MAE 5776

ANÁLISE MULTIVARIADA

Júlia M Pavan Soler

pavan@ime.usp.br

1º Sem/2020 - IME

Análise Multivariada

$$Y_{n\times p} = \left(Y_{ij}\right) \in \Re^{n\times p}$$

Já vimos ☺

- ✓ Estatísticas descritivas multivariadas, Episóides de Concentração, Boxplot Bivariado
- ✓ Distribuição N_p, Distribuições Amostrais (T^2 e W_p)
- ✓ $N_p(\mu_g;\Sigma_g)$: Inferências sobre μ_g (T^2, MANOVA, ICS, Correções para Múltiplos testes

Decomposições: SS_T e Y_{nxp}

Técnicas Multivariadas:

Já vimos ☺

- ✓ 1. Análise de Componentes Principais (CP)
- ✓ 2. Escalonamento Multidimensional (CoP)
- ✓ 3. Análise de Correspondência
- ✓ 4. Análise Fatorial

Quais são os critérios de otimização?

- Análise Discriminante (MANOVA)
- Análise de Agrupamento
- Análise de Correlação Canônica

Análise Discriminante

- ✓ Solução de Fisher: G=2 (solução explícita)
  G>2 (maximizar $\Sigma^{-1}\Sigma_B$)

Regra geral de Classificação de Observações)

Análise Discriminante

Foco está na classificação de observações, mais do que na obtenção de vetores reducionistas

Problema Geral de Classificação

Caso de Duas Populações (G=2)

Suposição: Uma população está estratificada em 2 subpopulações, $\tau 1$ e $\tau 2$, e de cada subpopulação é retirada uma amostra de tamanho n1 e n2, respectivamente.

Com base na amostra, para encontrar uma regra de discriminação de observações de cada população, uma alternativa é particionar o espaço amostral Ω em duas regiões, R1 e R2, que favoreçam às populações $\tau 1$ e $\tau 2$, respectivamente, tal que, para uma observação Y_0 tem-se que, se

$Y_0 \in R_1 \Rightarrow$ a observação é de $\tau 1$

$Y_0 \in R_2 \Rightarrow$ a observação é de $\tau 2$

Regra discriminante

Como determinar R1 e R2 ?

Análise Discriminante

Problema Geral de Discriminação/Classificação - Solução Probabilística

Caso de Duas Populações

Probabilidades a priori: $\tau_1 \Rightarrow p_1(y) \qquad \tau_2 \Rightarrow p_2(y) \qquad p_1 + p_2 = 1$

Função densidade de probabilidades: $\tau_1 \Rightarrow f_1(y) \qquad \tau_2 \Rightarrow f_2(y)$

Probabilidade de Classificação Errada:

$$P(2|1) = P(Y_i \in R_2 \mid \tau_1) = \int_{R_2 = \Omega - R_1} f_1(y)\, dy$$

$$P(1|2) = P(Y_i \in R_1 \mid \tau_2) = \int_{R_1 = \Omega - R_2} f_2(y)\, dy$$

Probabilidade de Classificação Correta:

$$P(1|1) = P(Y_i \in R_1 \mid \tau_1) = \int_{R_1 = \Omega - R_2} f_1(y)\, dy$$

$$P(2|2) = P(Y_i \in R_2 \mid \tau_2) = \int_{R_2 = \Omega - R_1} f_2(y)\, dy$$

Análise Discriminante

Problema Geral de Classificação - Caso de Duas Populações

Notação

Probabilidade de Classificação

Verdade	Predito τ_1	Predito τ_2
τ_1	P(1,1)	P(2,1)
τ_2	P(1,2)	P(2,2)

Custo de Classificação

Verdade	Predito τ_1	Predito τ_2
τ_1	0	c(2|1)
τ_2	c(1|2)	0

$$P(1,1) = P(1|1)p_1 \qquad P(2,1) = P(2|1)p_1$$

$$P(1,2) = P(1|2)p_2 \qquad P(2,2) = P(2|2)p_2$$

Logo, o custo esperado de classificação errada (*CECE*) é dado por:

$$\begin{aligned} CECE &= c(2|1)P(2,1) \quad + \quad c(1|2)P(1,2) \\ &= c(2|1)P(2|1)p_1 \quad + \quad c(1|2)P(1|2)p_2 \end{aligned}$$

$\Rightarrow$ Obter R1 e R2 que minimizem *CECE*

Análise Discriminante

Problema Geral de Classificação - Caso de Duas Populações

Minimizar o custo esperado de classificação errada:

$$CECE = c(2|1)\,P(2|1)p_1 \quad + \quad c(1|2)\,P(1|2)p_2$$

$$= c(2|1)\,p_1 \int_{R_2} f_1(y)\,dy \quad + \quad c(1|2)\,p_2 \int_{R_1} f_2(y)\,dy$$

$$= c(2|1)\,p_1 \quad + \int_{R_1} [c(1|2)\,p_2\ f_2(y) - c(2|1)\,p_1\ f_1(y)]\,dy$$

