

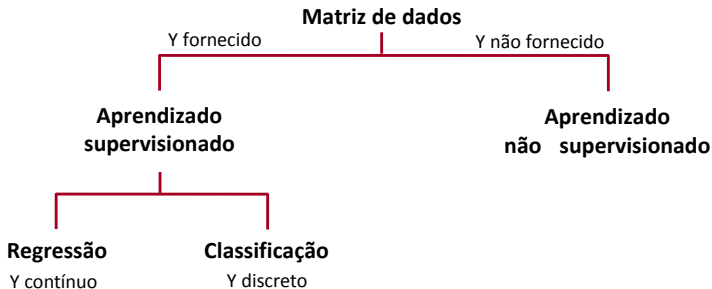


*Universidade Federal do Paraná*  
*Laboratório de Estatística e Geoinformação - LEG*

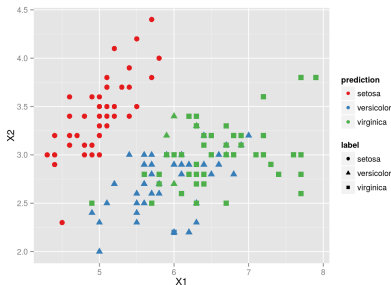


# Classificação

Eduardo Vargas Ferreira



- Em muitos problemas, a variável  $Y$  assume valores em um conjunto não ordenado  $\mathcal{C}$ , por exemplo:
  - ★ E-mail  $\in \{\text{spam, ham}\}$ ;
  - ★ Dígito  $\in \{0, 1, \dots, 9\}$ ;
  - ★ Alzheimer  $\in \{\text{com Alzheimer, sem Alzheimer}\}$ ;
- Nestes casos, estamos diante de um **problema de classificação**;



- Considere um problema binário, em que  $Y$  assume somente dois valores,  $c_1$  ou  $c_2$ . Para um dado  $x$ , escolheremos  $c_1$  quando

$$P(Y = c_1|x) \geq P(Y = c_2|x),$$

- Tal classificador é conhecido como **Classificador de Bayes**. Escolhemos nossa função, tal que,

$$h(x) = \underset{d \in \{c_1, c_2\}}{\operatorname{argmax}} P(Y = d|x).$$

**O classificador de Bayes é um padrão ouro inalcançável!**

- A solução é então estimar  $P(Y = c|\mathbf{x})$ , para  $i \in \mathcal{C}$ , ou seja

★ Estimamos  $P(Y = c|\mathbf{x})$  para cada categoria  $c \in \mathcal{C}$ ;

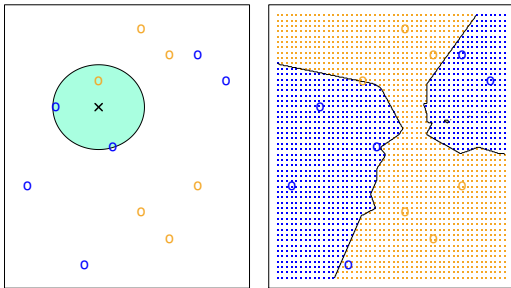
★ Tomamos  $\hat{h}(\mathbf{x}) = \underset{c \in \mathcal{C}}{\operatorname{argmax}} \hat{P}(Y = c|\mathbf{x})$ .

- Essa abordagem é conhecida como **plug-in classifier**.

# K-Nearest Neighbors

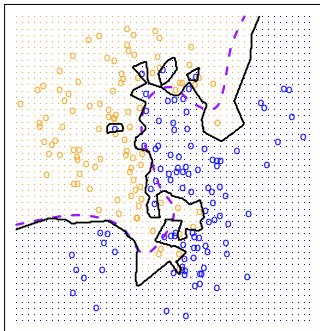
- O KNN estima a distribuição condicional de  $Y|X$  de acordo com as classes dos  $K$  vizinhos de determinada observação  $x_0$ , ou seja:

$$P(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} \mathbb{I}(y_i = j).$$

