

# Fundamentos de Estatística e Aprendizado de Máquina

Capítulo 6. Regressão Logística

Prof. Máiron Chaves

# Fundamentos de Estatística e Aprendizado de Máquina

---

Aula 6.1. Classificação Binária e Interpretação dos Coeficientes

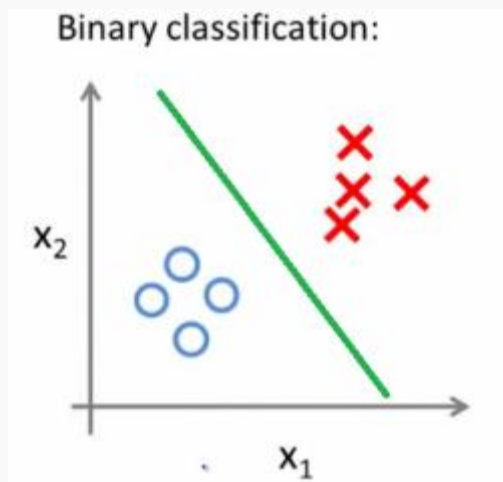
Prof. Máiron Chaves

# Nesta aula



- ☐ Regressão Linear vs Regressão Logística.
- ☐ Eventos Binários.
- ☐ Interpretando os coeficientes.

# Classificação Binária



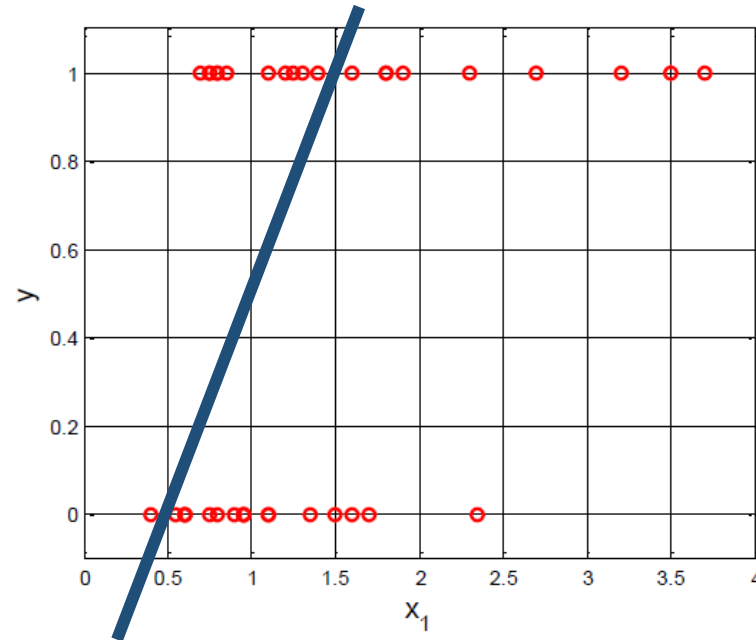
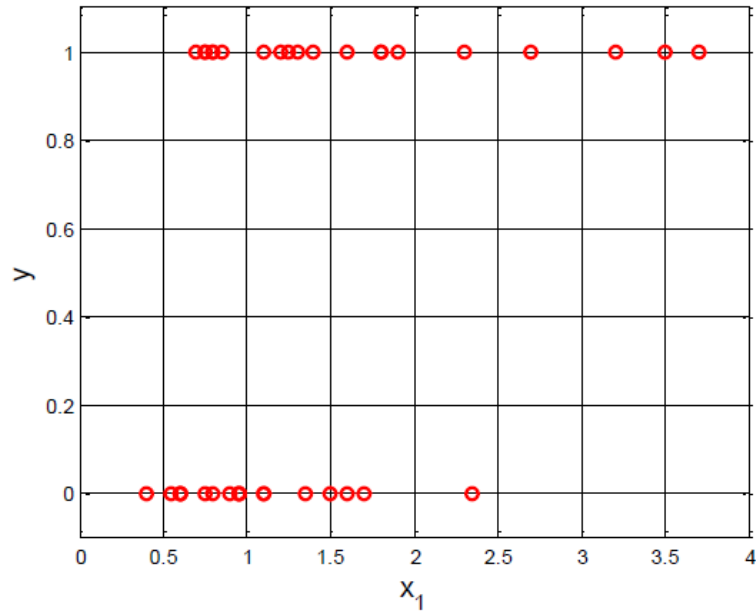
- Cliente Compra/Não Compra.
- Cliente Cancela/Não Cancela.
- Cliente Pagará/Não Pagará.
- Equipamento Estragará/Não Estragará.
- Contratar/Não Contratar um candidato.
- Tumor benigno/Tumor maligno.

