Universidade Federal do Paraná

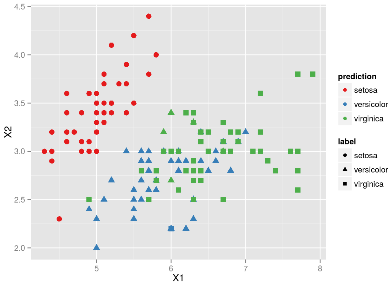
Laboratório de Estatística e Geoinformação - LEG

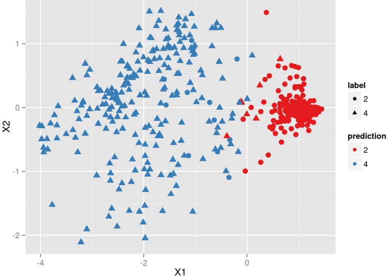
Classificação

Profs.: Eduardo Vargas Ferreira
Walmes Marques Zeviani

Introdução

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





Introdução

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

$$\mathrm{P}(Y = c_1|\boldsymbol{x}) \geq \mathrm{P}(Y = c_2|\boldsymbol{x}),$$

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

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

- Note que agora, o custo baseado na distância entre a resposta observada e estimada não faz mais sentido. Ao invés dele, é comum utilizar

$$J(h) = \mathrm{P}[Y \neq h(\boldsymbol{X})].$$

- Assim, ainda que $h(\boldsymbol{x}) \in \mathbb{R}^+$, ela representará a escolha por uma classe.

Plug-in classifier

- Entretanto, não conhecemos tais probabilidades:

O classificador de Bayes é um padrão ouro inalcançável!

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

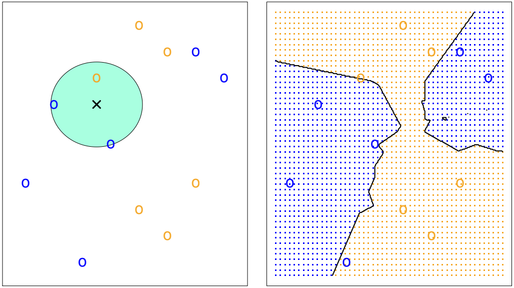
> ⋆ Estimamos $\mathrm{P}(Y = c|\boldsymbol{x})$ para cada categoria $c \in \mathcal{C}$;
>
> ⋆ Tomamos $h(\boldsymbol{x}) = \underset{c \in \mathcal{C}}{argmax}\ \widehat{\mathrm{P}}(Y = c|\boldsymbol{x})$.

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

K-Nearest Neighbors

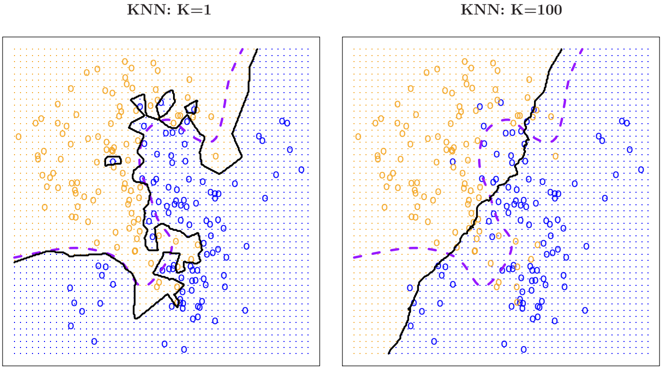
- O KNN estima a distribuição condicional de $Y|X$ de acordo com as classes dos K vizinhos de determinada observação x_0, ou seja:

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



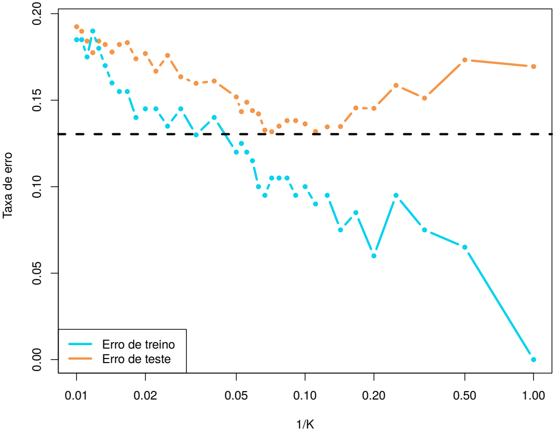
K-Nearest Neighbors

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



K-Nearest Neighbors

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



Exemplo: Inadimplência no cartão de crédito

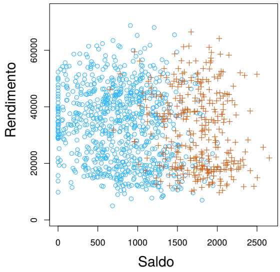
- Neste exemplo, nosso objetivo é prever se um cliente será ou não inadimplente no próximo mês;

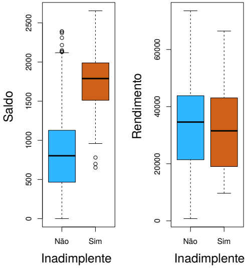
- Para tanto, temos três variáveis explicativas:

  ⋆ `Estudante`: se o cliente é ou não estudante;

  ⋆ `Rendimento`: rendimento anual do cliente;

  ⋆ `Saldo`: o valor devido no mês atual.





Podemos utilizar regressão linear?

- Suponha que para classificação da variável `Inadimplente` codificamos da forma:

$$Y = \begin{cases} 0, & \text{se } \texttt{Não} \ , \\ 1, & \text{se } \texttt{Sim} \ . \end{cases}$$

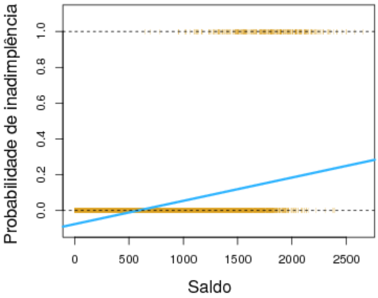
- Podemos simplesmente realizar uma regressão linear de Y em X e classificar como `Sim` se $\hat{Y} > 0.5$?

  - ⋆ Considerando o fato de que $\mathrm{E}\left(Y|\boldsymbol{X} = \boldsymbol{x}\right) = \mathrm{P}\left(Y = 1|\boldsymbol{X} = \boldsymbol{x}\right)$, podemos pensar que regressão é ótima para isto!

  - ⋆ No caso de resposta binária, regressão linear faz um bom trabalho (equivalente à **análise de discriminante linear**);

  - ⋆ Entretanto, ela pode produzir probabilidades menores do que 0 ou maiores do que 1. **Regressão logística** é mais apropriada.

Regressão logística



- Denotando por $p(\boldsymbol{X}) = P(Y = 1|\boldsymbol{X})$. A regressão logística utiliza a forma

$$p(\boldsymbol{X}) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

- Assim, não importa os valores de β_0 e β_1 ou $\boldsymbol{X}$, $p(\boldsymbol{X}) \in (0, 1)$.

Regressão logística

- Com um pouco de algebrismo chegamos em

$$log\left(\frac{p(\boldsymbol{X})}{1-p(\boldsymbol{X})}\right) = \beta_0 + \beta_1 X.$$

- Que é chamada **log odds** ou transformação **logit** em $p(\boldsymbol{X})$.

Variável	Coeficiente	Desvio padrão	Estatística t	p-valor
`Intercepto`	-10,6513	0,3612	-29,5	$< 0,0001$
`Saldo`	0,0055	0,0002	24,9	$< 0,0001$

- Qual é a probabilidade estimada de `Inadimplente` para um cliente com `Saldo` de \$1000?

$$\hat{p}(\boldsymbol{X}) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10{,}6513+0{,}0055\times 1000}}{1+e^{-10{,}6513+0{,}0055\times 1000}} = 0,006.$$

Regressão logística

- Vamos repetir o processo anterior, agora com `Estudante` como preditor;

