PRO2514 - Pesquisa Quantitativa em Gestão de Operações

Análise Discriminante

Prof. Dr. Renato de Oliveira Moraes

Bibliotecas a serem usadas nessa aula

- MASS
- tidyverse
- caret

Exemplo 1

Variável dependente com 2 categorias

X11 - Especificação da compra

0: não usa análise valor

1: usa análise de valor para cada compra

Variáveis independentes

X1: Delivery Speed

X2: Price Level

X3: Price Flexibility

X4: Manufacturer Image

X5: Service

X5: Salesforce Image

X7: Product Quality

Passos

- Importar a base de dados Hatco.XLSX
- Instalar a biblioteca MASS:
  install.packages("MASS")
  library (MASS)
- Construir o modelo de análise discriminante com funções lineares (supõe que matriz de covariância é a mesma entre os grupos):
  mod_discr <- lda (x11 ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data =Hatco)
- Olhar o modelo: print(mod_discr)

print(mod_discr)

Prior probabilities of groups:

0	1
0.4	0.6

Group means:

	x1	x2	x3	x4	x5	x6	x7
0	2,500	2,988	6,803	5,300	2,715	2,625	8,293
1	4,192	1,948	8,622	5,213	3,050	2,692	6,090

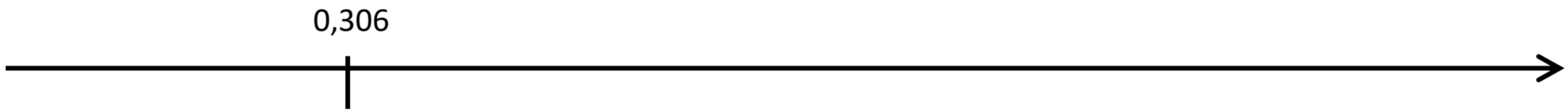
print(mod_discr) – continuação

Coefficients of linear discriminants:

	LD1
x1	0,608
x2	0,238
x3	0,507
x4	-0,057
x5	-0,429
x6	0,381
x7	-0,592

Função discriminante

$$0{,}608\,x1 + 0{,}238\,x2 + 0{,}507\,x3 - 0{,}057\,x4 - 0{,}429\,x5 + 0{,}381\,x6 - 0{,}592\,x7$$



Centro do Grupo 0: Empresas que não usam análise de valor

Cálculo do centroide dos grupos pela função discriminante

Valores médios das variáveis em cada um dos dois grupos

Grupo	x1	x2	x3	x4	x5	x6	x7
0	2,500	2,988	6,803	5,300	2,715	2,625	8,293
1	4,192	1,948	8,622	5,213	3,050	2,692	6,090

Função discriminante

0,608 x1 + 0,238 x2 + 0,507 x3 - 0,057 x4 - 0,429 x5 + 0,381 x6 - 0,592 x7

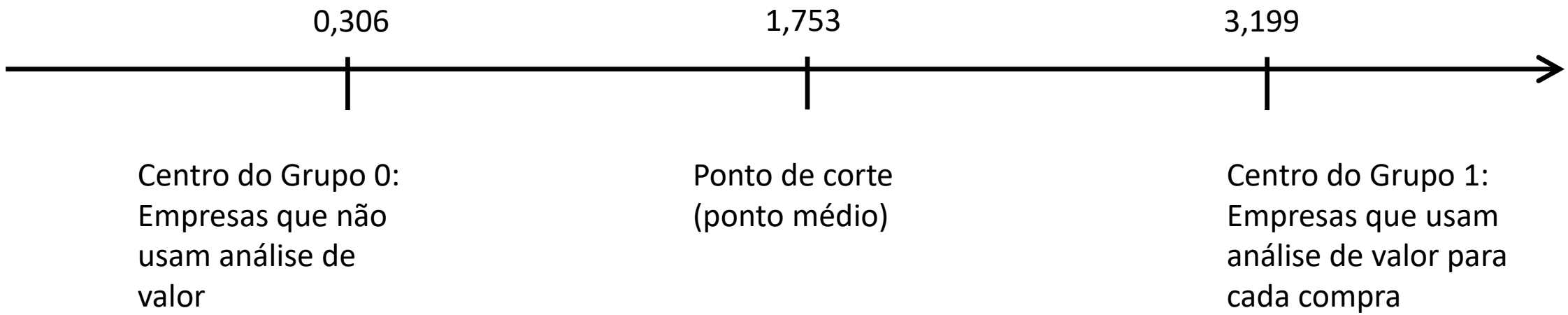
Centroides dos grupos – valor da função discriminante nos centroides dos grupos