# Regressão Logística



$Y \in \{1,0\}$

$$\hat{y} = \beta_0 + \beta_1 * X_1 + \dots + \beta_p * X_p$$

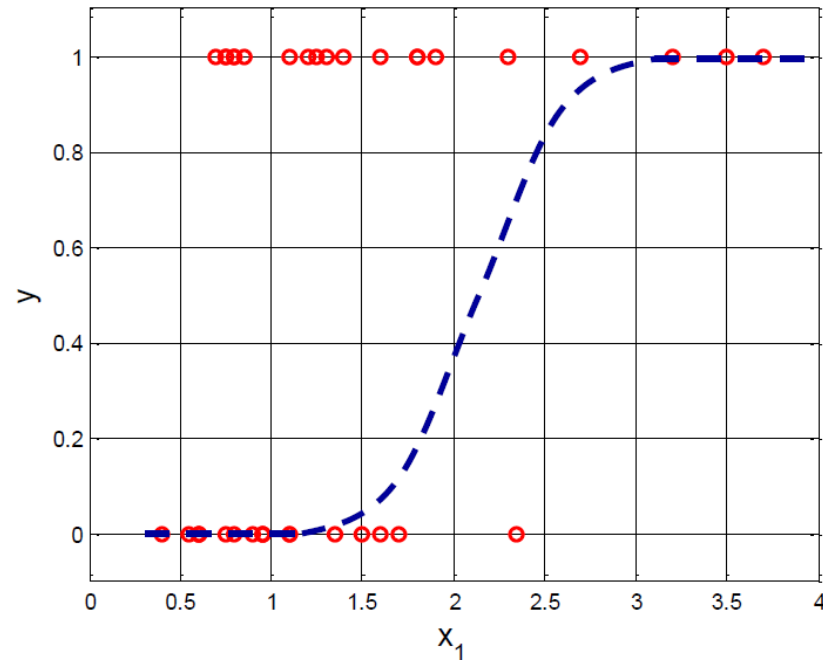
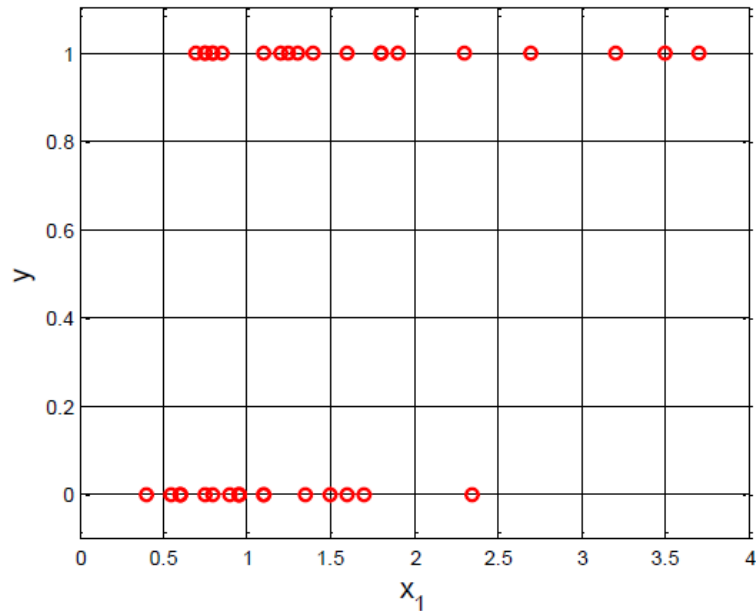