Variável	Coeficiente	Desvio padrão	Estatística t	p-valor
`Intercepto`	-3,5041	0,0707	-49,55	$< 0,0001$
`Estudante[Sim]`	0,4049	0,1150	3,52	0,0004

$$\widehat{\mathrm{P}}\left(\texttt{Inadimplente=Sim}|\texttt{Estudante=Sim}\right) = \frac{e^{-3{,}5041+0{,}4049\times 1}}{1+e^{-3{,}5041+0{,}4049\times 1}} = 0,0431.$$

$$\widehat{\mathrm{P}}\left(\texttt{Inadimplente=Sim}|\texttt{Estudante=Não}\right) = \frac{e^{-3{,}5041+0{,}4049\times 0}}{1+e^{-3{,}5041+0{,}4049\times 0}} = 0,0292.$$

Regressão logística com várias variáveis

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

$$log\left(\frac{p(\boldsymbol{X})}{1-p(\boldsymbol{X})}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p.$$

e

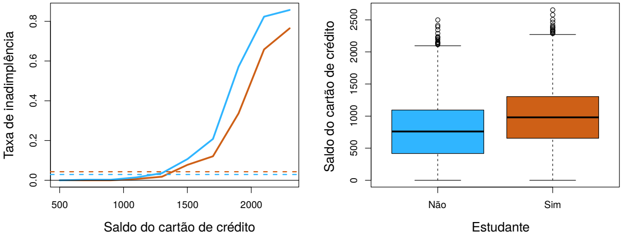
$$p(\boldsymbol{X}) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\beta_p X_p}}.$$

Variável	Coeficiente	Desvio padrão	Estatística t	p-valor
`Intercepto`	-10,8690	0,4923	-22,08	$< 0,0001$
`Saldo`	0,0057	0,0002	24,74	$< 0,0001$
`Rendimento`	0,0030	0,0082	0,37	0,7115
`Estudante[Sim]`	-0,6468	0,2362	-2,74	0,0062

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

Confundimento

- Os resultados são diferentes, especialmente quando existe correlação entre os preditores (veja o gráfico da direita);



- `Estudantes[Sim]` tendem a ter maior `Saldo` do cartão de crédito;
- Assim, marginalmente a taxa de `Inadimplência` é maior do que não `Estudantes[Não]`;
- Por outro lado, para cada nível do `Saldo` mensal, a inadimplência dos estudantes é menor (gráfico da esquerda).

Outra abordagem

- Uma alternativa para estimar $P(Y|X)$ consiste em modelar a distribuição de X em cada classe separadamente;

- E utilizar o **Teorema de Bayes** para obter $P(Y|X)$;

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

- 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

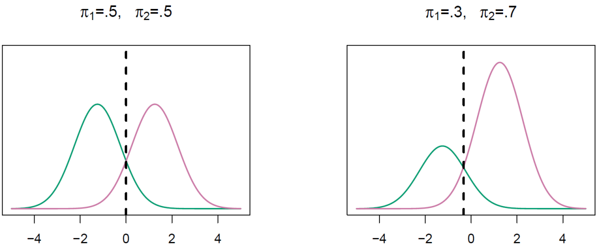
$$\delta_k(x) \propto argmax\ \pi_k f_k(x)$$

Outra abordagem

- $f_k(x) = P(\boldsymbol{X} = \boldsymbol{x} | Y = k)$ é a **densidade** para X na classe k (diferentes distribuições levam a diferentes métodos);

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

- Para diferentes prioris em cada classe, temos diferentes decisões;



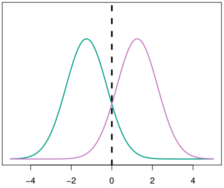
Análise de discriminante

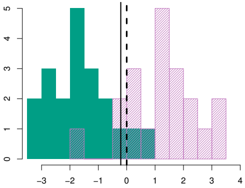
- Ao considerarmos para $f_k(x)$ a distribuição Normal em cada classe, nos leva à **análise de discriminante linear** ou **quadrática**, pois

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

- $\langle x-\mu_k, \Sigma_k^{-1}(x-\mu_k)\rangle$ é a **Distância de Mahalanobis** de x e μ_k;

- Por exemplo, seja $\mu_1 = -1.5, \mu_2 = 1.5, \pi_1 = \pi_2 = 0.5$ e $\sigma^2 = 1$





Análise de discriminante

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

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

- Note a ocorrência do termo quadrático na distância de Mahalanobis;

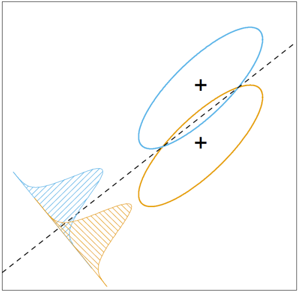
- Se todas as classes compartilharem o mesmo $\Sigma = \sum_k \frac{n_k-1}{n-K}\hat{\Sigma}_k$, estamos diante da **análise de discriminante linear (ADL)**

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

- Em ADL, o termo quadrático é cancelado.

Análise de discriminante

- Utilizamos, assim, os dados de treino para estimar tais quantidades e incorporar à regra de decisão, da seguinte forma



$$
\begin{aligned}
\hat{\pi}_k &= \frac{n_k}{n} \\
\hat{\mu}_k &= \frac{1}{n_k} \sum_{i:y_i=k} x_i \\
\hat{\Sigma}_k &= \frac{1}{n_k - 1} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^t
\end{aligned}
$$

Regressão logística versus ADL

- Regressão logística e análise de discriminante linear diferem-se na forma de estimar os parâmetros:

  - ⋆ Regressão logística maximiza a **verossimilhança condicional**

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

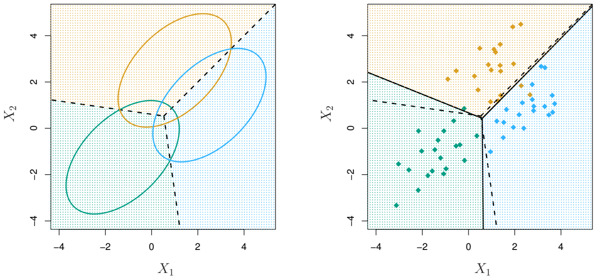
  - ⋆ ADL maximiza a **verossimilhança completa**

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

- Mas na prática, os resultados são similares.

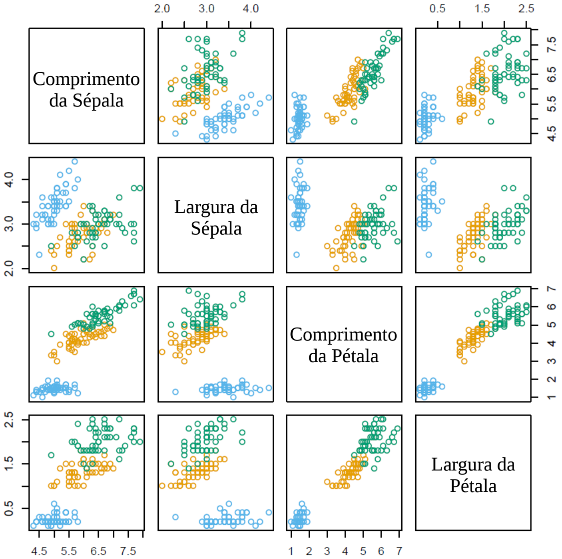
Ilustração: $p = 2$ e $k = 3$ classes

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



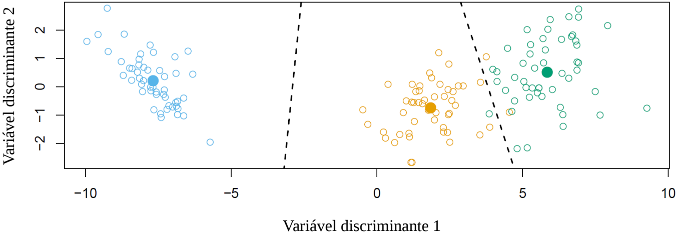
- A linha pontilhada é conhecida como **fronteira de decisão de Bayes** (*Bayes decision boundaries*);

Exemplo: Iris Data



Exemplo: `Iris Data`

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

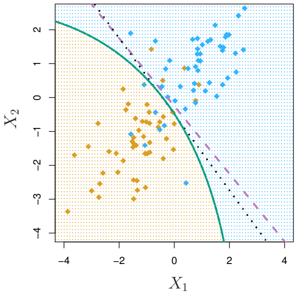


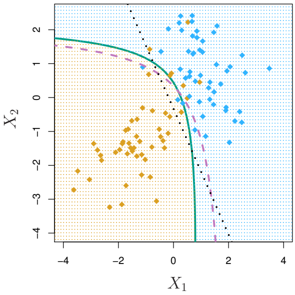
- Análise de discriminante linear classifica corretamente 147/150 observações dos dados de treino.

Exemplo simulado: Bayes, ADL e ADQ

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

- No gráfico da esquerda $\mathbf{\Sigma_1} = \mathbf{\Sigma_2}$ e o da direita $\mathbf{\Sigma_1} \neq \mathbf{\Sigma_2}$;



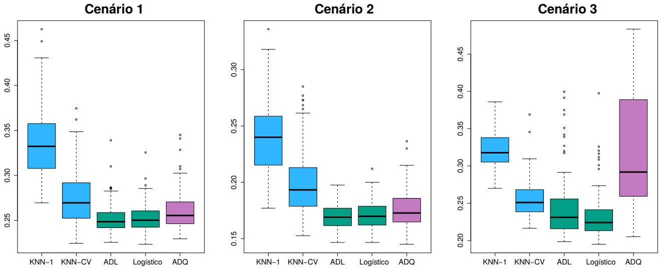


Qual classificador escolher?

- **Cenário 1:** 20 observações em cada classe. Todas não correlacionadas e normalmente distribuídas;

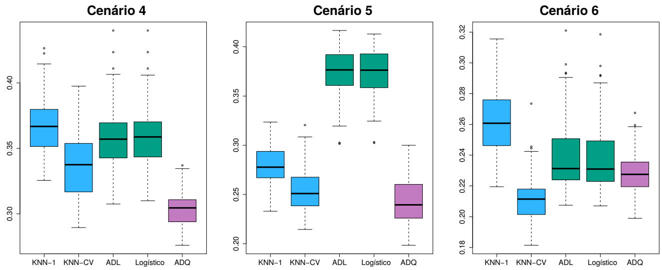
- **Cenário 2:** Semelhante ao cenário 1, mas em cada classe, os preditores têm correlação de -0,5;

- **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

- Vimos que quando $f_k(x)$ tem distribuição Normal com mesma variância $\boldsymbol{\Sigma}$ temos ADL. E se temos variâncias diferentes em cada classe temos ADQ;

- Agora, se supusermos que as componentes de $\boldsymbol{x}$ são independentes **condicionalmente à classe Y** estamos diante do **Naive Bayes**;

- Naive Bayes assume distribuição normal, com $\boldsymbol{\Sigma}_k$ 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 (Naive = Ingênuo) ela é conveniente, e leva a bons classificadores.

Tipos de erro

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

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

Falso positivo: fração de exemplos negativos classificados como positivo;

Falso negativo: fração de exemplo positivo classificado como negativo;

- Construímos esta tabela classificando a classe como `Sim` se

$$\widehat{\mathrm{P}}(\texttt{Inadimplência} = \texttt{Sim} \mid \texttt{Saldo, Estudante}) \geq 0,5.$$

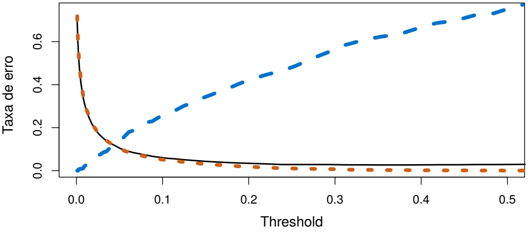
- Será que o limiar de 0,5 é a melhor opção?

Variando o threshold

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

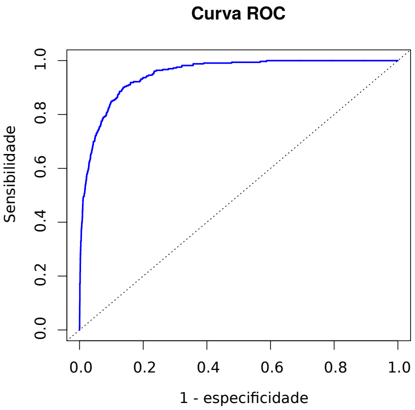
$$\widehat{\mathrm{P}}(\texttt{Inadimplência} = \texttt{Sim} \mid \texttt{Saldo, Estudante}) \geq \text{threshold}.$$

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



Curva ROC

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



Referências

- 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