$\leq 0 \Rightarrow$ mínimo *CECE*

Somando e subtraindo $c(2|1)\,p_1 \int_{R_1} f_1(y)\,dy$ tem-se:

R1 e R2 são conjuntos de valores Y $\in \Re^p$ para os quais:

$$R_1: \quad \frac{f_1(y)}{f_2(y)} \geq \frac{c(1|2)p_2}{c(2|1)p_1} \qquad R_2: \quad \frac{f_1(y)}{f_2(y)} < \frac{c(1|2)p_2}{c(2|1)p_1}$$

Discriminação sob Estimação

Problema Geral de Classificação - Caso de Duas Populações Normais

Função densidade de probabilidades:

heterocedasticidade

$$\tau_g \Rightarrow f_g(y) = \frac{1}{(2\pi)^{p/2}\left|\Sigma_g\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(Y-\mu_g\right)'\Sigma_g^{-1}\left(Y-\mu_g\right)\right\}; \quad g=1,2; Y \in \Re^p$$

Classificar uma observação em $\tau 1$ se $Y \in \Re^p$ pertencer à região R1 dada por:

$$R_1: \quad -\frac{1}{2}Y'\left(\Sigma_1^{-1}-\Sigma_2^{-1}\right)Y + \left(\mu_1'\Sigma_1^{-1}-\mu_2'\Sigma_2^{-1}\right)Y - c \geq \ln\left[\frac{c(1|2)\,p_2}{c(2|1)\,p_1}\right]$$

em que, $c = \frac{1}{2}\ln\left(\frac{\left|\Sigma_1\right|}{\left|\Sigma_2\right|}\right) + \frac{1}{2}\left(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2\right)$

R2 é dada pelo complementar de R1 em Ω.

Sob Heterocedasticidade $\Rightarrow$ Função Discriminante Quadrática (em $Y \in \Re^p$)

Regra de Discriminação Amostral: obter estimador de MVS

Análise Discriminante

Problema Geral de Classificação - Caso de Duas Populações Normais

Regra de discriminação: (os parâmetros são substituídos por sua estimativas)

Alocar Y_0 em $\tau 1$ se X_0^Q

$$-\frac{1}{2}Y_0'\left(S_1^{-1}-S_2^{-1}\right)Y_0+\left(\bar{X}_1'S_1^{-1}-\bar{X}_2'S_2^{-1}\right)Y_0-\hat{c}_Q \geq \ln\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right]$$

Alocar Y_0 em $\tau 2$ caso contrário

em que, $\hat{c}_Q=\frac{1}{2}\ln\left(\frac{|S_1|}{|S_2|}\right)+\frac{1}{2}\left(\bar{Y}_1'S_1^{-1}\bar{Y}_1'-\bar{Y}_2'S_2^{-1}\bar{Y}_2'\right)$

Função discriminante quadrática

Critério flexível: permite heterocedasticidade, custos e prioris diferentes

Análise Discriminante

Problema Geral de Classificação - Caso de Duas Populações Normais

$$Y_i \in \tau_k \; ; Y_i \overset{iid}{\sim} N_p\left(\mu_g ; \Sigma_g\right) \quad g = 1,2$$

$\Rightarrow$ **Suposição:** $\Sigma_1 = \Sigma_2 = \Sigma$

Regra de discriminação:

Alocar Y_0 em $\tau 1$ se

$$\underbrace{\left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1} Y_0}_{X_0^L} - \underbrace{\frac{1}{2}\left(\bar{Y}_1 - \bar{Y}_2\right)' S_c^{-1}\left(\bar{Y}_1 + \bar{Y}_2\right)}_{\hat{c}} \geq \ln\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right]$$

Alocar Y_0 em $\tau 2$ caso contrário

$\Rightarrow$ Note que a função discriminante X_0^L é linear em Y_0

Análise Discriminante

Problema Geral de Classificação - Caso de Duas Populações Normais

$$Y_i \in \tau_k \; ; Y_i \overset{iid}{\sim} N_p\left(\mu_g ; \Sigma\right) \quad g = 1,2$$

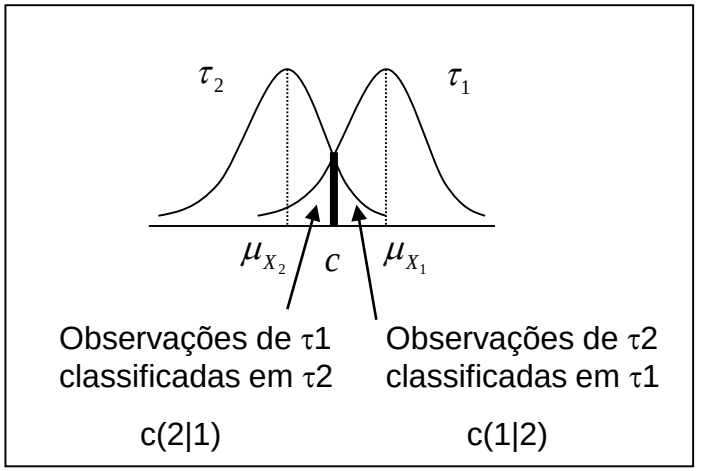
homocedasticidade

$\Rightarrow$ Função Discriminante Linear

Alocar Y_0 em $\tau 1$ se

$$X_0 - \hat{c} \geq \ln\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right]$$