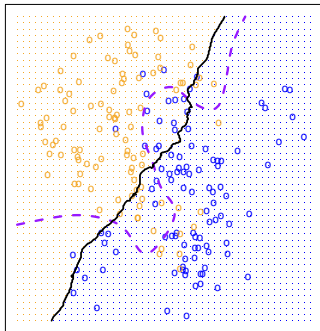


- A escolha de  $K$  tem um efeito drástico no classificador KNN obtido

KNN:  $K=1$



KNN:  $K=100$

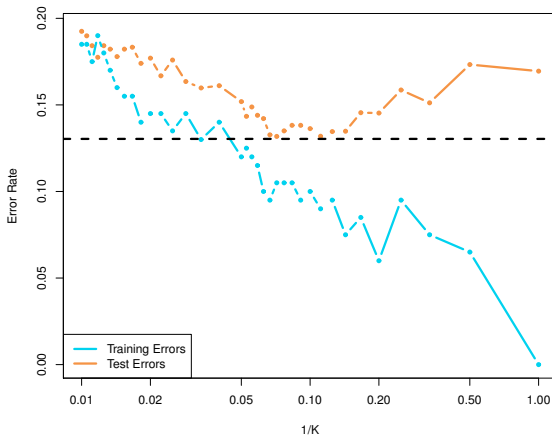




# K-Nearest Neighbors



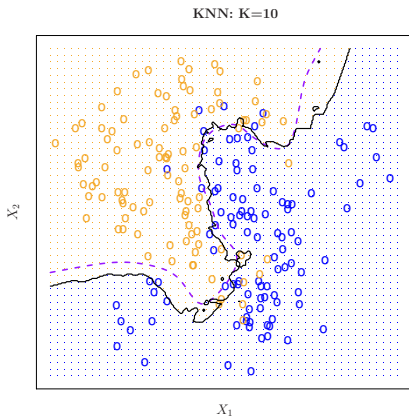
- Temos que escolhê-lo de acordo com o resultado do teste. A linha pontilhada representa o classificador de Bayes.



# K-Nearest Neighbors



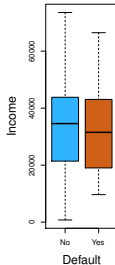
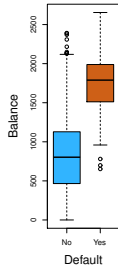
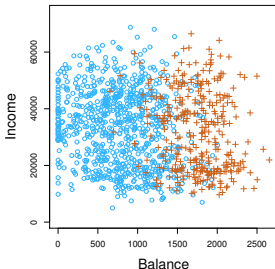
- Temos que escolhê-lo de acordo com o resultado do teste. A linha pontilhada representa o classificador de Bayes.



# Regressão logística

# Exemplo: Inadimplência no cartão de crédito

- Nosso objetivo é prever se um cliente será ou não inadimplente no próximo mês. Para tanto, temos três variáveis explicativas:
  - ★ **Student**: se o cliente é ou não estudante;
  - ★ **Income**: rendimento anual do cliente;
  - ★ **Balance**: o valor devido no mês atual.



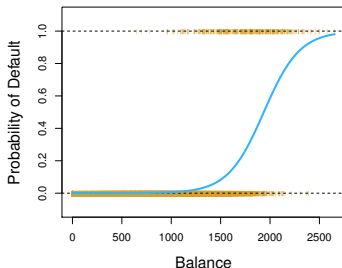
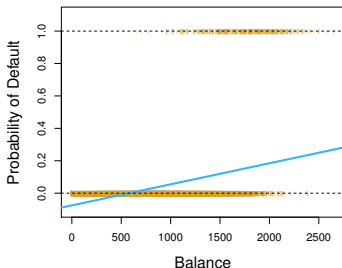
# Podemos utilizar regressão linear?



- Suponha que para classificação da variável **Default** codificamos da forma:

$$Y = \begin{cases} 0, & \text{se No ,} \\ 1, & \text{se Yes .} \end{cases}$$

- Podemos simplesmente realizar uma regressão linear de  $Y$  em  $X$  e classificar como **Yes** se  $\hat{Y} > 0.5$ ?



- A regressão logística utiliza a forma

$$P(Y = 1|X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}.$$

- Com um pouco de algebrismo, chegamos em

$$\log \left[ \frac{P(Y = 1|X)}{1 - P(Y = 1|X)} \right] = \beta_0 + \beta_1 X.$$

Variável	Coefficiente	Erro padrão	Estatística <i>t</i>	p-valor
Intercepto	-3,5041	0,0707	-49,55	< 0,0001
Student [Yes]	0,4049	0,1150	3,52	0,0004

$$\log \left[ \frac{P(\text{Default} = \text{Yes} \mid \text{Student})}{1 - P(\text{Default} = \text{Yes} \mid \text{Student})} \right] = -3,5241 + 0,4049 \cdot \text{Student [Yes]}$$