# Regressão Logística



$Y \in \{1,0\}$

$$\hat{y} = \Pr(y = 1 | x) = \frac{e^{x^T \beta + \beta_0}}{1 + e^{x^T \beta + \beta_0}}$$

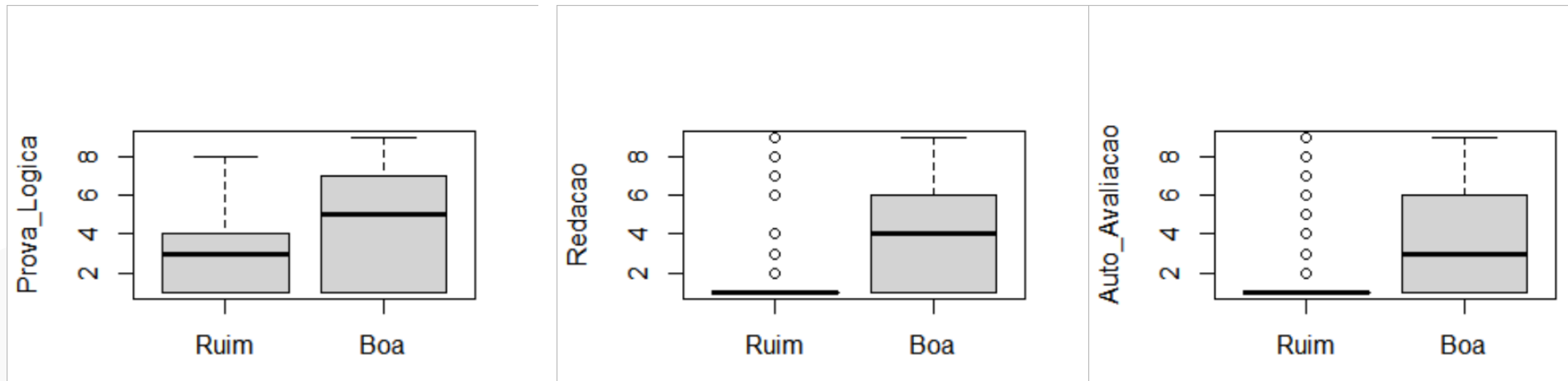


# Regressão Logística



Prova_Logica	Redacao	Psicotecnico	Dinamica_Grupo	Fit_Cultural	Ingles	Avaliacao_RH	Auto_Avaliacao	Demograficos	Estado_Civil	Escolaridade	Classe
2	1	1	1	1	2	1	2	1	1 Divorciado	Superior	Ruim
2	1	1	1	1	2	1	3	1	1 Divorciado	Segundo_Grau	Ruim
5	1	1	1	1	2	1	2	1	1 Casado	Superior	Ruim
5	4	6	8	4	1	8	1	1	1 Divorciado	Superior	Boa
5	3	3	1	2	1	2	1	1	1 Solteiro	Segundo_Grau	Ruim
2	3	1	1	3	1	1	1	1	1 Solteiro	Pos_Graduacao	Ruim
3	5	7	8	8	9	7	1	1	7 Divorciado	Segundo_Grau	Boa
1	5	6	1	6	1	7	7	7	1 Casado	Segundo_Grau	Boa
1	9	8	7	6	4	7	1	1	3 Casado	Segundo_Grau	Boa
4	1	1	1	2	1	3	1	1	1 Casado	Segundo_Grau	Ruim
5	1	1	1	2	1	3	1	1	1 Casado	Pos_Graduacao	Ruim
8	1	1	1	3	6	3	9	9	1 Solteiro	Pos_Graduacao	Boa
1	1	3	1	2	1	2	1	1	1 Divorciado	Superior	Ruim
1	1	1	2	1	1	1	1	1	1 Solteiro	Pos_Graduacao	Ruim
3	4	5	2	6	8	4	1	1	1 Solteiro	Superior	Boa
4	3	3	1	2	1	3	3	3	1 Divorciado	Segundo_Grau	Ruim
3	3	2	1	3	1	3	6	6	1 Solteiro	Segundo_Grau	Ruim
2	1	1	1	2	1	1	1	1	1 Solteiro	Superior	Ruim
1	1	1	1	2	1	2	1	1	1 Solteiro	Segundo_Grau	Ruim
1	1	1	1	2	1	1	1	1	1 Casado	Segundo_Grau	Ruim
8	1	1	1	5	1	8	1	1	6 Solteiro	Segundo_Grau	Boa
8	7	4	4	5	3	5	1	1	1 Solteiro	Pos_Graduacao	Boa
1	1	1	1	1	1	2	1	1	1 Casado	Segundo_Grau	Ruim
2	1	1	1	2	1	1	1	1	1 Solteiro	Segundo_Grau	Ruim
1	8	8	4	1	1	8	1	1	1 Casado	Superior	Boa
5	1	1	2	2	1	2	1	1	1 Casado	Superior	Ruim
3	1	1	1	2	1	2	1	1	2 Divorciado	Pos_Graduacao	Ruim
3	1	1	1	2	1	3	1	1	1 Casado	Superior	Ruim

# Regressão Logística



Prova_Logica	Redacao	Auto_Avaliacao
Min. :1.000	Min. :1.000	Min. :1.000
1st Qu.:1.000	1st Qu.:1.000	1st Qu.:1.000
Median :3.000	Median :1.000	Median :1.000
Mean :3.531	Mean :2.272	Mean :2.082
3rd Qu.:5.000	3rd Qu.:3.000	3rd Qu.:2.000
Max. :9.000	Max. :9.000	Max. :9.000



# Regressão Logística



Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-3.84564	0.27820	-13.823	< 2e-16	***
Prova_Logica	0.21409	0.05184	4.130	3.63e-05	***
Redacao	0.69158	0.07437	9.299	< 2e-16	***
Auto_Avaliacao	0.42409	0.06728	6.303	2.91e-10	***

$$\hat{y} = -3,8458 + (0,2143 * Prova\_Logica) + (0,6914 * Redacao) + (0,4238 * Auto\_Avaliacao)$$

# Regressão Logística



$\beta_1 = \text{Prova\_Logica} = \exp(0.2143) = 2,7182^{0,2143} = \underline{1,23}$  – Ou seja, mantendo as demais variáveis constantes, para cada ponto a mais na prova de lógica, o candidato aumenta em média 1,23 vezes as chances de pertencer à classe ‘Boa’.

$\beta_2 = \text{Redacao} = \exp(0.6914) = 2,7182^{0,6914} = \underline{1,99}$  – Ou seja, mantendo as demais variáveis constantes, para cada ponto a mais na prova de redação, o candidato aumenta em média, 1,99 vezes as chances de pertencer à classe ‘Boa’.

$\beta_3 = \text{Auto\_Avaliacao} = \exp(0,4238) = 2,7182^{0,4238} = \underline{1,5278}$  – Ou seja, mantendo as demais variáveis constantes, para cada ponto a mais na auto avaliação, o candidato aumenta em 1,52 vezes as chances de pertencer à classe ‘Boa’.

# Conclusão



- ✓ Classificação Binária.
- ✓ Interpretando coeficientes da Regressão Logística.

# Na próxima aula



- ❑ Realizando previsões com a Regressão Logística.

# Fundamentos de Estatística e Aprendizado de Máquina

---

Aula 6.2. Realizando Predição com a Regressão Logística

Prof. Máiron Chaves

# Nesta aula



- ☐ Função logística.
- ☐ Realizando predições.
- ☐ Seleção automática de preditores.

# Regressão Logística



$$\hat{y} = -3,8458 + (0,2143 * Prova\_Logica) + (0,6914 * Redacao) + (0,4238 * Auto\_Avaliacao)$$

$$\hat{y} = \frac{e^{-3,8458+(0,2143*Prova\_Logica)+(0,6914*Redacao)+(0,4238*Auto\_Avaliacao)}}{1 + e^{-3,8458+(0,2143*Prova\_Logica)+(0,6914*Redacao)+(0,4238*Auto\_Avaliacao)}}$$

Supondo que o candidato tire **3** em Prova\_Logica, **5** em Redacao e **1** em Auto\_Avaliacao.

A probabilidade dele(a) pertencer à classe de interesse 'Boa' fica:

$$\hat{y} = \frac{e^{-3,8458+(0,2143*3)+(0,6914*5)+(0,4238*1)}}{1 + e^{-3,8458+(0,2143*3)+(0,6914*5)+(0,4238*1)}}$$

$$\hat{y} = \frac{1,9697}{2,9697}$$

$$\hat{y} = 0,6632 \text{ (ou 66,32\%)}$$

# Regressão Logística



Prova_Logica	Redacao	Auto_Avaliacao	Classe	probabilidade
2	1	1	Ruim	0.09097517
2	1	1	Ruim	0.09097517
5	1	1	Ruim	0.15982827
5	4	1	Boa	0.60234090
5	3	1	Ruim	0.43134889
2	3	1	Ruim	0.28523727
3	5	1	Boa	0.66343073
1	5	7	Boa	0.94240236
1	9	1	Boa	0.95332501
4	1	1	Ruim	0.13312577
5	1	1	Ruim	0.15982827



# Regressão Logística



Ponto de Corte: 0,5

Prova_Logica	Redacao	Auto_Avaliacao	Classe	probabilidade	Predicao
2	1	1	Ruim	0.09097517	Ruim
2	1	1	Ruim	0.09097517	Ruim
5	1	1	Ruim	0.15982827	Ruim
5	4	1	Boa	0.60234090	Boa
5	3	1	Ruim	0.43134889	Ruim
2	3	1	Ruim	0.28523727	Ruim
3	5	1	Boa	0.66343073	Boa
1	5	7	Boa	0.94240236	Boa
1	9	1	Boa	0.95332501	Boa
4	1	1	Ruim	0.13312577	Ruim
5	1	1	Ruim	0.15982827	Ruim

# Regressão Logística - Stepwise



Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-7.29948	0.63721	-11.455	< 2e-16	***
Prova_Logica	0.23650	0.07342	3.221	0.00128	**
Redacao	0.20571	0.10626	1.936	0.05288	.
Psicotecnico	0.40900	0.10350	3.952	7.75e-05	***
Dinamica_Grupo	0.19739	0.09318	2.118	0.03414	*
Fit_Cultural	0.28630	0.10435	2.744	0.00608	**
Ingles	-0.08452	0.03998	-2.114	0.03452	*
Avaliacao_RH	0.64018	0.09103	7.033	2.03e-12	***
Auto_Avaliacao	0.13054	0.08629	1.513	0.13031	
Demograficos	0.77714	0.15599	4.982	6.30e-07	***
Estado_CivilDivorciado	0.09922	0.38401	0.258	0.79611	
Estado_Civilsolteiro	-0.36401	0.40339	-0.902	0.36686	
Escolaridadessegundo_Grau	0.07021	0.37122	0.189	0.84998	
Escolaridadesuperior	-0.56484	0.38455	-1.469	0.14188	

---

signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Regressão Logística - Stepwise



Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	-7.35786	0.55520	-13.253	< 2e-16	***
Prova_Logica	0.22033	0.07196	3.062	0.002201	**
Redacao	0.22923	0.10710	2.140	0.032335	*
Psicotecnico	0.39374	0.10252	3.841	0.000123	***
Dinamica_Grupo	0.20777	0.09192	2.260	0.023793	*
Fit_Cultural	0.26802	0.10217	2.623	0.008708	**
Ingles	-0.08332	0.03936	-2.117	0.034275	*
Avaliacao_RH	0.62276	0.08834	7.049	1.80e-12	***
Auto_Avaliacao	0.13426	0.08499	1.580	0.114167	
Demograficos	0.73248	0.15198	4.820	1.44e-06	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Conclusão



- ✓ Realizando previsões com a Regressão Logística.
- ✓ Stepwise na Regressão Logística.

# Na próxima aula



- ☐ Avaliando a capacidade preditiva do modelo ajustado.

# Fundamentos de Estatística e Aprendizado de Máquina

---

Aula 6.3. Avaliando a Capacidade Preditiva

Prof. Máiron Chaves

# Nesta aula



- ☐ Acurácia.
- ☐ Sensibilidade.
- ☐ Especificidade.
- ☐ Curva ROC.

# Regressão Logística



Ponto de corte: 0,5

Classe_Predita	Classe_Original	
	Boa	Ruim
Boa	169	23
Ruim	72	435

$$\text{Acurácia} = \frac{169 + 435}{169 + 23 + 72 + 435} = 86,40\%$$

$$\text{Sensitividade} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Negativos}} = \frac{169}{169 + 72} = 70,12\%$$

$$\text{Especificidade} = \frac{\text{Verdadeiros Negativos}}{\text{Verdadeiros Negativos} + \text{Falsos Positivos}} = \frac{435}{435 + 23} = 94,97\%$$



# Regressão Logística



Ponto de corte: 0,9

Classe_Predita	Classe_Original	
	Boa	Ruim
Boa	98	9
Ruim	143	449

$$Acurácia = \frac{98 + 449}{98 + 9 + 143 + 449} = 78,25\%$$

$$Sensitividade = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Negativos} = \frac{98}{98 + 143} = 40,66\%$$

$$Especificidade = \frac{Verdadeiros\ Negativos}{Verdadeiros\ Negativos + Falsos\ Positivos} = \frac{449}{449 + 9} = 98,03\%$$

# Regressão Logística



Ponto de corte: 0,1

Classe_Predita	Classe_Original	
	Boa	Ruim
Boa	229	291
Ruim	12	167

$$Acurácia = \frac{229 + 167}{229 + 167 + 12 + 291} = 56,65\%$$

$$Sensitividade = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Negativos} = \frac{229}{229 + 12} = 95,02\%$$

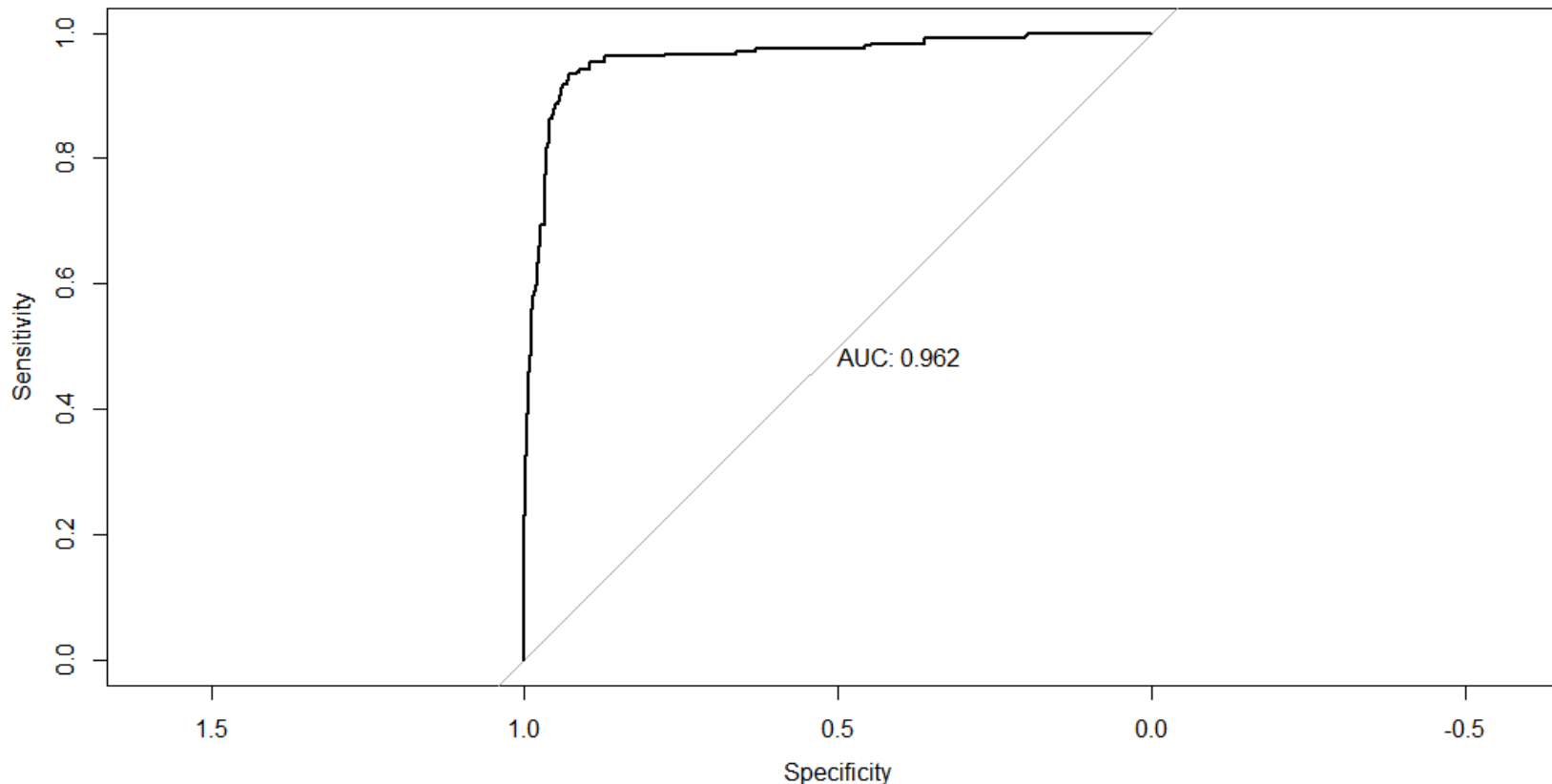
$$Especificidade = \frac{Verdadeiros\ Negativos}{Verdadeiros\ Negativos + Falsos\ Positivos} = \frac{167}{167 + 291} = 36,46\%$$

# Regressão Logística



Curva ROC 'receiver operating characteristic curve'

Em vez de verificar manualmente os pontos de corte, podemos criar uma curva ROC que irá varrer todos os cortes possíveis e traçar a sensibilidade e especificidade.



# Conclusão



- ✓ Acurácia.
- ✓ Sensibilidade.
- ✓ Especificidade.
- ✓ Curva ROC.
- ✓ Área sobre a curva ROC.

# Na próxima aula



- ❑ Análise de Sensibilidade e Especificidade

# Fundamentos de Estatística e Aprendizado de Máquina

---

Aula 6.4. Análise de Sensibilidade e Especificidade

Prof. Máiron Chaves

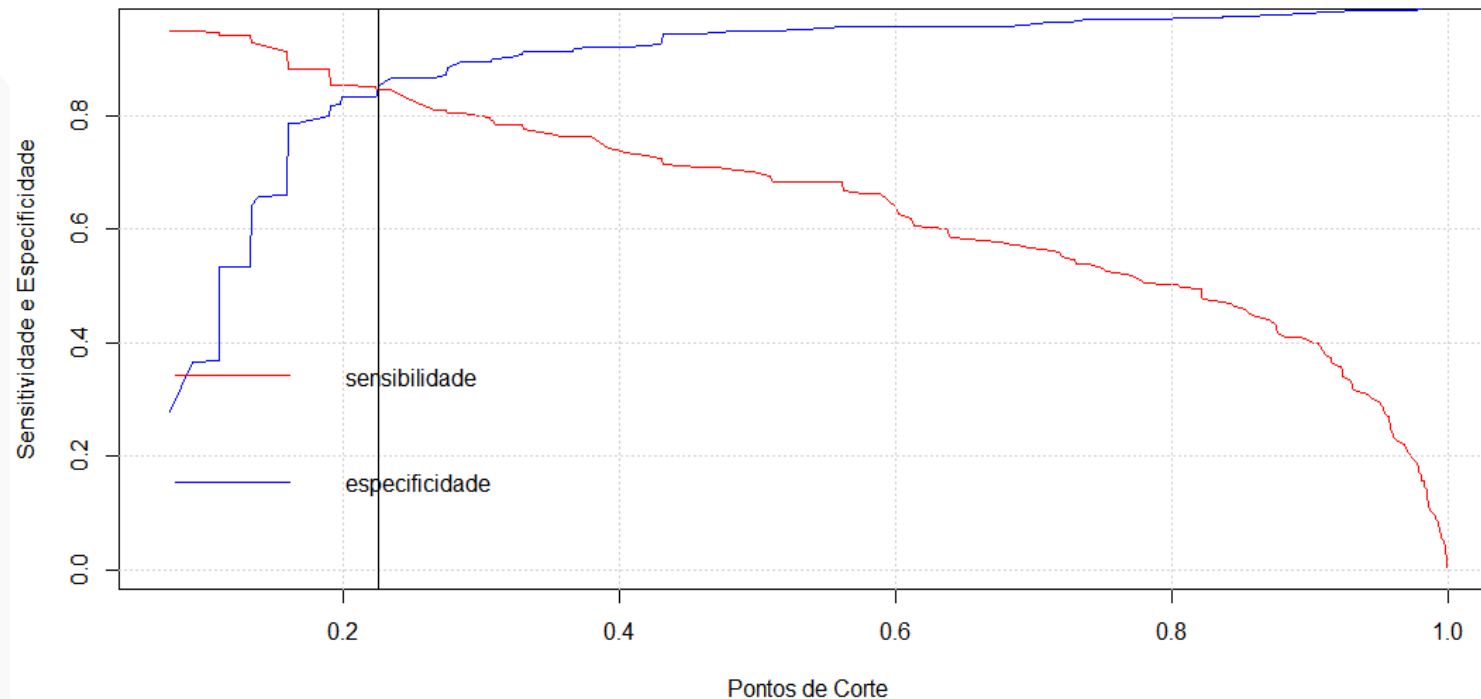
# Nesta aula



- ☐ Gráfico de Sensibilidade e Especificidade.
- ☐ Identificando o ponto de corte 'ótimo'.

# Regressão Logística

A ideia da Análise de Sensibilidade e Especificidade é simular várias matrizes de confusão, através de vários pontos de corte diferentes e identificar aquela matriz de confusão que nos dará tanto a maior Sensibilidade quanto a maior Especificidade.





# Regressão Logística



Ponto de corte: 0,225

Classe_Predita	Classe_Original	
	Boa	Ruim
Boa	204	75
Ruim	37	383

$$Acurácia = \frac{204 + 383}{204 + 75 + 37 + 383} = 83,97\%$$

$$Sensitividade = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Negativos} = \frac{204}{204 + 37} = 84,64\%$$

$$Especificidade = \frac{Verdadeiros\ Negativos}{Verdadeiros\ Negativos + Falsos\ Positivos} = \frac{383}{383 + 75} = 83,62\%$$

# Conclusão



- ✓ Análise de Sensibilidade e Especificidade.

# Na próxima aula



- ❑ Regressão Logística com o R.

# Fundamentos de Estatística e Aprendizado de Máquina

---

Aula 6.5. Regressão Logística com o R

Prof. Máiron Chaves

# Nesta aula



- ❑ Regressão Logística com o R.

# Regressão Logística com o R



INSTITUTO DE GESTÃO E TECNOLOGIA DA INFORMAÇÃO

## Estatística Computacional – Regressão Logística no R

```
#####
```

```
#####      Regressao Logística      #####
```

```
##      AED - Capitulo 06 - Prof. Máiron Chaves      ####
```

```
#####
```

```
#Copie este código, cole no seu R e execute para ver os resultados
```

```
rm(list = ls()) #Limpa memória do R
```

```
#Instala e carrega biblioteca para gerar a curva ROC
```

```
install.packages('pROC') #Instala
```

```
library(pROC) #Carrega
```

```
#Monte o dataset
```

```
dados <- data.frame(Prova_Logica = c(2, 2, 5, 5, 5, 2, 3, 2, 1, 4,
```

# Conclusão



- ✓ Regressão Logística com o R.