Parte 5. TÉCNICAS DE DISCRIMINAÇÃO E DE CLASSIFICAÇÃO DE DADOS

Introdução

Técnicas estatísticas de análise baseadas em discriminantes são utilizadas normalmente para a separação de dados em diferentes grupos, a partir de um determinado grupo de dados experimentais. Baseiam-se na criação de um "discriminante", ou seja, um critério quantitativo para separar observações em diferentes grupos, com máxima distância entre si. São utilizadas normalmente com a finalidade de explorar conjuntos de dados.

Técnicas de classificação de dados consistem na aplicação de técnicas estatísticas para estabelecer regras, ou critérios para alocar uma dada observação em diferentes grupos, os quais são definidos previamente.

O texto a seguir apresenta primeiramente um exemplo ilustrativo da aplicação de um discriminante para um conjunto de dados. Em seguida, são apresentadas e ilustradas as técnicas de classificação.

Discriminação

Um exemplo de criação de discriminante para separar dados é apresentado a seguir, para um caso em que há duas populações de dados (dois grupos), apresentados na Tabela 1 e na Figura 1.

Tabela 1. Conjunto de dados bivariados, pertencentes a dois grupos

	Grupo 1			Grupo 2	
N.	X1	X2	N.	X1	X2
1	0.158	0.182	1	-0.012	-0.031
2	0.210	0.206	2	0.036	0.053
3	0.207	0.188	3	0.038	0.036
4	0.280	0.236	4	-0.063	-0.074
5	0.197	0.193	5	-0.054	-0.119
6	0.227	0.173	6	0.000	-0.005
7	0.148	0.196	7	0.005	0.039
8	0.254	0.212	8	0.091	0.122
9	0.079	0.147	9	-0.036	-0.072
10	0.149	0.128	10	0.045	0.064
11	0.200	0.150	11	-0.026	-0.024
12	0.187	0.191	12	0.016	0.026

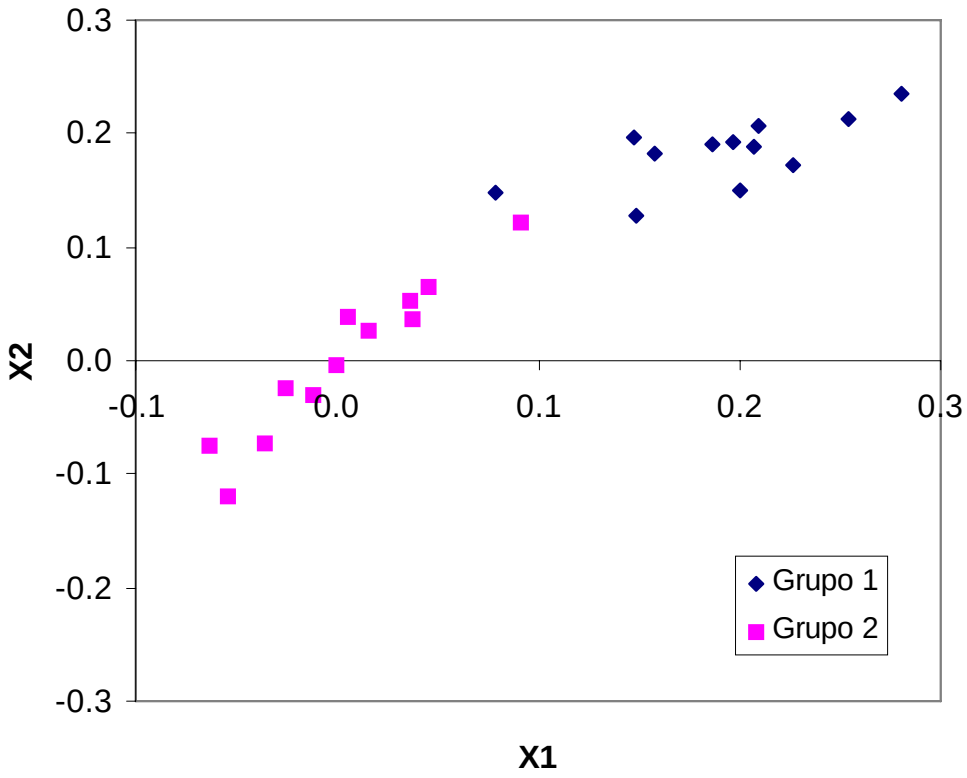

Figura 1. Representação gráfica dos dados da Tabela 1.