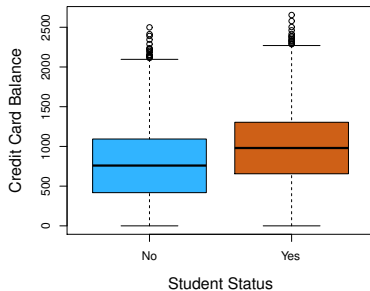
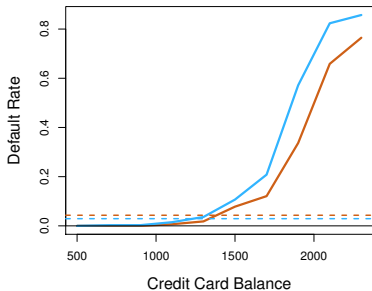
- Agora o caso de mais de um preditor, o modelo geral torna-se

$$\log \left[ \frac{P(Y = 1|\mathbf{X})}{1 - P(Y = 1|\mathbf{X})} \right] = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p.$$

Variável	Coefficiente	Erro padrão	Estatística <i>t</i>	p-valor
Intercepto	-10,8690	0,4923	-22,08	< 0,0001
Balance	0,0057	0,0002	24,74	< 0,0001
Income	0,0030	0,0082	0,37	0,7115
Student [Yes]	-0,6468	0,2362	-2,74	0,0062

- Por que o coeficiente de **Student** é negativo agora, enquanto era positivo anteriormente? **Confundimento**.

- Os resultados são diferentes, especialmente quando existe correlação entre os preditores.





# Regressão multinomial

- Até agora, discutimos o caso de regressão logística com duas classes. É fácil generalizar para mais classes

$$P(Y = k|\mathbf{X}) = \frac{e^{\beta_{0k} + \beta_{1k}X_1 + \dots + \beta_{pk}X_p}}{\sum_{l=1}^K e^{\beta_{0l} + \beta_{1l}X_1 + \dots + \beta_{pl}X_p}}$$

- Por exemplo, podemos classificar um paciente na sala de emergência de acordo com seu sintoma

$$Y = \begin{cases} 1, & \text{se AVC,} \\ 2, & \text{se overdose de droga,} \\ 3, & \text{se ataque epilético.} \end{cases}$$

# Outra abordagem

- Uma alternativa para estimar  $P(Y|X)$  consiste em modelar a distribuição de  $X$ , em cada classe separadamente, utilizando o **Teorema de Bayes**:

$$P(Y = k|X = x) = \frac{P(Y = k)P(X = x|Y = k)}{P(X = x)}$$

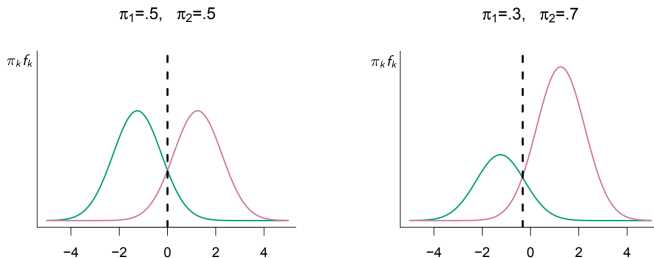
- Que escrevendo de outra forma fica

$$P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)}$$

- Então temos que

$$\delta_k(x) \propto \operatorname{argmax} \pi_k f_k(x)$$

- $\pi_k = P(Y = k)$  é a **probabilidade marginal** ou **priori** para classe  $k$ . Pode ser estimada utilizando as proporções amostrais em cada classe.

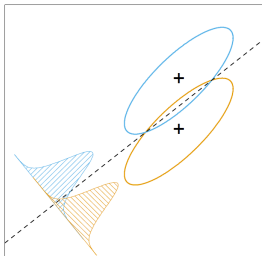


- $f_k(x) = P(X = x|Y = k)$  é a **densidade** para  $X$  na classe  $k$  (diferentes distribuições levam a diferentes métodos).