Alocar Y_0 em $\tau 2$ caso contrário



- Se os custos e as prioris são iguais $\Rightarrow$ função discriminante linear de Fisher
- Se $c(2|1) > c(1|2)$ e $p_1 = p_2$ $\Rightarrow$ o limite "c" é deslocado para a esquerda
- Se $p_1 < p_2$ e $c(2|1) = c(1|2)$ $\Rightarrow$ o limite "c" é deslocado para a direita

Análise Discriminante

Banco	Condição	Y1	Y2	Y3	Y4
B1	1	0,8888	0,7391	1,0255	0,3938
B2	1	1,6655	0,7268	0,878	0,0004
B3	1	2,2111	0,9166	0,9492	0,342
B4	1	1,4351	0,9133	0,9577	0,2325
B5	1	2,1414	0,002	1,0245	0,3966
B6	1	1,192	0,4972	1,034	0,3095
B7	1	1,5895	0,2593	1,0453	0,557
B8	1	1,3272	0,4126	1,0448	0,3482
B9	1	1,8847	0,388	0,9864	0,0337
B10	1	0,5229	0,9473	1,1244	0,118
n		10	10	10	10
Média		1,4852	0,5802	1,007	0,2732
D.P.		0,533	0,319	0,0674	0,1762
B11	2	0,4922	0,3166	1,1127	0,1628
B12	2	1,4427	0,0589	0,9019	0,1355
B13	2	0,5438	0,5358	1,03	0,1481
B14	2	0,1904	0,7087	0,9917	0,2625
B15	2	0,1102	0,7378	1,528	0,0783
B16	2	2,006	0,014	1,0321	0,0816
B17	2	0,2321	0,9234	0,9753	0,0045
B18	2	0,9019	0,1634	1,1414	0,5485
B19	2	1,9757	0,3395	0,9997	0,0751
B20	2	0,7276	0,3139	1,1077	0,2957
n		10	10	10	10
Média		0,862	0,4112	1,0821	0,1793
D.P.		0,712	0,3055	0,1726	0,1567

Condição:

1: Com problemas

2: Sem problemas

Objetivo:

Obter uma função de discriminação com base nas 4 variáveis de indicadores econômicos

⇒Obtenha a função discriminante linear e quadrática

⇒ Quais suposições estão implícitas em cada caso?

Análise Discriminante

Dados dos Bancos

$$\bar{Y}_{g=1} = \begin{pmatrix} 1{,}486 \\ 0{,}580 \\ 1{,}007 \\ 0{,}273 \end{pmatrix} \qquad S_{g=1} = \begin{pmatrix} 0{,}284 & & & \\ -0{,}070 & 0{,}102 & & \\ -0{,}021 & -0{,}004 & 0{,}005 & \\ 0{,}008 & -0{,}022 & 0{,}004 & 0{,}031 \end{pmatrix}$$

$$\bar{Y}_{g=2} = \begin{pmatrix} 0{,}862 \\ 0{,}414 \\ 1{,}082 \\ 0{,}179 \end{pmatrix} \qquad S_{g=2} = \begin{pmatrix} 0{,}505 & & & \\ -0{,}164 & 0{,}091 & & \\ -0{,}051 & 0{,}014 & 0{,}030 & \\ -0{,}012 & -0{,}016 & 0{,}002 & 0{,}025 \end{pmatrix}$$

Usar o teste M de Box para verificar a homocedasticidade!

$$S_c = \begin{pmatrix} 0{,}395 & & & \\ -0{,}117 & 0{,}096 & & \\ -0{,}036 & 0{,}005 & 0{,}017 & \\ -0{,}002 & -0{,}019 & 0{,}003 & 0{,}028 \end{pmatrix}$$

Análise Discriminante

Dados dos Bancos

homocedasticidade

Suposição: $Y_i \in \tau_g \; ; Y_i \overset{iid}{\sim} N_p(\mu_g ; \Sigma_g) \quad g = 1,2 \qquad \Sigma_1 = \Sigma_2 = \Sigma$

Custos de classificação Errada e Prioris iguais para as populações

$\Rightarrow$ Função Discriminante Linear de Fisher

$$X_0 - c \geq \ln\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right] \quad \Rightarrow \quad X_0 - c - \ln\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right] \geq 0$$

$$\underbrace{(\bar{Y}_1 - \bar{Y}_2)' S_c^{-1} Y_0}_{X_0^L} \underbrace{- \frac{1}{2}(\bar{Y}_1 - \bar{Y}_2)' S_c^{-1}(\bar{Y}_1 + \bar{Y}_2)}_{c} \underbrace{- \ln\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right]}_{=0}$$

4.36Y1 + 8.91Y2 + 0.52Y3 + 9.71Y4 -12.27

Análise Discriminante

Dados dos Bancos

Suposição: $Y_i \in \tau_g \; ; Y_i \overset{iid}{\sim} N_p\left(\mu_g ; \Sigma_g\right) \quad g = 1,2$ heterocedasticidade