Podem ser aplicadas técnicas estatísticas convencionais para verificar se os dados podem ser considerados pertencentes a dois grupos. Por exemplo, pode-se utilizar teste de hipótese para a diferença entre as médias dos grupos:

$$H_0 : \mu_1 - \mu_2 = 0$$

$$\left(\bar{X}_{11} - \bar{X}_{12} \right) - t_\nu . s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \left(\mu_1 - \mu_2 \right) \leq \left(\bar{X}_{11} - \bar{X}_{12} \right) + t_\nu . s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

em que a variância combinada dos dois grupos é calculada por:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Na Tabela a seguir, são apresentados os valores de médias e desvios padrão da amostra de dados considerada.

	Grupo 1			Grupo 2			
N.	**X1**	**X2**	**N.**	**X1**	**X2**	**X11 - X12**	**X21-X22**
1	0.158	0.182	1	-0.012	-0.031	0.170	0.213
2	0.210	0.206	2	0.036	0.053	0.174	0.153
3	0.207	0.188	3	0.038	0.036	0.169	0.152
4	0.280	0.236	4	-0.063	-0.074	0.343	0.310
5	0.197	0.193	5	-0.054	-0.119	0.251	0.312
6	0.227	0.173	6	0.000	-0.005	0.227	0.178
7	0.148	0.196	7	0.005	0.039	0.143	0.157
8	0.254	0.212	8	0.091	0.122	0.163	0.090
9	0.079	0.147	9	-0.036	-0.072	0.115	0.219
10	0.149	0.128	10	0.045	0.064	0.104	0.064
11	0.200	0.150	11	-0.026	-0.024	0.226	0.174
12	0.187	0.191	12	0.016	0.026	0.171	0.165
média	**0.191**	**0.184**		**0.003**	**0.001**	**0.188**	**0.182**
s	**0.053**	**0.030**		**0.045**	**0.069**	**0.065**	**0.074**

Exemplo para X1:

t (0.95, ν = 22) = 2.074 e:

$$0.146 \leq \mu_1 - \mu_2 \leq 0.230$$

Portanto, com 95% de certeza pode-se afirmar que as médias entre os dois grupos são diferentes, e a hipótese nula é rejeitada.

Pode-se, também, construir um discriminante para os dois grupos, baseado em algum critério estatístico. Uma das técnicas, por exemplo, consiste na criação de novas variáveis discriminantes para os dados. Para os dados do exemplo, pode-se criar um discriminante que seja uma combinação linear das variáveis originais e utilizar um valor de corte para separar os dados:

$$d = w_1 x_1 + w_2 x_2 \, .$$

Neste caso, cada valor de *d* é o ponto em uma reta calculada com os pesos w_p, expressos como:

$$w_1 = \cos\theta \, ; \qquad w_1 = \text{sen}\,\theta$$

e θ é o ângulo de inclinação da reta em relação ao eixo da variável x_1, como mostrado na Figura 2.