# Análise de discriminante

- Ao considerarmos para  $f_k(x)$  a distribuição Normal em cada classe, nos leva à **análise de discriminante linear** ou **quadrática**, pois

$$\begin{aligned}\delta_k(x) &\propto \operatorname{argmax} \pi_k f_k(x) \\ &= \operatorname{argmax} \left\{ \log \pi_k - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} \langle x - \mu_k, \Sigma_k^{-1} (x - \mu_k) \rangle \right\}.\end{aligned}$$



$$\hat{\pi}_k = \frac{n_k}{n}$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

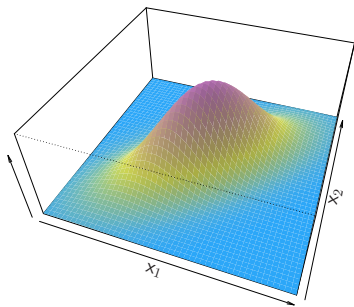
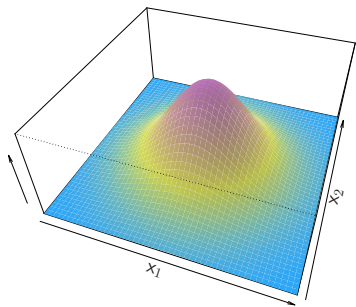
$$\hat{\Sigma}_k = \frac{1}{n_k - 1} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^t$$

- Quando  $f_k(x)$  possui matriz de covariância,  $\Sigma_k$ , diferente em cada classe, temos a **análise de discriminante quadrático (ADQ)**

$$\begin{aligned}\delta_k(x) &\propto \operatorname{argmax} \pi_k f_k(x) \\ &= \operatorname{argmax} \left\{ \log \pi_k - \frac{1}{2} \log |\Sigma_k| - \frac{1}{2} (x - \mu_k)^t \Sigma_k^{-1} (x - \mu_k) \right\}.\end{aligned}$$

- Se todas as classes compartilharem o mesmo  $\Sigma = \sum_k \frac{n_k - 1}{n - K} \hat{\Sigma}_k$ , estamos diante da **análise de discriminante linear (ADL)**

$$\begin{aligned}\delta_k(x) &\propto \operatorname{argmax} \pi_k f_k(x) \\ &= \operatorname{argmax} \left\{ \log \pi_k - \frac{1}{2} \mu_k^t \Sigma^{-1} \mu_k + x^t \Sigma^{-1} \mu_k \right\}.\end{aligned}$$

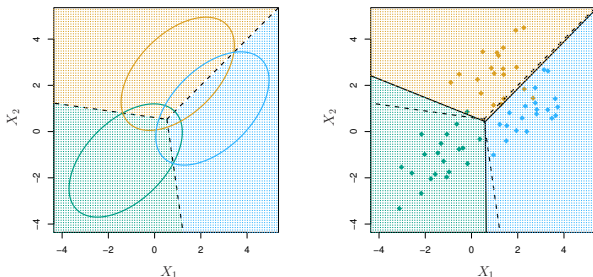




# Ilustração: $p = 2$ e $k = 3$ classes

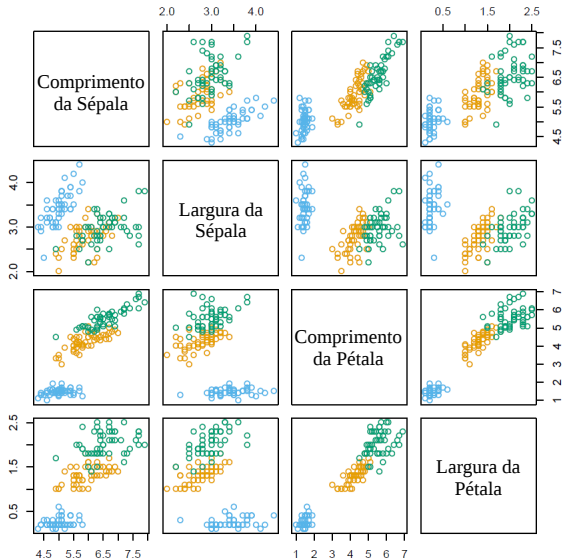


- No exemplo abaixo, temos  $\pi_1 = \pi_2 = \pi_3 = 1/3$ ;

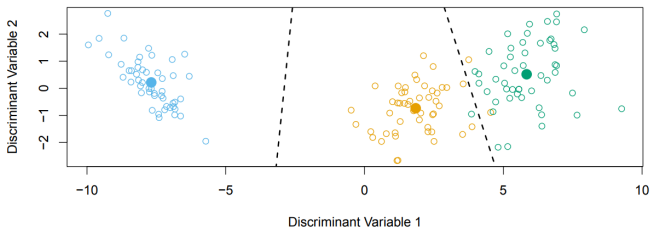


- A linha pontilhada é conhecida como **fronteira de decisão de Bayes** (*Bayes decision boundaries*);

