A** , nº**2**: 179–185.

[45] Lachenbruch, P. A. (1975) *Discriminant Analysis.* Hafner, New York.

[46] Lachenbruch, P. A. e Mickey, M. R. (1968) Estimation of Error Rates in Discriminant Analysis. *Technometrics*, **10**: 1–11.

[47] Landwehr, J. M.; Pregibon D. e Shoemaker, A. C. (1984) Graphical Methods for Assessing Logistic Regression Models. *JASA*, **79**: 61–83.

[48] Mardia, K. V.; Kent, J. T. e Bibby, J. M. (1979) *Multivariate Analysis*. Academic Press.

[49] Matusita, K. (1972) Discrimination and the Affinity of Distributions in Discriminant Analysis and Applications. *T. Cacoullos*. Academic, New York: 213–223.

[50] Morgan, J. N. e Sonquist, J. A. (1963) Problems in the Analysis of Survey Data, and a Proposal. *JASA*, **58**: 415–434.

[51] Murphy, B. J. e Moran, M. A. (1986) Parametric and Kernel Density Methods in Discriminant Analysis: Another Comparison. *Comp. and Maths. with Appls.*, **vol.12A**, nº**2**: 197–207.

[52] Olkin, I e Tate, R. F. (1961) Multivariate Correlation Models with Mixed Discrete and Continuous Variables. *Ann. Math. Stat.*, **32**: 448–465.

[53] Parzen, E. (1962) On Estimation of a Probability Density Function and Model. *Ann. Math. Stat.*, **33**: 1065–1076.

[54] Peduzzi, P. N.; Hardy, R. J. e Holford, T. R. (1980) A Stepwise Variable Selection Procedure for Non-Linear Regression Models. *Biometrics*, **36**: 551–516.

[55] Prentice, R. L. e Pyke, R. (1979) Logistic Disease Incidence Models and Case-Control Studies. *Biometrika*, **66**: 403–411.

[56] Press, S. J. e Wilson, S. (1978) Choosing between Logistic Regression and Discriminant Analysis. *JASA*, **73**: 699–705.

[57] Rao, C. R. (1973) *Linear Statistical Inference and Its Applications*. Wiley, 2ª ed.

[58] Reis, I. M. (1989) *Modelos Log-Lineares para Análise de Dados Categóricos*. São Paulo. Tese de Mestrado - Instituto de Matemática e Estatística da Universidade de São Paulo.

[59] *SAS User's Guide: Statistical Analysis System* (1979). SAS Institute Inc., Cary, North Carolina.

[60] Santner, T. J. e Duffy, D. E. (1986) A Note on A. Albert and J. A. Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models. *Biometrika*, **73**: 755–758.

[61] Seber, G. A. F. (1984) *Multivariate observations.* Wiley.

[62] Silva, J. P. (1988) *Análise e Decisão de Crédito.* Atlas.

[63] Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

[64] Titterington, D. M.; Murray, G. D.; Murray, L. S.; Spiegelhalter, D. J.; Skene, A. M.; Habbema, J. D. F. e Gelpke, G. J. (1981) Comparison of Discrimination Techniques Applied to a Complex Data Set of Head Injured Patients. *J. R. Statist. Soc. A.*, **144**: 145–175.

[65] Truett, J.; Cornfield, J. e Kannel, W. (1967) A Multivariate Analysis of the Risk of Coronary Heart Disease in Framinghan. *J. Chron. Dis.*, **20**: 511–524.

[66] Vlachonikolis, I. G. e Marriott, F. H. C. (1982) Discrimination with Mixed Binary and Continuous Data. *Appl. Statist.*, **31**, nº**1**: 23–31.