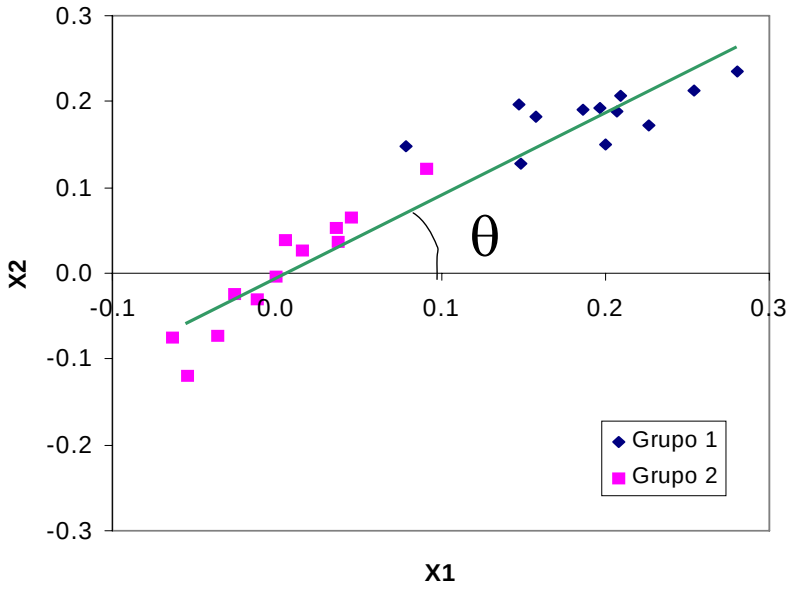

Figura 2. Representação gráfica de um discriminante linear para as variáveis x_1 e x_2.

Pode-se, assim, utilizar um valor limite, ou de corte, em *d*, para separar os dois grupos de dados. O valor do ângulo θ é calculado de modo a se maximizar a distância entre os dois grupos de dados e minimizar a distância entre os dados de um mesmo grupo. Para isso, utilizam-se as matrizes **B** e **W**, definidas a seguir para uma população de dados com *n* observações das *p* variáveis, x_p, e:

G grupos na população (*g* = 1,...,*G*);

n_g o número de observações no grupo *g*;

$\bar{\mathbf{x}}_{(px1)}$ o vetor das médias das *p* variáveis, para toda a população;

$\bar{\mathbf{x}}_{g(px1)}$ o vetor das médias das *p* variáveis no grupo *g*,

A matriz $\mathbf{B}_{(pxp)}$ é a soma ponderada dos quadrados das distâncias quadráticas entre grupos, é obtida pela Eq. 1:

$$\mathbf{B} = \sum_{g=1}^{G} n_g \left(\bar{\mathbf{x}}_g - \bar{\mathbf{x}}\right)\left(\bar{\mathbf{x}}_g - \bar{\mathbf{x}}\right)^T \tag{1}$$

Essa matriz é obtida fazendo-se o produto das diferenças entre vetores para cada grupo e, então, fazendo-se a soma ponderada para todos os grupos (soma de *G* matrizes *pxp*).

A matriz $\mathbf{W}_{(pxp)}$ é a soma das distâncias quadráticas entre cada observação e a média de todas observações em cada grupo, somada para todos os grupos, indicada pela Eq. 2:

$$\mathbf{W} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \left(\mathbf{x}_{i,g} - \bar{\mathbf{x}}_g\right)\left(\mathbf{x}_{i,g} - \bar{\mathbf{x}}_g\right)^T \tag{2}$$

A matriz $\mathbf{T}_{(pxp)}$ é a soma das distâncias quadráticas entre cada observação e a média de todas as observações:

$$\mathbf{T} = \sum_{i=1}^{n} \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)^T \tag{3}$$

Pode-se demonstrar que:

$$\mathbf{T} = \mathbf{W} + \mathbf{B} \tag{4}$$

Por exemplo, pode-se adotar o quociente entre ambas as matrizes, ou seja, o valor do produto $\mathbf{W}^{-1}\mathbf{B}$, como critério de seleção dos coeficientes do discriminante. Para os dados da Tabela 1, sejam, respectivamente, *SSb*, *SSw* e *SSt* as somas das distâncias quadráticas entre grupos, em cada grupo e a soma total, sendo: *SSt* = *SSb* + *SSw*. Sejam:

$$\lambda_1 = \frac{SSb}{SSw}; \quad \lambda_2 = \frac{SSb}{SSt}$$

Então, pode-se variar o ângulo θ de modo a selecionar o ângulo que maximize λ_1, ou λ_2, como ilustrado na Figura 3. No caso, o máximo ocorre para $\theta = 21^o$.

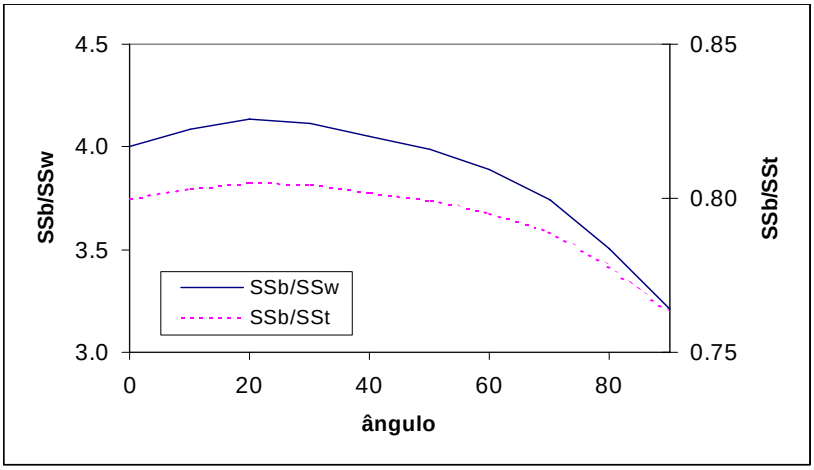

Figura 3. Variação dos quocientes entre distâncias quadráticas λ_1 e λ_2 em função do ângulo de inclinação da reta discriminante.

A Figura 4 mostra os valores do discriminante *d* que maximizam a relação entre distâncias quadráticas "entre grupos" e "internas aos grupos". Observa-se que há sobreposição entre as distribuições dos dados dos dois grupos, o que impossibilita separar os dados entre os grupos usando um discriminante linear. Este fato ilustra o fato de que a capacidade de discriminar os dados depende não apenas da distância entre os valores médios dos grupos, mas também da distribuição dos dados, como ilustrado na Figura 5.

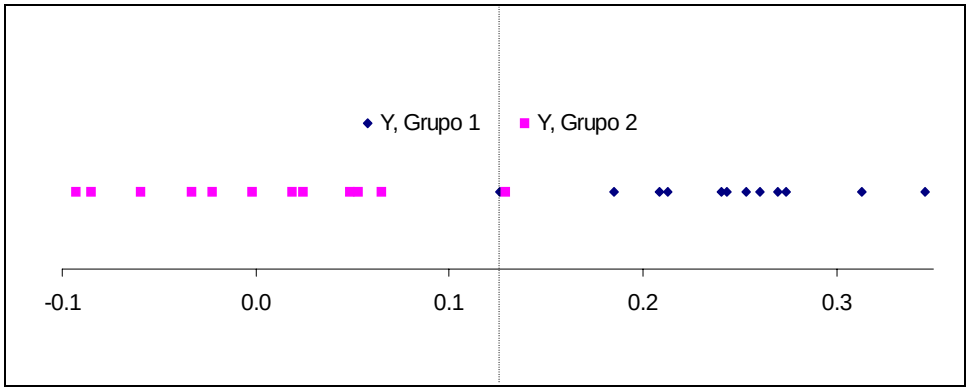

Figura 4. Valores do discriminante *d* para os dados dos dois grupos.

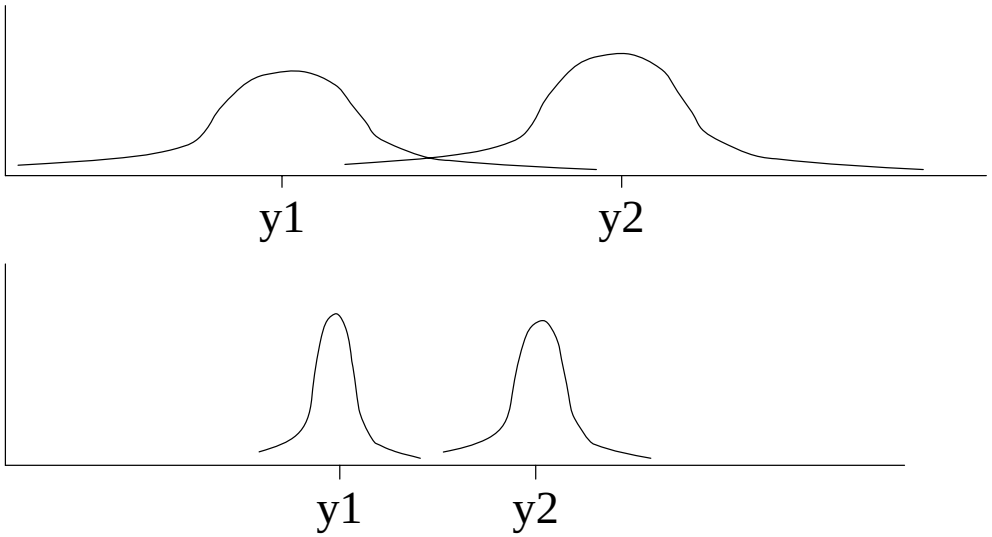

Figura 5. Superior: médias distantes, distribuições sobrepostas de dados; inferior: médias dos dois grupos próximas, com menor dispersão dos dados.

Discriminante linear de Fisher

Dado um conjunto de vetores centrados na média **X**, define-se o discriminante linear *Y* como:

$$Y = \mathbf{X}^T \boldsymbol{\gamma} \tag{4}$$

em que $\boldsymbol{\gamma}$ é um vetor de pesos determinado segundo o critério de máximo quociente entre a distância quadrática do discriminante "*entre grupos* : *dentro dos grupos*". O quadrado do discriminante é:

$$Y^2 = \left(\mathbf{X}^T \boldsymbol{\gamma}\right)^T \left(\mathbf{X}^T \boldsymbol{\gamma}\right) = \boldsymbol{\gamma}^T \mathbf{X}\mathbf{X}^T \boldsymbol{\gamma} \tag{5}$$

Somando-se para todas as *n* observações:

$$\sum_{i=1}^{n} Y^2 = \boldsymbol{\gamma}^T \left(\sum_{i=1}^{n} \mathbf{X}\mathbf{X}^T \right) \boldsymbol{\gamma} = \boldsymbol{\gamma}^T \mathbf{B} \boldsymbol{\gamma} + \boldsymbol{\gamma}^T \mathbf{W} \boldsymbol{\gamma} \tag{6}$$

Deve-se obter o discriminante maximizando-se o escalar:

$$\lambda = \frac{\boldsymbol{\gamma}^T \mathbf{B} \boldsymbol{\gamma}}{\boldsymbol{\gamma}^T \mathbf{W} \boldsymbol{\gamma}} \tag{7}$$

Para isso, calcula-se a derivada em relação ao vetor $\boldsymbol{\gamma}$:

$$\frac{\partial \lambda}{\partial \boldsymbol{\gamma}} = \frac{2\mathbf{B}\boldsymbol{\gamma}\left(\boldsymbol{\gamma}^T\mathbf{W}\boldsymbol{\gamma}\right) - \left(\boldsymbol{\gamma}^T\mathbf{B}\boldsymbol{\gamma}\right)2\mathbf{W}\boldsymbol{\gamma}}{\left(\boldsymbol{\gamma}^T\mathbf{W}\boldsymbol{\gamma}\right)^2} = 0 \qquad \text{(p x 1)} \qquad (8)$$

Usando-se a Eq. (7):

$$\mathbf{B}\boldsymbol{\gamma}\left(\boldsymbol{\gamma}^T\mathbf{W}\boldsymbol{\gamma}\right) - \lambda\left(\boldsymbol{\gamma}^T\mathbf{W}\boldsymbol{\gamma}\right)\mathbf{W}\boldsymbol{\gamma} = 0 \qquad (9)$$

Ou:

$$\mathbf{B}\boldsymbol{\gamma} - \lambda\mathbf{W}\boldsymbol{\gamma} = 0 \qquad (10)$$

Pré-multiplicando-se por $\mathbf{W}^{-1}$:

$$\left(\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I}\right)\boldsymbol{\gamma} = 0 \qquad (11)$$

As soluções não triviais são os autovalores e autovetores da matriz $\mathbf{W}^{-1}\mathbf{B}$.

Assim, obtêm-se *p* discriminantes. O número de discriminantes a ser adotado é:

$$d = \min(G - 1; p) \qquad (12)$$

Exemplo:

3 populações bivariadas, 3 grupos, com mesma variância:

G1: $\begin{bmatrix} -2 & 5 \\ 0 & 3 \\ -1 & 1 \end{bmatrix}$ G2: $\begin{bmatrix} 0 & 6 \\ 2 & 4 \\ 1 & 2 \end{bmatrix}$ G3: $\begin{bmatrix} 1 & -2 \\ 0 & 0 \\ -1 & -4 \end{bmatrix}$

Centróides de cada grupo:

$\bar{\mathbf{x}}_1 = \begin{bmatrix} -1 \\ 3 \end{bmatrix}$ $\bar{\mathbf{x}}_2 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$ $\bar{\mathbf{x}}_3 = \begin{bmatrix} 0 \\ -2 \end{bmatrix}$

Centróide de tudo: $\bar{\mathbf{x}} = \begin{bmatrix} 0 \\ 5/3 \end{bmatrix}$

Cálculo das matrizes:

$$\mathbf{B} = \sum_{g=1}^{3} 3\left(\bar{\mathbf{x}}_\mathbf{g} - \bar{\mathbf{x}}\right)\left(\bar{\mathbf{x}}_\mathbf{g} - \bar{\mathbf{x}}\right)^T = \begin{bmatrix} 6 & 3 \\ 3 & 62 \end{bmatrix} \qquad \mathbf{W} = \sum_{g=1}^{3}\sum_{j=1}^{G}\left(\bar{\mathbf{x}}_{gj} - \overline{\mathbf{x}_g}\right)\left(\bar{\mathbf{x}}_{gj} - \overline{\mathbf{x}_g}\right)^T = \begin{bmatrix} 6 & -2 \\ -2 & 24 \end{bmatrix}$$

Inversa de **W**:

$$\mathbf{W} = \mathbf{W}^{-1} = \frac{1}{140}\begin{bmatrix} 24 & 2 \\ 2 & 6 \end{bmatrix} \qquad \mathbf{W}^{-1}\mathbf{B} = \begin{bmatrix} 1{,}071 & 1{,}4 \\ 0{,}214 & 2{,}7 \end{bmatrix}$$

Autovalores e autovetores da matriz:

$\lambda_1 = 2{,}87$; $\lambda_2 = 0{,}90$

$$\boldsymbol{\gamma}_1 = \begin{bmatrix} 0{,}386 \\ 0{,}495 \end{bmatrix} \qquad \boldsymbol{\gamma}_2 = \begin{bmatrix} 0{,}938 \\ -0{,}112 \end{bmatrix}$$

Os autovetores são comumente normalizados fazendo:

$$\boldsymbol{\gamma}^T S \boldsymbol{\gamma} = 1$$

Em que *S* é a variância ponderada na forma:

$$S = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2 + (n_3 - 1)S_3}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1)}$$

Critério de alocação de novas observações:

1) Calcula-se o valor de cada discriminante Y_g para a nova observação;

2) Calcula-se o discriminante para a centróide de cada grupo:

3) Aloca-se com base na mínima distância quadrática entre o valor dos discriminantes em relação à centróide de cada grupo:

$$\min\left[\sum_{d=1}^{D}\left(Y_d - \overline{Y}_{d,g}\right)^2\right], \text{ para } g = 1,\ldots, G$$

Classificação

Podem-se utilizar as funções discriminantes, como os discriminantes lineares de Fisher, para classificar uma nova observação em um dos grupos previamente conhecidos. O critério mais usado para classificação nesses casos é a soma dos desvios quadráticos entre discriminantes, ou seja: aloca-se uma observação qualquer no grupo para o qual a soma dos desvios quadráticos dos discriminantes é a menor entre os grupos. O procedimento é:

1) calcular os valores dos *D* discriminantes para a nova observação, $\mathbf{X}_{obs}$:

$$y_{d,obs} = \gamma_d{}^T \mathbf{X}_{obs}\text{ , } d = 1,...D;$$

2) calcular os valores dos discriminantes para os centróides dos *G* grupos:

$$\bar{y}_{d,g} = \gamma_d{}^T \bar{\mathbf{X}}_g\text{ , }\quad d = 1,\ldots D,\ g = 1,\ldots G;$$

3) buscar a menor distância quadrática:

$$\min\left[\sum_{d=1}^{D}\left(y_{d,obs} - \bar{y}_{d,g}\right)^2\right],\ \text{para}\ g = 1,\ldots G$$

Alocar $\mathbf{X}_{obs}$ no grupo com mínima distância quadrática.

<u>Critérios estatísticos de classificação entre duas populações</u>:

Considerando-se duas populações (ou grupos) de observações x_i, i =1,...n, com probabilidades de ocorrência a priori dadas por p_1 e p_2, de modo que:

$$p_1 + p_2 = 1,$$

e que as funções densidade de probabilidade, $f_1(x)$ e $f_2(x)$, têm a forma ilustrada na Figura 6, pode-se dizer que a probabilidade de uma observação qualquer, x_0, pertencente a um grupo m, ser alocada em um dado grupo g, $P(g|m)$, é expressa por:

$$P(g|m) = P(x_0 \in Rg|Gm) = \int_{Rg} f_m(x)dx$$

Para o caso aqui considerado, a probabilidade de alocar x_0 erradamente é:

$$P(2|1) = P(x_0 \in R2|G1) = \int_{R2} f_1(x)dx$$

e:

$$P(1|2) = P(x_0 \in R1|G2) = \int_{R1} f_2(x)dx$$

e a probabilidade de alocar x_0 corretamente é:

$$P(1|1) = P(x_0 \in R1|G1) = \int_{R1} f_1(x)dx$$

e:

$$P(2|2) = P(x_0 \in R2|G2) = \int_{R2} f_2(x)dx \quad .$$

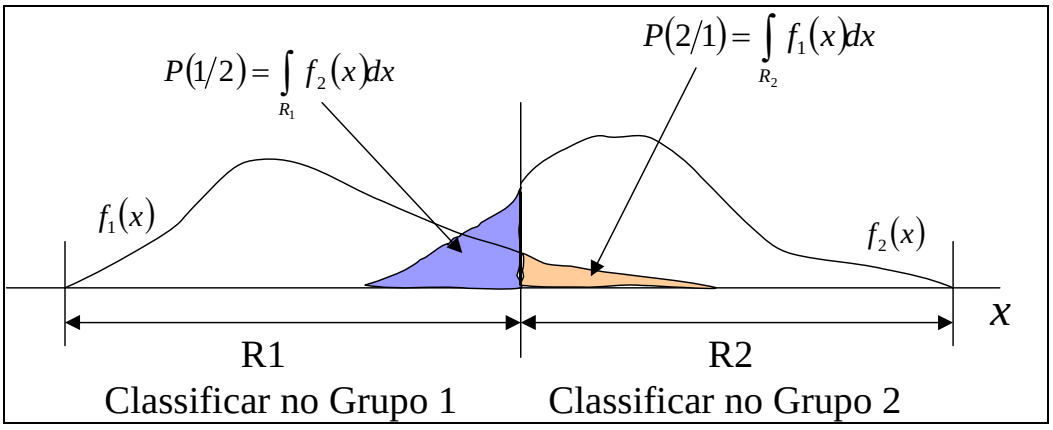

Figura 6. Ilustração da distribuição de observações de dois grupos.

Essas probabilidades podem ser expressas, então, por:

para os acertos:

$$P(x_0 \in R1|G1) = P(1|1).p(1)$$

$$P(x_0 \in R2|G2) = P(2|2).p(2)$$

e para os erros:

$$P(x_0 \in R2|G1) = P(2|1).p(1)$$

e:

$$P(x_0 \in R1|G2) = P(1|2).p(2)$$

Podem-se atribuir atribuir “custos”, ou “punições”, pelos erros de classificação, na forma de uma “matriz de custo”:

	Classificação	
Grupo	*G1*	*G2*
G1	0	*C(2|1)*
G2	*C(1|2)*	0

Define-se a função “custo esperado de falhas”, *ECM* (para “expected cost of missclassification”), como:

$$ECM = C(2|1).P(2|1).p_1 + C(1|2).P(1|2).p_2 .$$

Os algoritmos de classificação baseiam-se na minimização dessa função, a qual pode ser escrita como:

$$ECM = C(2|1).p_1 \int_{R2} f_1(x)dx + C(1|2).p_2 \int_{R1} f_2(x)dx$$

como

$$\int_{R1} f_1(x)dx + \int_{R2} f_1(x)dx = \int_{R1+R2} f_1(x)dx = 1$$

então

$$ECM = C(2|1).p_1\left[1 - \int_{R1} f_1(x)dx\right] + C(1|2).p_2 \int_{R1} f_2(x)dx$$

ou

$$ECM = \int_{R1}\left[C(1|2).p_2 f_2(x) - C(2|1).p_1 f_1(x)\right]dx + C(2|1).p_1$$

Como o último termo à direita é constante e positivo, a função *ECM* só diminui na região *R1* se o integrando for negativo. Assim, pode-se estabelecer o seguinte critério de classificação:

Alocar x_0 em *R1* se:

$$\frac{f_1(x_0)}{f_2(x_0)} \geq \left(\frac{C(1|2)}{C(2|1)}\right).\left(\frac{p_2}{p_1}\right)$$

Para a região *R2*, fazendo-se a mesma substituição:

Alocar x_0 em *R2* se:

$$\frac{f_1(x_0)}{f_2(x_0)} < \left(\frac{C(1|2)}{C(2|1)}\right).\left(\frac{p_2}{p_1}\right).$$

Classificações baseadas em populações com distribuição normal multivariada:

Para *G* grupos de observações multivariadas (com dimensão *p*) a função densidade de distribuição normal de probabilidade das observações em um grupo *g* qualquer é expressa como:

$$f_g(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\left|\sum\right|_g^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_g)^T \sum_g^{-1}(\mathbf{x}-\boldsymbol{\mu}_g)\right] \quad g = 1, 2,... G$$

Supondo que: a) as variâncias dos grupos não sejam iguais; b) os custos de alocação correta, *C(g|g)*, sejam iguais a zero; c) os custos de alocação errada, *C(g|m)*, sejam iguais a 1, pode-se definir um critério de alocação similar ao anterior, baseado no produto:

$$p_g\, f_g(\mathbf{x})$$

Para isso, normalmente se utiliza a função densidade de distribuição normal de probabilidade na forma linearizada, ficando o produto na forma:

$$\ln\left[p_g\, f_g(\mathbf{x})\right] = \ln(p_g) - \left(\frac{p}{2}\right)\ln(2\pi) - \frac{1}{2}\ln\left|\sum_g\right| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_g)^T \sum_g^{-1}(\mathbf{x}-\boldsymbol{\mu}_g) \quad g = 1, 2,... G$$

Aloca-se uma observação qualquer,$\mathbf{x}_0$, no grupo para o qual essa expressão for máxima. Como o segundo termo do lado direito da equação é o mesmo para todos os grupos, a comparação entre os grupos baseia-se nos demais termos. Assim, define-se o "discriminante quadrático", expresso como:

$$discr.Q_g = \ln(p_g) - \frac{1}{2}\ln\left|\sum_g\right| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_g)^T \sum_g^{-1}(\mathbf{x}-\boldsymbol{\mu}_g)$$

Por esse critério, aloca-se $\mathbf{x}_0$ no grupo *g* se $discr.Q_q$ for máximo para esse grupo, em comparação com os demais grupos. O discriminante é denominado "quadrático" devido à distância estatística quadrática, presente na equação.