# Exemplo: Iris Data

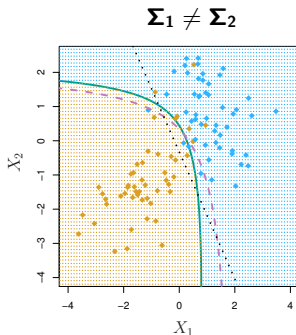
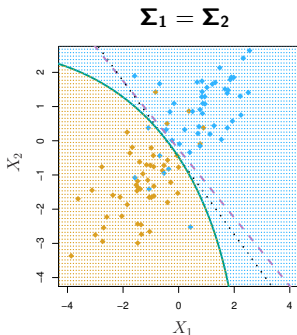


- Temos 4 variáveis, 3 espécies com 50 observações em cada classe;



- Análise de discriminante linear classifica corretamente 147/150 observações dos dados de treino.

- No exemplo, temos a fronteira de decisão de Bayes em rosa, ADL pontilhado e ADQ em verde, em um problema com 2 classes;



- Regressão logística maximiza a **verossimilhança condicional**

$$\prod_i p(x_i, y_i) = \underbrace{\prod_i p(y_i|x_i)}_{\text{logística}} \underbrace{\prod_i g(x_i)}_{\text{ignorado}}$$

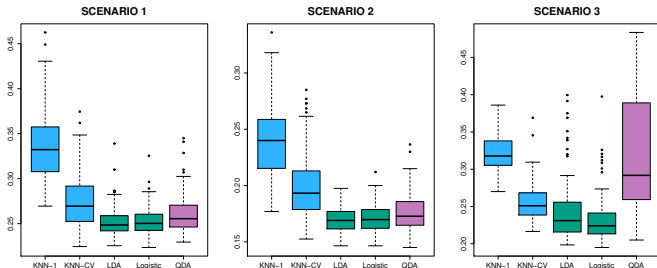
- ADL maximiza a **verossimilhança completa**

$$\prod_i p(x_i, y_i) = \underbrace{\prod_i p(x_i|y_i)}_{\text{normal } f_k} \underbrace{\prod_i p(y_i)}_{\text{bernoulli } \pi_k}$$

# Qual classificador escolher?



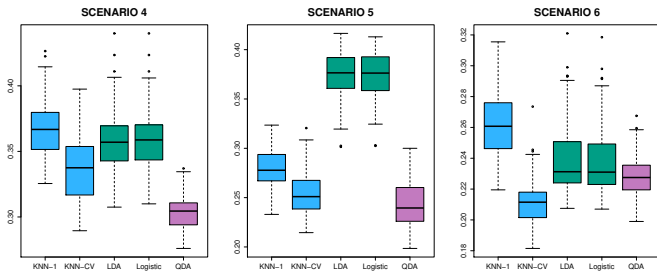
- **Cenário 1:** 20 observações em cada classe. Todas não correlacionadas e normalmente distribuídas;
- **Cenário 2:** Semelhante ao cenário 1, mas em cada classe, os preditores têm correlação de -0,5;
- **Cenário 3:** Semelhante ao cenário 1, mas com distribuição *t de student*.



# Qual classificador escolher?



- **Cenário 4:** Os dados são normalmente distribuídos, com correlação de 0,5 em uma classe e -0,5 em outra;
- **Cenário 5:** As respostas foram geradas utilizando os preditores:  $X_1^2$ ,  $X_2^2$  e  $X_1 \times X_2$  (ou seja, limite de decisão quadrático);
- **Cenário 6:** As respostas foram geradas utilizando funções não lineares mais elaboradas.



# Naive bayes

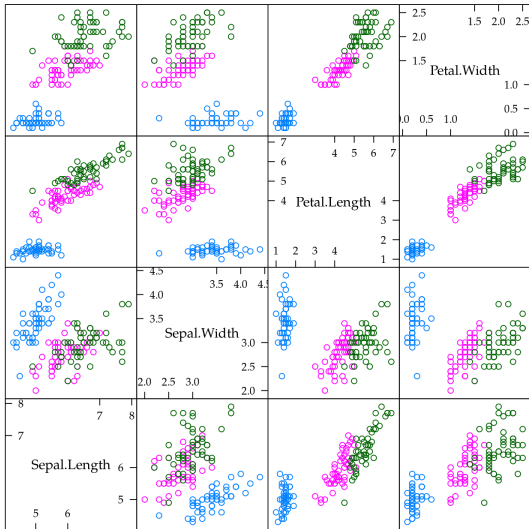


- Se supusermos que as componentes de  $x$  são independentes **condicionalmente à classe  $Y$**  estamos diante do **Naive Bayes**;
- Naive Bayes assume distribuição normal, com  $\Sigma_k$  diagonal:

$$\delta_k(x) \propto \log \left[ \pi_k \prod_{j=1}^p f_{kj}(x_j) \right] = -\frac{1}{2} \sum_{j=1}^p \frac{(x_j - \mu_{kj})^2}{\sigma_{kj}^2} + \log(\pi_k).$$

