

# Fundamentos de Machine Learning

Paulo C. Marques F.

Insper

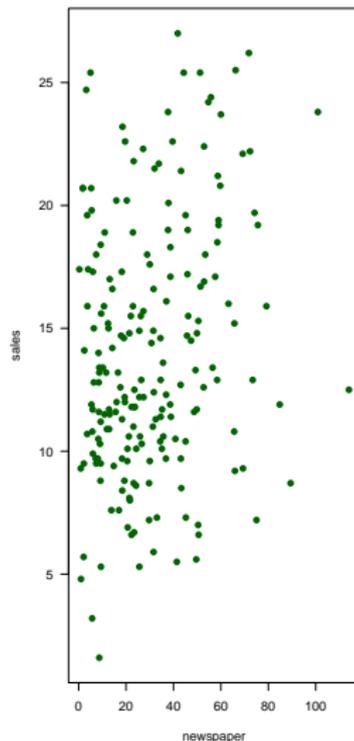
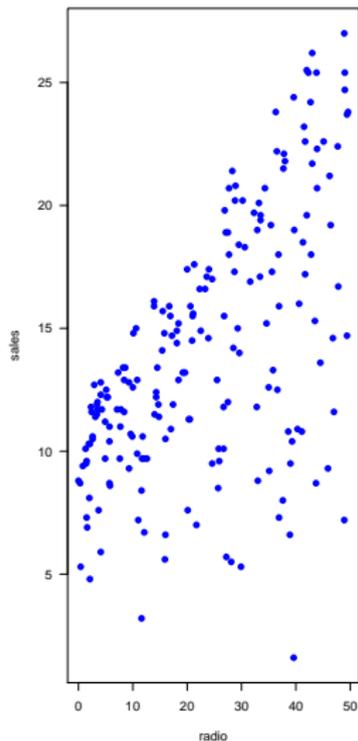
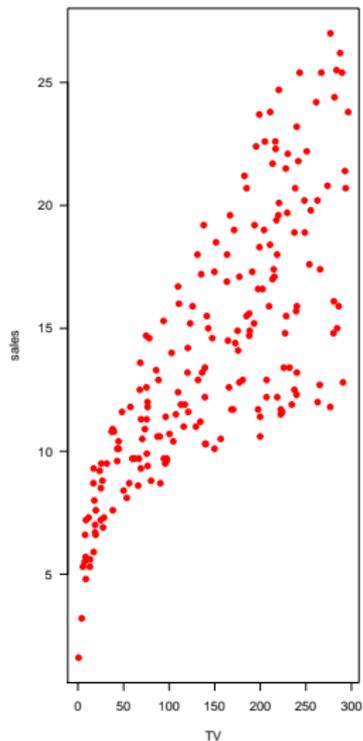
Para um determinado fenômeno, temos um vetor de  $p \geq 1$  *variáveis preditoras*  $X = (X_1, \dots, X_p) \in \mathbb{R}^p$  e uma *variável resposta*  $Y$ .

Informalmente, o objetivo da *aprendizagem supervisionada* é entender as relações entre  $X$  e  $Y$ .

Exemplo: no conjunto de dados **Advertising** temos informações sobre as vendas de um produto em 200 mercados.

Em cada mercado, temos três variáveis preditoras, **TV**, **radio** e **newspaper**, que determinam quais os gastos em propaganda, em milhares de dólares, feitos pelo fabricante do produto em cada uma destas mídias.

A variável resposta **sales** consiste no número de vendas do produto, medido em milhares de unidades.



Na literatura, as variáveis preditoras também são denominadas *entradas*, *variáveis independentes* (péssimo nome), *variáveis explicativas* e *características* (*features*).

A variável resposta também é denominada *saída*, *variável dependente* e *variável explicada*.

Nos problemas de *aprendizagem não supervisionada* não temos uma variável resposta.

Dois problemas não supervisionados exemplares são a análise de conglomerados (*clusters*) e os procedimentos de redução dimensional, tais como a Análise de Componentes Principais (*Principal Components Analysis* ou PCA).

Por hora, trataremos dos problemas de aprendizagem supervisionada.

Em um problema de *regressão*, a variável resposta é quantitativa:  
 $Y \in \mathbb{R}$ .

Em um problema de *classificação*<sup>1</sup>, a variável resposta é qualitativa e pertence a um conjunto finito de *rótulos* ou *classes*:  $Y \in \{0, 1, \dots, c\}$ .

A diferença é aparentemente inócua, mas veremos que há resultados teóricos específicos para cada tipo de problema.

Começaremos examinando os fundamentos dos problemas de regressão.

---

<sup>1</sup>Na literatura há um pequeno “bug” terminológico, pois *regressão logística* se refere a um método de classificação.

Suponha que o par  $(X, Y)$  tenha distribuição conjunta  $F_{X,Y}$ , que não conhecemos, pois estamos fazendo inferência.

## Problema

Queremos uma função  $\psi : \mathbb{R}^p \rightarrow \mathbb{R}$  que minimize o *erro de predição esperado*  $E[(Y - \psi(X))^2]$ .

## Observações:

1. Em Aprendizagem Estatística, o erro de predição esperado também é denominado *erro de teste esperado*, ou apenas *erro de teste*. Utilizaremos estes termos sem diferenciação.
2. Para aqueles familiarizados com a Teoria da Decisão, estamos trabalhando com uma função de perda quadrática. É possível refazer as análises a seguir com outras funções de perda, obtendo conclusões similares, porém os cálculos são muito mais difíceis.