Grupo 0: 0,306

Grupo 1: 3,199

Função discriminante

$$0{,}608\,x1 + 0{,}238\,x2 + 0{,}507\,x3 - 0{,}057\,x4 - 0{,}429\,x5 + 0{,}381\,x6 - 0{,}592\,x7$$

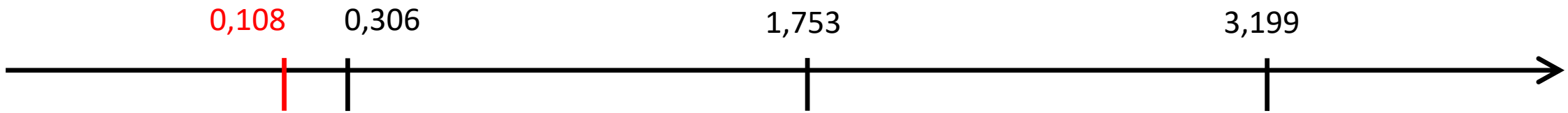


Onde classificar uma nova observação

x1	x2	x3	x4	x5	x6	x7
2	3	6	7	3	4	8

$$0{,}608\,x1 + 0{,}238\,x2 + 0{,}507\,x3 - 0{,}057\,x4 - 0{,}429\,x5 + 0{,}381\,x6 - 0{,}592\,x7 = 0{,}108$$

Nova observação classificada no Grupo 0



Centro do Grupo 0: Empresas que não usam análise de valor

Ponto de corte (ponto médio)

Centro do Grupo 1: Empresas que usam análise de valor para cada compra

Analisar a qualidade do modelo

- Instalar a biblioteca tidyverse
  install.packages("tidyverse")
  library (tidyverse)
- Fazer a predição:
  predicao = mod_discr %>% predict (Hatco)
- Olhar os resultados:
  head(predicao$class, 10) // classes atuais
  head(predicao$posterior, 10) // probabil de pertencer a cada grupo
  head(predicao$x, 10) // valor da função discriminante

Análise da taxa geral de acertos

```
l  = mean(predicao$class == Hatco$x11)

print(l)
```

0,9 // 90% de acertos

Construção da tabela de confusão

```
table(predicao$class, Hatco$x11, dnn=c("previsto","Real"))
```

	Real	
Previsto	0	1
0	37	7
1	3	53

Medidas de classificação binária

		Realidade	
		Positivo (1)	Negativo (0)
Previsão	Positivo (1)	VP – Verdadeiro Positivo	FP – Falso Positivo
	Negativo (0)	FN – Falso Negativo	VN – Verdadeiro Negativo

$$Eficácia = \frac{(VP + VN)}{(VP + FN + FP + VN)}$$

$$Precisão = \frac{VP}{(VP + FP)}$$

$$Sensibilidade = \frac{VP}{(VP + FN)}$$

$$Especificidade = \frac{VN}{(FP + VN)}$$

$$Eficácia = \frac{(VP + VN)}{(VP + FN + FP + VN)}$$

		Realidade	
		Positivo (1)	Negativo (0)
Previsão	Positivo (1)	**VP – Verdadeiro Positivo**	FP – Falso Positivo
	Negativo (0)	FN – Falso Negativo	**VN – Verdadeiro Negativo**

Taxa geral de acerto do modelo

$$Precisão = \frac{VP}{(VP + FP)}$$

		Realidade	
		Positivo (1)	Negativo (0)
Previsão	Positivo (1)	**VP – Verdadeiro Positivo**	FP – Falso Positivo
	Negativo (0)	FN – Falso Negativo	VN – Verdadeiro Negativo

A probabilidade de uma previsão positiva ser verdadeira

$$Sensibilidade = \frac{VP}{(VP + FN)}$$

		Realidade	
		Positivo (1)	Negativo (0)
Previsão	Positivo (1)	**VP – Verdadeiro Positivo**	FP – Falso Positivo
	Negativo (0)	FN – Falso Negativo	VN – Verdadeiro Negativo

A frequência com que o modelo identifica os casos positivos

$$Especificidade = \frac{VN}{(FP + VN)}$$

		Realidade	
		Positivo (1)	Negativo (0)
Previsão	Positivo (1)	VP – Verdadeiro Positivo	FP – Falso Positivo
	Negativo (0)	FN – Falso Negativo	**VN – Verdadeiro Negativo**

A frequência com que o modelo identifica os casos negativos

Construção da tabela de confusão

	Real	
Previsto	0	1
0	37	7
1	3	53

$$Eficácia = \frac{(37 + 53)}{100} = 0{,}90$$

$$Precisão = \frac{37}{(37 + 7)} \cong 0{,}841$$

$$Sensibilidade = \frac{37}{(37 + 3)} \cong 0{,}925$$

$$Especificidade = \frac{53}{(7 + 53)} \cong 0{,}883$$

- Instalar pacote caret
  install.packages('caret')
  library (caret)
- Preparar dados da classificação e predicação
  observado = as.factor (Hatco$x11)
  predito = predicao$class
- Gerar matriz de confusão
  confusionMatrix (predito, observado)
  caret::confusionMatrix

confusionMatrix (predito, observado)

```
          Reference
Prediction  0  1
         0 37  7
         1  3 53

               Accuracy : 0.9
                 95% CI : (0.8238, 0.951)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : 2.339e-11

                  Kappa : 0.7951

Mcnemar's Test P-Value : 0.3428

            Sensitivity : 0.9250
            Specificity : 0.8833
         Pos Pred Value : 0.8409
         Neg Pred Value : 0.9464
             Prevalence : 0.4000
         Detection Rate : 0.3700
   Detection Prevalence : 0.4400
      Balanced Accuracy : 0.9042
```

Eficácia

Sensibilidade

Especificidade

Precisão

Outras formar de avaliar o modelo discriminante

- Separação da base de dados em duas partes: desenvolvimento e teste. Essa abordagem é muito comum em ciências de dados.
  - Base de desenvolvimento é usada para construção do modelo – 70% a 80% da base original
  - Base de teste é usada para avaliar os acertos de classificação do modelo – de 20% a 30% da base original
- Validação cruzada. São construídos N modelos, onde N é o número de observações da base de dados. A cada construção de modelo, uma observação e ignorada e usada para avaliar a capacidade de classificação do modelo

Exemplo 2

Variável dependente com 3 categorias

x14: situação compra

- 1: nova
- 2: primeira recompra
- 3: outras

Variáveis independentes

X1: Delivery Speed

X2: Price Level

X3: Price Flexibility

X4: Manufacturer Image

X5: Service

X5: Salesforce Image

X7: Product Quality

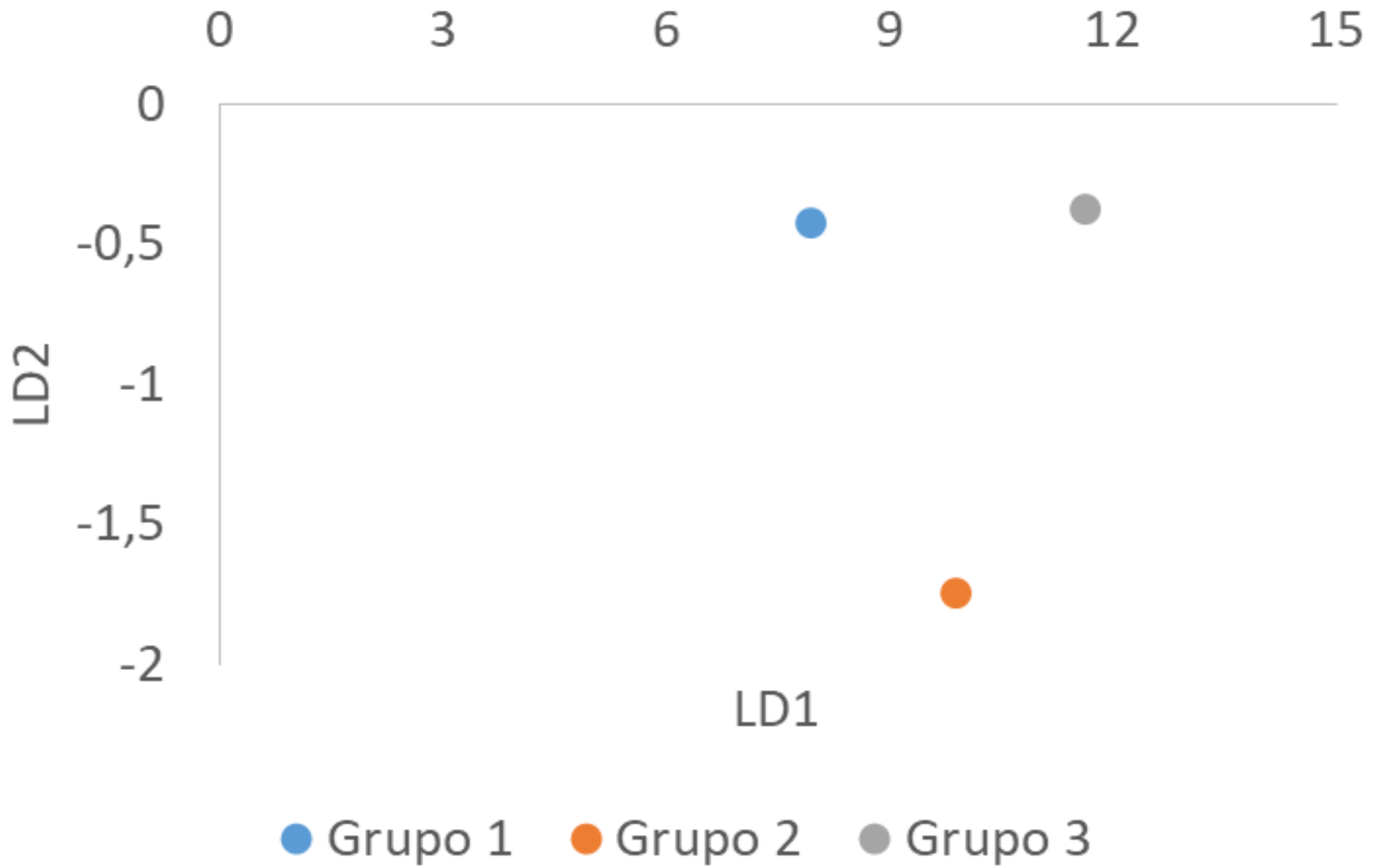
```
mod2_discr <- lda (x14 ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data =Hatco )

print(mod2_discr)
```

print(mod2_discr)

```
Call:
lda(x14 ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data = Hatco)

Prior probabilities of groups:
   1    2    3
0.34 0.32 0.34

Group means:
         x1       x2       x3       x4       x5       x6       x7
1 2.482353 2.094118 7.135294 4.958824 2.229412 2.614706 7.614706
2 3.421875 3.181250 7.296875 5.565625 3.284375 2.712500 7.315625
3 4.635294 1.864706 9.214706 5.238235 3.255882 2.670588 6.002941
```

print(mod2_discr)

```
Coefficients of linear discriminants:
           LD1         LD2
x1 -0.04164167 -0.89853715
x2 -0.14117700 -1.47110540
x3  0.78796904  0.28424773
x4  0.42300427 -0.49658943
x5  1.67295097  1.26898251
x6 -0.62939446  0.68743446
x7 -0.18984722  0.09056349

Proportion of trace:
   LD1    LD2
0.8545 0.1455
```


0
3
6
9
12
15
0
-0,5
-1
-1,5
-2
LD2
LD1
Grupo 1
Grupo 2
Grupo 3

```
predicao2 = mod2_discr %>% predict (Hatco)
mean(predicao2$class == Hatco$x14)
      0.86

table(predicao2$class, Hatco$x14, dnn=c("Previsto","Real"))

                Real
Previsto        1    2    3
           1    30   1    0
           2    3    24   2
           3    1    7    32
```

Gerar matriz de confusão e qualidade do modelo de classificação

```
observado2 = as.factor (Hatco$x14)
predito2 = predicao2$class

confusionMatrix (predito2, observado2)
```

confusionMatrix (predito2, observado2)

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3
         1 30  1  0
         2  3 24  2
         3  1  7 32
```

```
Overall Statistics

               Accuracy : 0.86
                 95% CI : (0.7763, 0.9213)
    No Information Rate : 0.34
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.7897

 Mcnemar's Test P-Value : 0.1888
```

confusionMatrix (predito2, observado2)

```
Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.8824   0.7500   0.9412
Specificity            0.9848   0.9265   0.8788
Pos Pred Value         0.9677   0.8276   0.8000
Neg Pred Value         0.9420   0.8873   0.9667
Prevalence             0.3400   0.3200   0.3400
Detection Rate         0.3000   0.2400   0.3200
Detection Prevalence   0.3100   0.2900   0.4000
Balanced Accuracy      0.9336   0.8382   0.9100
```

Exercício – Análise Discriminante Cereais Matinais

Variáveis

- Brand
- Manufacturer : G, K e Q
- Calories
- Protein
- Fat
- Sodium
- Fiber
- Carbohydrates
- Sugar
- Potassium

Construir um modelo discriminante

Variável Dependente: Manufacturer (G, K e Q)

Variáveis Independentes:

- Calories
- Protein
- Fat
- Sodium
- Fiber
- Carbohydrates
- Sugar
- Potassium