Custos de classificação Errada e Probabilidades a Priori iguais para as populações

$\Rightarrow$ Função Discriminante Quadrática

$$\underbrace{-\frac{1}{2} Y_0' \left(S_1^{-1} - S_2^{-1}\right) Y_0 + \left(\bar{Y}_1' S_1^{-1} - \bar{Y}_2' S_2^{-1}\right) Y_0 - \hat{c}_Q}_{X_0^Q} \geq \ln\left[\frac{c(1|2)p_2}{c(2|1)p_1}\right] = 0$$

$$-0,214Y_1^2 + 14,535Y_2^2 - 204,116Y_3^2 + 14,038Y_4^2$$
$$+9,332Y_1Y_2 + -38,603Y_1Y_3 + 16,846Y_1Y_4 - 35,125Y_2Y_3 + 31,732Y_2Y_4 + 43,362Y_3Y_4$$
$$+38,194Y_1 + 17,076Y_2 + 478,004Y_3 - 73,415Y_4 - 273,776$$

Análise Discriminante

Dados dos Bancos – Avaliação empírica da Regra de Classificação

Banco	Condição	Regra Linear		Regra Quadrática	
		X	Grupo	X	Grupo
B1	1	3,336	1	6,329	1
B2	1	2,701	1	4,108	1
B3	1	10,223	1	27,94	1
B4	1	5,825	1	13,043	1
B5	1	1,403	1	2,861	1
B6	1	1,425	1	3,046	1
B7	1	3,146	1	6,822	1
B8	1	1,539	1	3,359	1
B9	1	0,635	1	1,247	1
B10	1	1,222	1	1,151	1
B11	2	-4,74	2	-1,808	2
B12	2	-3,57	2	-5,13	2
B13	2	-2,514	2	-2,313	2
B14	2	-1,223	2	-4,802	2
B15	2	-2,862	2	-39,071	2
B16	2	-1,862	2	-1,397	2
B17	2	-1,4	2	-3,083	2
B18	2	-0,792	2	0,713	1
B19	2	0,945	1	1,411	1
B20	2	-2,484	2	-1,175	2

Regra Linear:

Real	Predito 1	2
1	10	0
2	1	9

Regra Quadrática:

Real	Predito 1	2
1	10	0
2	2	8

Função linear classifica melhor!

Testar a igualdade das matrizes de covariância (Teste de Box). Decidir pela função linear (de Fisher) no caso da não rejeição de $H_0 : \Sigma_1 = \Sigma_2$.

```
Chi-Sq (approx.) = 13.733, df = 10, p-value = 0.1855
```

Análise Discriminante

Problema Geral de Classificação – Caso de Muitas Populações

As Regiões de Classificação que minimizam *CEEC* são definidas por alocar Y_0 à população τ_k, k=1,2,…,G, que atinge o mínimo erro de classificação, dado por:

$$\sum_{\substack{g=1 \\ g \neq k}}^{G} p_g \; f_g(y)\, c(k \mid g)$$

Logo, se todos os custos são iguais, devemos alocar Y_0 à população τ_k se:

$$p_k \; f_k(y) > p_g \; f_g(y) \qquad g = 1,\ldots,G; g \neq k$$

ou, equivalentemente: $\ln p_k \; f_k(y) > \ln p_g \; f_g(y) \qquad g = 1,\ldots,G; g \neq k$

Análise Discriminante

Problema Geral de Classificação – Caso de Muitas Populações

Alocar Y_0 a τ_k se: $\ln p_k \ f_k(y) > \ln p_g \ f_g(y) \qquad g = 1,\ldots,G; g \neq k$

Caso Especial (N_p): $Y_i \overset{iid}{\sim} N_p(\mu_g; \Sigma_g)$ heterocedasticidade

$$f_g(y) = \frac{1}{(2\pi)^{p/2}|\Sigma_g|^{1/2}} \exp\left\{-\frac{1}{2}(Y-\mu_g)'\Sigma_g^{-1}(Y-\mu_g)\right\}, \quad g = 1,2,\ldots,G$$

$$\ln p_k f_k(y) = \ln p_k - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(Y-\mu_k)'\Sigma_k^{-1}(Y-\mu_k) = \max_g \ln p_g f_g(y)$$

Define-se Escore Discriminante Quadrático (de $Y \in \Re^p$) para a g-ésima população:

$$d_g^Q(y) = -\frac{1}{2}\ln|\Sigma_g| - \frac{1}{2}(Y-\mu_g)'\Sigma_g^{-1}(Y-\mu_g) + \ln p_g \qquad g = 1,\ldots,G$$

Análise Discriminante

Problema Geral de Classificação – Caso de Muitas Populações

$$\ln p_k \; f_k(y) > \ln p_g \; f_g(y) \qquad g = 1,...,G; g \neq k$$

heterocedasticidade

$Y_i \overset{iid}{\sim} N_p(\mu_g; \Sigma_g) \Rightarrow$ Alocar Y a τ_k se o escore quadrático $d_k^Q(y)$ é maior que os demais

em que, $d_k^Q(y) = -\frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(Y-\mu_k)'\Sigma_k^{-1}(Y-\mu_k) + \ln p_k \quad k = 1,...,G$

homocedasticidade

Se $Y_i \overset{iid}{\sim} N_p(\mu_g; \Sigma)$, isto é, $\Sigma_1 = ... = \Sigma_g$

$$d_k^Q(y) \Rightarrow d_k(y) = \mu_k' \Sigma^{-1} Y - \frac{1}{2}\mu_k' \Sigma^{-1}\mu_k + \ln p_k \quad k = 1,...,G$$

Escore discriminante linear para a população τ_k

Análise Discriminante

Problema Geral de Classificação – Caso de Muitas Populações

$$Y_i \overset{iid}{\sim} N_p(\mu_g; \Sigma_g)$$

escore discriminante quadrático máximo

$$d_k^Q(y)$$

$$-\frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(Y-\mu_k)'\Sigma_k^{-1}(Y-\mu_k) + \ln p_k$$

$$Y_i \overset{iid}{\sim} N_p(\mu_g; \Sigma)$$

escore discriminante linear máximo

$$d_k(y)$$

$$\mu_k'\,\Sigma^{-1}Y - \frac{1}{2}\mu_k'\,\Sigma^{-1}\mu_k + \ln p_k$$

O Escore Discriminante linear pode ser comparado para duas populações, de tal modo que, a condição " $d_k(y)$ é maior ", fica equivalente a:

$$0 \le d_k(y) - d_g(y) = (\mu_k - \mu_g)'\,\Sigma^{-1}Y - \frac{1}{2}(\mu_k - \mu_g)'\,\Sigma^{-1}(\mu_k + \mu_g) + \ln\left(\frac{p_k}{p_g}\right)$$

Alocar Y a τ_k se: $(\mu_k - \mu_g)'\,\Sigma^{-1}Y - \frac{1}{2}(\mu_k - \mu_g)'\,\Sigma^{-1}(\mu_k + \mu_g) \ge \ln\left(\frac{p_g}{p_k}\right)$

Função de Fisher $\qquad$ c

Análise Discriminante

Validação Empírica de um Algoritmo de Classificação Amostral

Métricas de validação via a Matriz de Classificação

Matriz de Classificação (ou de Confusão)

	Predito		
Verdade	τ_1 +	τ_2 -	
τ_1 +	n_{1c} V+	n_{1M} F-	n_1
τ_2 -	n_{2M} F+	n_{2c} V-	n_2

- Taxa de Erro Aparente (proporção de itens mal classificados):

$$TxErro = \frac{n_{1M} + n_{2M}}{n_1 + n_2} = \frac{F_+ + F_-}{n}$$ Estima Pr(classificação errada)

- Acurácia: $$Acurácia = \frac{n_{1C} + n_{2C}}{n_1 + n_2} = \frac{V_+ + V_-}{n}$$ Estima Pr(classificação correta)

Métricas de Validação via a Matriz de Classificação

Matriz de Classificação (ou de Confusão)

	Predito		
Verdade	τ_1 +	τ_2 -	
τ_1 +	n_{1c} V+	n_{1M} F-	n_1
τ_2 -	n_{2M} F+	n_{2c} V-	n_2

- Sensibilidade = $\frac{V_+}{V_+ + F_-}$ = Pr(classificação + | +)
- Especificidade = $\frac{V_-}{F_+ + V_-}$ = Pr(classificação - | -)

Poder Preditivo via a Curva ROC: Sensibilidade x (1-Especificidade)

- Preditivo Positivo = $\frac{V_+}{F_+ + V_+}$ Precisão do classificador
- Preditivo Negativo = $\frac{V_-}{F_- + V_-}$
- Escore F1 = $2\frac{\mathrm{Pr}\,ecisão * Sensibilidade}{\mathrm{Pr}\,ecisão + Sensibilidade}$

Média harmônica da precisão e sensibilidade

Análise Discriminante

Validação de um Algoritmo de Classificação

Matriz de Classificação (ou de Confusão)

	Predito		
Verdade	τ_1	τ_2	
τ_1	n_{1c} **V+**	n_{1M} **F+**	n_1
τ_2	n_{2M} **F-**	n_{2c} **V-**	n_2

TxErro subestima a Probabilidade de erro de classificação (populacional), assim como as demais métricas:

$$TxErro = \frac{n_{1M} + n_{2M}}{n_1 + n_2} = \frac{F_+ + F_-}{n} \quad \rightarrow \quad p_1 \int_{R_2} f_1(y)dy + p_2 \int_{R_1} f_2(y)dy$$

Alternativas

- Método de Particionamento (*Data Split*): particiona os dados em Amostra de Treinamento e Amostra de Validação (Teste)
- Método de "Validação Cruzada" (Cross-validation)

Análise Discriminante

Validação de um Algoritmo de Classificação Amostral

Validação Cruzada pelo método Leave-One-Out (*Fold*=N)

1. Inicie com as observações de τ_1 . Omita uma obs deste grupo e obtenha a função de classificação baseada nos remanescentes N-1=(n1-1)+n2 observações (supondo G=2)
2. Classifique a obs omitida usando a função calculada no passo 1
3. Repetir os passos 1 e 2 até que todas as obs de τ_1 tenham sido classificadas. Calcule o número de erros de classificação neste grupo
4. Repita os passos de 1 a 3 para as observações do grupo 2.

Taxa de Erro de Classificação esperada é dada por:

$$TxErro = \frac{n_{1M}^{Cross} + n_{2M}^{Cross}}{n_1 + n_2}$$

Algoritmos de CV podem usar *Fold*=k

Análise Discriminante
Normalização de Variáveis

	Unidades Amostrais	Variáveis 1	2	…	j	…	p
G1	1	Y_{111}	Y_{112}		Y_{11j}		Y_{11p}
	2	Y_{121}	Y_{122}		Y_{12j}		Y_{12p}
	…	…	…	…	…		…
	n_1	Y_{1n11}	Y_{1n12}		Y_{1n1j}		Y_{1n1p}
G2	1	Y_{211}	Y_{212}		Y_{21j}		Y_{21p}
	2	Y_{221}	Y_{222}		Y_{22j}		Y_{22p}
	…	…	…	…	…		…
	n_2	Y_{2n21}	Y_{2n22}		Y_{2n2j}		Y_{2n2p}

$\bar{Y}_{1\,p\times 1}$ $S_{1\,p\times p}$

$\bar{Y}_{2\,p\times 1}$ $S_{2\,p\times p}$

$\bar{Y}_{p\times 1}$ $S_{c\,p\times p}$

Na AD a normalização das variáveis é usada com a finalidade de facilitar a interpretação dos pesos das variáveis na função discriminante e no cálculo de "*c*". O lda do R adota a "normalização" das variáveis para calcular as funções discriminantes, mas o linda não. A normalização da variável j avaliada no indivíduo i do grupo g é dada por:

$$Y^*_{gij} = \left(\frac{Y_{gij} - \bar{Y}_j}{s_{gj}} \right)$$

Média: para cada j independente de grupo

Variância: para cada grupo g e variável j

$$\bar{Y}_j = \frac{1}{n_1 + n_2} \sum_{g=1}^{2} \sum_{i=1}^{n_g} Y_{gij}$$

Média da variável j (j=1,..,p), independente de grupo

$$S_{gj} = \frac{1}{n_g - 1} \sum_{g=1}^{2} \left(Y_{gij} - \bar{Y}_{gj} \right)^2$$

Variância da variável j no grupo g

Dados dos Bancos - Funções Discriminantes Lineares - Classes Preditas

- Dados originais (linDA), prioris iguais
  1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 1 2

- Dados normalizados (lda), prioris amostrais
  1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 1 2

- Dados normalizados (lda), prioris iguais
  1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 1 2

- Dados normalizados (lda), prioris proporcionais: p1=2p2, p1+p2=1
  1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 1 2

- Dados normalizados (lda), prioris iguais, CV (Leave-One-Out)
  1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 1 1 2

- Dados normalizados, prioris iguais, *Data Split*: 70% Treinamento, 30% Teste (set.seed(1314):obs preditas 3,4,6,14,16,19)
  1 1 1 2 1 1

Dados dos Bancos - Função Discriminante Quadrática - Classes Preditas

- Dados normalizados (lda), prioris iguais
  1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 1 1 2

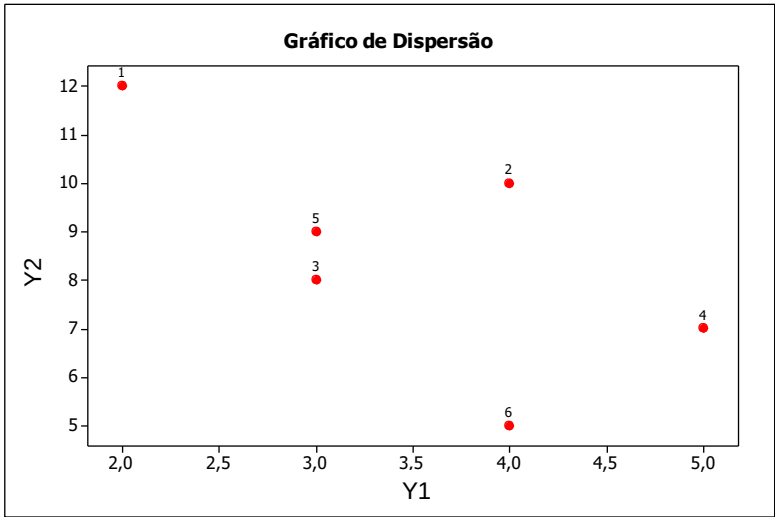
Compare as regras de classificação!

Análise Discriminante

Considere os dados (hipotéticos) a seguir em que duas variáveis foram observadas em três indivíduos do grupo 1 e em três indivíduos do grupo 2:

$$G_1 = \begin{pmatrix} 2 & 4 & 3 \\ 12 & 10 & 8 \end{pmatrix} \qquad G_2 = \begin{pmatrix} 5 & 3 & 4 \\ 7 & 9 & 5 \end{pmatrix}$$

1. Calcule a função discriminante de Fisher para a diferença entre os grupos. Qual é a regra de classificação de observações? Que suposições são feitas?
2. Calcule também o escore discriminante para cada grupo via o método geral de classificação. Suponha que *p1=p2*. E se *p1=2p2*?
3. Calcule a taxa observada de erro de classificação. Classifique a observação (4,7).
4. Calcule a taxa de erro de classificação via validação cruzada.
5. Obtenha a função discriminante para os dados “normalizados”.



Diferentes formulações da função discriminante linear.

Grupo	Y1	Y2	X=-0,33Y1+0,67Y2	X*=0,33Y1+0,67Y2-4,54	X1=7,33Y1+4,33Y2-32,67	X2=7,67Y1+3,67Y2-28,17	Grupo Pred
1	2	12	7,38	2,84	33,95	31,21	1
1	4	10	5,38	0,84	39,95	39,21	1
1	3	8	4,37	-0,17	23,96	24,2	2
2	5	7	3,04	-1,5	34,29	35,87	2
2	3	9	5,04	0,5	28,29	27,87	1
2	4	5	2,03	-2,51	18,3	20,86	2
			c=4,54				

Solução usando pacote `lda do R`: valores Y estão normalizados para ter variância 1

X = 0.2182179 Y1* - 0.4364358 Y2*
LD1= X-c; c=((-3.7097)+(-2.1821))/2=-2.946
LD1 <= 0 grupo1, cc grupo2
> fit.values$class
[1] 1 1 2 2 1 2

	LD1	
1	-1.8548521	alocar G1
2	-0.5455447	alocar G1
3	0.1091089	alocar G2
4	0.9819805	alocar G2
5	-0.3273268	alocar G1
6	1.6366342	alocar G2

Note as seguintes formulações da função discriminante linear (G=2):

$$X = \left(\bar{Y}_k - \bar{Y}_g\right)' S_c^{-1} Y; \qquad c = \frac{1}{2}\left(\bar{Y}_k - \bar{Y}_g\right)' S_c^{-1}\left(\bar{Y}_k + \bar{Y}_g\right)$$

$$X^* = \left(\bar{Y}_k - \bar{Y}_g\right)' S_c^{-1} Y - c$$

$$X_{gi} = \bar{Y}_g' S_c^{-1} Y_i - \frac{1}{2}\bar{Y}_g' S_c^{-1}\bar{Y}_g = d_g(y_i)$$

Note que, sob normalidade e homocedasticidade, $Y \mid \tau_g \sim N_p\left(\mu_g; \Sigma\right)$, $g = 1, 2$:

$$X_c = \left(\mu_1 - \mu_2\right)' \Sigma^{-1}\left(Y - \mu\right); \qquad \mu = \frac{1}{2}\left(\mu_1 + \mu_2\right)$$

$$X_c \mid Y \in \tau_1 \sim N\left(\frac{1}{2}d_M^2; d_M^2\right),$$

$$X_c \mid Y \in \tau_2 \sim N\left(-\frac{1}{2}d_M^2; d_M^2\right); \quad d_M^2 = \left(\mu_1 - \mu_2\right)'\Sigma^{-1}\left(\mu_1 - \mu_2\right)$$

Assim, a probabilidade de classificação errada é,

$$P\left(Y \text{ alocado em} \tau_1 \mid Y \in \tau_2\right) = P\left(X_c(y) > 0 \mid Y \in \tau_2\right) = P\left(Z > \frac{1}{2}d_M\right) = \Phi\left(-\frac{1}{2}d_M\right)$$

```
##Comandos R

#Análise discriminante
dat<-matrix(c(2,4,3,5,3,4,12,10,8,7,9,5,1,1,1,2,2,2),6,3)
xbar<-colMeans(dat[,1:2])
xbar1<-colMeans(dat[1:3,1:2])
xbar2<-colMeans(dat[4:6,1:2])
cov1<-cov(dat[1:3,1:2])
cov2<-cov(dat[4:6,1:2])
library(biotools)
mt<-boxM(dat[,-3],dat[,3])

library(DiscriMiner)
fitlda<-linDA(dat[,-3],dat[,3])

library(MASS) ##outra alternativa de analise
fit<- lda(dat[,3] ~ dat[,1] + dat[,2],prior=c(1,1)/2)
#fit<- lda(dat[,3] ~ dat[,1] + dat[,2],prior=c(2,1)/3) ; #p1=2p2, p1+p2=1
fit.values <- predict(fit, data.frame(dat[,1:2]))
fit.values$x
fit.values$class
ct <- table(dat[,3],fit.values$class)  #tabela com as classificações
diag(prop.table(ct, 1))    # % de classif correta
sum(diag(prop.table(ct)))
fit$svd   # (SSB-\lambda SSW)a=0
mv<-aggregate(fit.values$x, data.frame(dat[,3]), FUN=mean)
colMeans(mv[2])
```