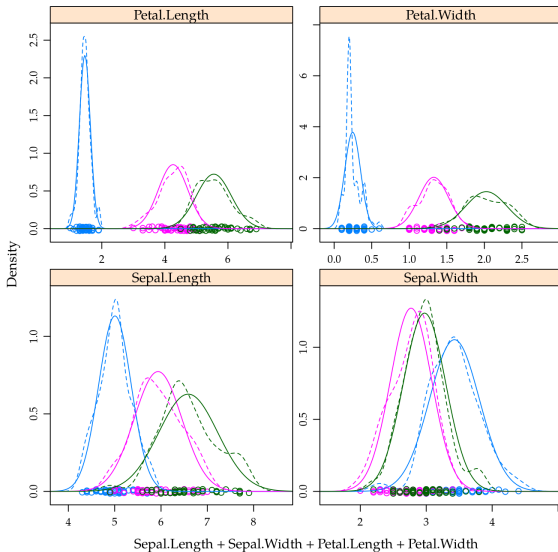
- Apesar de tal suposição não ser razoável, em muitos problemas ela é conveniente, e leva a bons classificadores.
- Lembre-se que estamos interessados classificar, e obter estimadores viciados não altera esta decisão.

# Exemplo: Iris Data



Scatter Plot Matrix

# Exemplo: Iris Data



# Tipos de erro

- Voltando ao exemplo do cartão de crédito, temos a seguinte situação:

		Default observado		
		Não	Sim	Total
Default predito	Não	9644	252	9896
	Sim	23	81	104
Total		9667	333	10000

- Tivemos  $\frac{23 + 252}{10000} = 2,75\%$  erros de classificação;
- Se classificarmos todos como **Não**, teríamos  $\frac{333}{10000} = 3,33\%$  de erro;

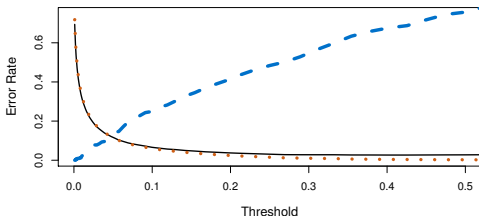
**Falso positivo:** fração de negativos classificados como positivo,  $\frac{23}{9667} = 0,2\%$ ;

**Falso negativo:** fração de positivos classificado como negativo,  $\frac{252}{333} = 75,7\%$ .

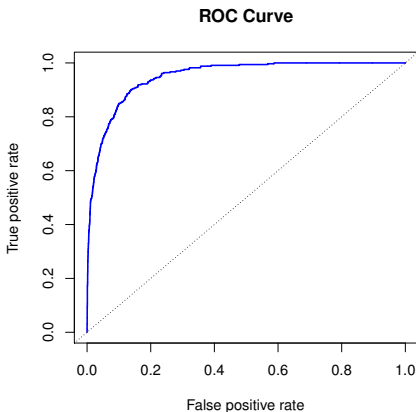
- Podemos mudar as taxas de erro, alterando a fronteira de decisão para algum valor  $\in [0, 1]$ :

$$\hat{P}(\text{Default} = \text{Yes} \mid \text{Balance}, \text{Student}) \geq \text{threshold.}$$

- Abaixo, em azul temos a taxa de falso negativo, em laranja falso positivo e em preto a taxa de erro total.



- A curva ROC (*receiver operator characteristic*) nos ajuda nesta escolha do *threshold*. Ela apresenta as duas taxas de erro ao mesmo tempo.



- James, G., Witten, D., Hastie, T. e Tibshirani, An Introduction to Statistical Learning, 2013;
- Hastie, T., Tibshirani, R. e Friedman, J., The Elements of Statistical Learning, 2009;
- Lantz, B., Machine Learning with R, Packt Publishing, 2013;
- Tan, Steinbach, and Kumar, Introduction to Data Mining, Addison-Wesley, 2005;
- Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani