

Análise Discriminante: parte 2

Prof. Caio Azevedo

- A classificação entre duas populações sob heterocedasticidade ($\Sigma_1 \neq \Sigma_2$) pode ser encontrada em Johnson & Wichern.
- A regra de classificação apresentada no supramencionado livro pode gerar resultados indesejados (altas taxas de erro), principalmente quando temos mais de duas variáveis e quando a suposição de normalidade multivariada não é válida.
- Não discutiremos a AD sob heterocedasticidade.

Classificação com várias populações baseado no CECE

- Temos g populações e cada unidade amostral pertence à uma e somente uma população. Defina:
 - Suporte da distribuição: $A = \{\mathbf{x} \in \mathfrak{R}^p, f(\mathbf{x}) > 0\}$ e a respectiva partição $A = \dot{\cup}_{i=1}^g A_i$.
 - p_i : probabilidade (à priori) de um determinado indivíduo pertencer à população i .
 - $c(k|i)$: custo de classificar o indivíduo na população k dado que ele pertence à população i , naturalmente $c(i|i) = 0, \forall i$.
 - $P(k|i) = \int_{A_k} f_i(\mathbf{x}) d\mathbf{x}$: probabilidade de classificar um indivíduo na população k dado que ele pertence à população i . Além disso $P(i|i) = 1 - \sum_{k=1, k \neq i}^g P(k|i)$.

Classificação com várias populações baseado no CECE

- A esperança condicional de classificar equivocadamente uma unidade pertencente à população 1 em uma outra população (2,3,...,g) é dado por:

$$\begin{aligned} ECE(1) &= P(2|1)c(2|1) + P(3|1)c(3|1) + \dots + P(g|1)c(g|1) \\ &= \sum_{k=1}^g P(k|1)c(k|1) \end{aligned} \quad (1)$$

- Este valor esperado condicional ocorre com probabilidade p_1 . Analogamente podemos definir $ECE(2), \dots, ECE(g)$ e, além disso, cada uma delas ocorre com probabilidade $p_k, k = 2, \dots, g$

Classificação com várias populações baseado no CECE

- Assim o custo esperado de classificação errada é dado por:

$$\begin{aligned} CECE &= p_1 ECE(1) + p_2 ECE(2) + \dots + p_g ECE(g) \\ &= \sum_{i=1}^g p_i \sum_{k=1, k \neq i}^g P(k|i) c(k|i) \end{aligned} \quad (2)$$

- A regra de classificação que minimiza o CECE (equação (2)) consiste em classificar uma dada observação (\mathbf{x}_0), associada à uma determinada unidade, na população k , $k = 1, 2, \dots, g$, para o qual

$$\sum_{i=1}^g p_i f_i(\mathbf{x}_0) c(k|i)$$

é mínimo. Para a demonstração veja: Anderson. T. W. An introduction to multivariate statistical analysis. New York: John Wiley, 2003.

Classificação com várias populações baseado no CECE com custos iguais

- Nesse caso a regra passa a ser: aloca-se a unidade amostral na população k , com base um vetor de observações \mathbf{x}_0 se

$$p_k f_k(\mathbf{x}_0) > p_i f_i(\mathbf{x}_0), \forall i \neq k \quad (3)$$

ou de modo equivalente, se

$$\ln(p_k f_k(\mathbf{x}_0)) > \ln(p_i f_i(\mathbf{x}_0)), \forall i \neq k$$

Classificação com várias populações baseado no CECE com custos iguais

- Um aspecto interessante é que a regra de classificação (3) equivale àquela que maximiza a probabilidade à posteriori $P(k|\mathbf{x}_0)$ (\mathbf{x}_0 vem da população k dado que \mathbf{x}_0 foi observado), a qual é dada por:

$$P(k|\mathbf{x}_0) = \frac{p_k f_k(\mathbf{x}_0)}{\sum_{i=1}^g p_i f_i(\mathbf{x}_0)} = \frac{\text{priori} \times \text{verossimilhança}}{\sum[\text{priori} \times \text{verossimilhança}]}$$

(Teorema de Bayes)

Método de Fisher para AD com várias populações

- Fisher também desenvolveu uma metodologia de AD considerando várias populações.
- A idéia é semelhante ao caso de duas populações, no sentido que ele buscou definir funções univariadas (função discriminante) com base nas observações multivariadas.
- Neste caso, também, a metodologia de Fisher é equivalente à regra do mínimo CECE, sob normalidade multivariada, homocedasticidade, custos de classificação errada e probabilidades à priori iguais (veja Johnson & Wichern).

Método de Fisher para AD com várias populações

- Características importantes:
 - Trabalhar com representações univariadas de dados multivariados.
 - Análises gráficas que permitem analisar de modo simples o comportamento das populações de interesse.
 - Não assume normalidade dos dados.
 - Considera a homocedasticidade.

Método de Fisher para AD com várias populações

- A idéia é trabalhar com combinações lineares das observações \mathbf{x} (\mathbf{x}_0), ou seja, $Y = \mathbf{a}'\mathbf{x}$. Assuma que $\mathcal{E}(\mathbf{X}|i) = \boldsymbol{\mu}_i$ e $\text{Cov}(\mathbf{X}|i) = \boldsymbol{\Sigma}$.
- Assim $\mathcal{E}(Y) = \mathbf{a}'\boldsymbol{\mu}_i$ (população i) e $\text{Cov}(Y) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$ (para todas as populações).
- Defina $(\bar{\boldsymbol{\mu}} = \frac{1}{g} \sum_{i=1}^g \boldsymbol{\mu}_i)$

$$V = \frac{\mathbf{a}' \left(\sum_{i=1}^g (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}) (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})' \right) \mathbf{a}}{\mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}} = \frac{\text{variabilidade entre grupos}}{\text{variabilidade dentro de cada grupo}}.$$

- A idéia de Fisher foi tentar separar ao máximo as populações em relação à medida V , ou seja, ele buscou maximizá-la.

Método de Fisher para AD com várias populações

- Resultado: defina $\mathbf{w} = \sum_{i=1}^g (n_i - 1) \mathbf{s}_i^2$ (variabilidade dentro de cada população), $\mathbf{b} = \sum_{i=1}^g (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'$, em que $\bar{\mathbf{x}}_i$ é a média amostral, \mathbf{s}_i^2 é a matriz de covariâncias da população i e $\bar{\mathbf{x}} = \frac{1}{g} \sum_{i=1}^g \bar{\mathbf{x}}_i$.
- Sejam $\tilde{\lambda}_1, \tilde{\lambda}_2, \dots, \tilde{\lambda}_s > 0$, em que $s \leq \min(g - 1, p)$ os autovalores diferentes de zero de $\mathbf{w}^{-1}\mathbf{b}$ e $\tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2, \dots, \tilde{\mathbf{e}}_s$ os respectivos autovetores, devidamente reescalados (veja Johnson & Wichern).

Método de Fisher para AD com várias populações

- Então o vetor de coeficientes $\tilde{\mathbf{a}}$ que maximiza a razão

$$\frac{\tilde{\mathbf{a}}' \tilde{\mathbf{b}}}{\tilde{\mathbf{a}}' \tilde{\mathbf{w}}}$$

é dado por $\tilde{\mathbf{a}}_1 = \tilde{\mathbf{e}}_1$. A combinação linear $\tilde{y}_1 = \tilde{\mathbf{a}}_1 \mathbf{x}$ é denominada de primeira função discriminante (amostral). Assim, a combinação linear $\tilde{y}_k = \tilde{\mathbf{a}}_k \mathbf{x}$ é denominada k -ésima função discriminante (amostral).

Método de Fisher para AD com várias populações

- Regra de classificação: aloca-se a observação \mathbf{x}_0 (ou seja, a unidade amostral), à população k , com base em r funções discriminantes, se:

$$\sum_{j=1}^r (\tilde{y}_j - \bar{y}_{kj}) = \sum_{j=1}^r [\tilde{\mathbf{a}}_j'(\mathbf{x}_0 - \bar{\mathbf{x}}_k)]^2 \leq \sum_{j=1}^r [\tilde{\mathbf{a}}_j'(\mathbf{x}_0 - \bar{\mathbf{x}}_i)]^2, \forall i \neq k$$

- Nota: assim como no caso de duas populações, a regra de classificação de Fisher coincide com a regra do CECE (sob normalidade), sob custos iguais e probabilidades de classificação iguais.

Voltando ao Exemplo 1 (considerando os três grupos)

- Resultados da classificação:

	S	VE	VI
S	25	0	0
VE	0	24	1
VI	0	1	24

- TEA (%) : 2,67.

Medidas resumo

Função discriminante 1

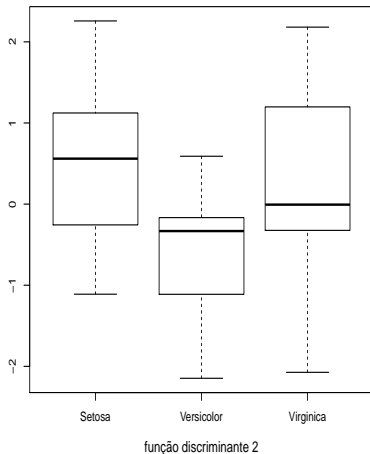
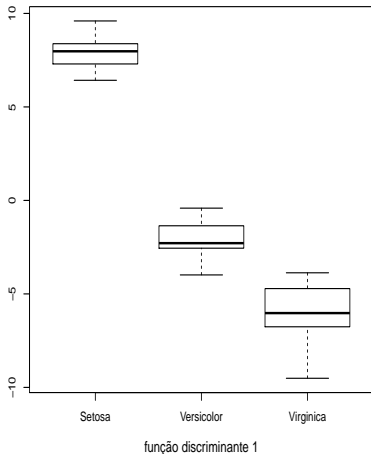
Grupo	Média	DP	Var.	Mínimo	Mediana	Máximo	n
Setosa	7,91	0,82	0,67	6,42	7,98	9,60	25
Versicolor	-2,09	0,94	0,89	-3,98	-2,29	-0,41	25
Virginica	-5,95	1,40	1,97	-9,51	-6,03	-3,87	25

Medidas resumo

Função discriminante 2

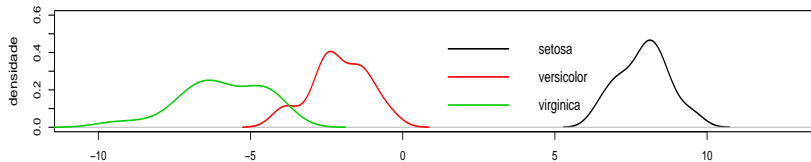
Grupo	Média	DP	Var.	Mínimo	Mediana	Máximo	n
Setosa	0,54	0,93	0,87	-1,11	0,56	2,26	25
Versicolor	-0,54	0,79	0,62	-2,15	-0,33	0,59	25
Virginica	0,25	1,09	1,19	-2,07	-0,01	2,18	25

Ex. 1: boxplots da função discriminante



Ex. 1: densidade estimada da função discriminante

Função discriminante 1



Função discriminante 2