Análise Discriminante via Modelo de Regressão Logística Dicotômica

$Y_i = 1$: se o indivíduo é do grupo $G=1$; $Y_i = 0$ se o indivíduo é do grupo $G=2$

$$E(Y \mid X, \beta) = P(Y = 1 \mid X, \beta) = p(X) = \frac{e^{\beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j}}$$

$$P(Y = 0 \mid X, \beta) = 1 - p(X) = \frac{1}{1 + e^{\beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j}}$$

$$odds = \frac{p(X)}{1 - p(X)}$$

$$\ln(odds) = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j$$

Estimação do vetor β via Máxima Verossimilhança:

$$L(\beta) = \prod_{i=1}^{n} p(X_i)^{Y_i} \left(1 - p(X)\right)^{1-Y_i} \quad \rightarrow \quad \hat{\beta}$$

Regra de classificação:

Outros valores podem ser adotados

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \sum_{j=1}^{p} X_{ij}\hat{\beta}_j}}{1 + e^{\hat{\beta}_0 + \sum_{j=1}^{p} X_{ij}\hat{\beta}_j}} \begin{cases} \geq 0.5 \rightarrow \text{classificar em G=1} \\ < 0.5 \rightarrow \text{classificar em G=2} \end{cases}$$

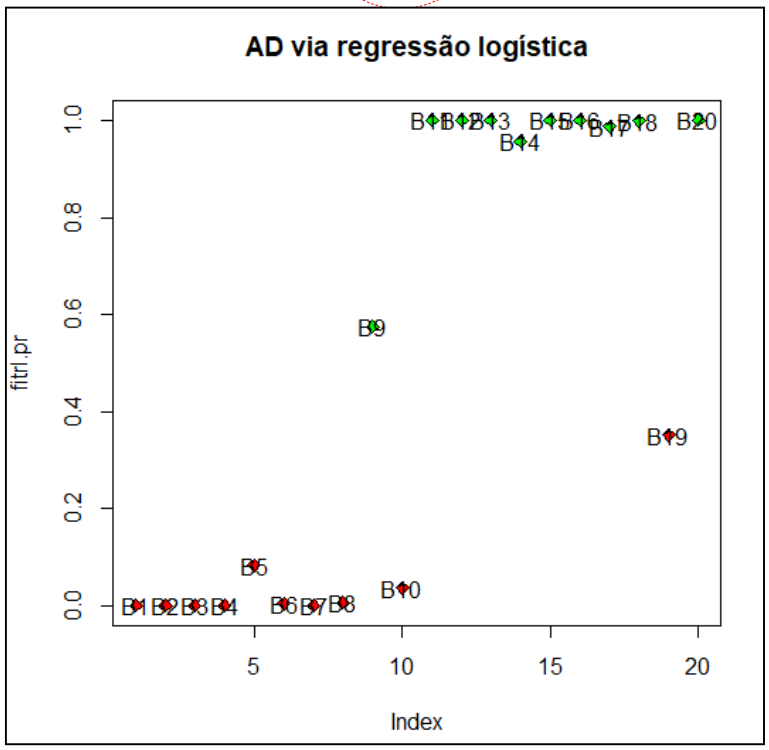
Pode ser estendido para regressão politômica (G>2)

Análise Discriminante via Modelo de Regressão Logística Dicotômica

Banco	Condição	Y1	Y2	Y3	Y4
B1	1	0,8888	0,7391	1,0255	0,3938
B2	1	1,6655	0,7268	0,878	0,0004
B3	1	2,2111	0,9166	0,9492	0,342
B4	1	1,4351	0,9133	0,9577	0,2325
B5	1	2,1414	0,002	1,0245	0,3966
B6	1	1,192	0,4972	1,034	0,3095
B7	1	1,5895	0,2593	1,0453	0,557
B8	1	1,3272	0,4126	1,0448	0,3482
B9	1	1,8847	0,388	0,9864	0,0337
B10	1	0,5229	0,9473	1,1244	0,118
n		10	10	10	10
Média		1,4852	0,5802	1,007	0,2732
D.P.		0,533	0,319	0,0674	0,1762
B11	2	0,4922	0,3166	1,1127	0,1628
B12	2	1,4427	0,0589	0,9019	0,1355
B13	2	0,5438	0,5358	1,03	0,1481
B14	2	0,1904	0,7087	0,9917	0,2625
B15	2	0,1102	0,7378	1,528	0,0783
B16	2	2,006	0,014	1,0321	0,0816
B17	2	0,2321	0,9234	0,9753	0,0045
B18	2	0,9019	0,1634	1,1414	0,5485
B19	2	1,9757	0,3395	0,9997	0,0751
B20	2	0,7276	0,3139	1,1077	0,2957
n		10	10	10	10
Média		0,862	0,4112	1,0821	0,1793
D.P.		0,712	0,3055	0,1726	0,1567

Cargas

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    25.10      63.20   0.397    0.691
Y1            -17.63      30.22  -0.584    0.560
Y2            -49.04      94.33  -0.520    0.603
Y3             29.56     148.89   0.199    0.843
Y4            -50.23      93.59  -0.537    0.591
```



Matriz de classificação

	p<0.5	p≥0.5
1	9	1
2	1	9

18/20 corretas