O erro de predição esperado

$$E[(Y - \psi(X))^2] = \int (y - \psi(x))^2 dF_{X,Y}(x, y)$$

pode ser reescrito utilizando as propriedades da esperança condicional:

$$\begin{aligned} E[(Y - \psi(X))^2] &= E[E[(Y - \psi(X))^2 | X]] \\ &= \int E[(Y - \psi(x))^2 | X = x] dF_X(x). \end{aligned}$$

Portanto, para resolvermos o problema, basta encontrar  $\psi$  que minimize o integrando  $E[(Y - \psi(x))^2 | X = x]$ .

Para obter uma solução do problema, defina  $t = \psi(x)$  e observe que

$$\begin{aligned} E[(Y - \psi(x))^2 | X = x] &= E[(Y - t)^2 | X = x] \\ &= E[Y^2 | X = x] - 2t E[Y | X = x] + t^2 =: g(t). \end{aligned}$$

Zerando a derivada de  $g$  em relação a  $t$  obtemos

$$g'(t) = -2E[Y | X = x] + 2t = 0$$

se e somente se  $t = E[Y | X = x]$ . Note também que  $g''(t) = 2 > 0$ .

Portanto, a função  $\psi$  que minimiza o erro de predição esperado é a *função de regressão*  $\psi(x) = E[Y | X = x]$ .

A solução obtida é formal, no sentido de que, uma vez que não conhecemos a distribuição conjunta  $F_{X,Y}$ , também não conhecemos a esperança condicional  $E[Y | X = x]$ .

Suponha que temos *dados de treinamento*  $(X_1, Y_1), \dots, (X_n, Y_n)$  tais que os pares  $(X_i, Y_i)$  são independentes e identicamente distribuídos com distribuição  $F_{X,Y}$ .

A idéia central da aprendizagem estatística supervisionada é utilizar a informação contida nestes dados de treinamento para construir uma função  $\hat{\psi} : \mathbb{R}^p \rightarrow \mathbb{R}$  que será nossa estimativa (nosso “chute”) para  $\psi$ .

Formalmente,  $\hat{\psi} = \mathcal{L}[(X_1, Y_1), \dots, (X_n, Y_n)]$ , em que  $\mathcal{L}$  é um funcional “arbitrariamente complicado”. Cada funcional  $\mathcal{L}$  corresponde a um método específico de aprendizagem.

Há dois propósitos, em geral antagônicos, quando se trata de aprender sobre  $\psi$ .

No contexto de *predição*, procuramos um método de aprendizagem que faça boas previsões no seguinte sentido.

Para um novo vetor de variáveis preditoras  $X$ , tratamos  $\hat{\psi}$  como uma “caixa-preta” e queremos  $\hat{\psi}$  que, para a variável resposta  $Y$  correspondente, produza um erro de predição esperado  $E[(Y - \hat{\psi}(X))^2]$  pequeno.

No contexto de *inferência*, queremos um método de aprendizagem que produza soluções interpretáveis.

Em inferência, a prioridade é entender a relação entre a variável resposta e cada uma das variáveis preditoras. Neste caso, não podemos tratar  $\hat{\psi}$  como uma “caixa-preta”: a forma funcional de  $\hat{\psi}$  precisa ser interpretável.

No contexto de inferência, as perguntas típicas são:

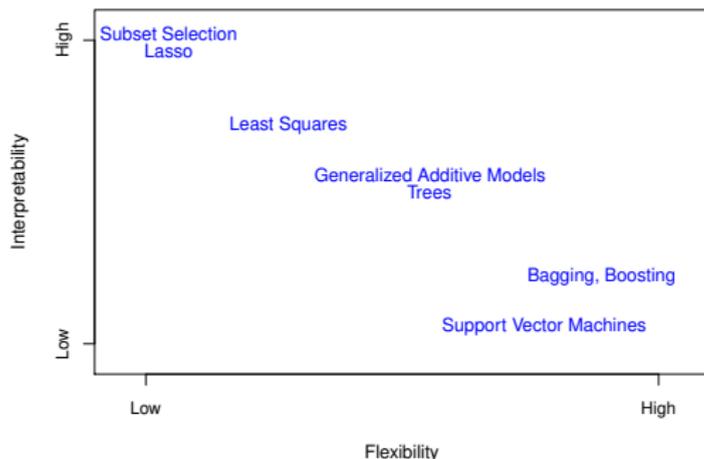
- ▶ Quais preditoras são associadas com a resposta?
- ▶ Qual a relação entre a resposta e *cada uma* das preditoras?

No exemplo dos dados **Advertising**, gostaríamos de responder perguntas inferenciais do tipo:

- ▶ Há uma associação entre **TV** e **sales**?
- ▶ Um aumento de 25% em **newspaper** levaria em média a que mudança em **sales**?

Nos modelos de aprendizagem estatística observa-se um perde-ganha (*trade-off*) entre predição e inferência.

Em geral, modelos mais flexíveis, que preveem bem, não são afeitos à interpretação (funcionam como “caixas-pretas”), enquanto modelos menos flexíveis e mais interpretáveis têm capacidade preditiva limitada.



Suponha que temos o modelo aditivo  $Y = \psi(X) + \epsilon$ , em que o vetor de preditoras  $X$  tem distribuição  $F_X$  e o *erro aleatório*  $\epsilon$  é uma variável aleatória, independente de  $X$ , cuja distribuição é tal que  $E[\epsilon] = 0$  e  $\text{Var}[\epsilon] = \sigma^2$  (desconhecida).

Note que este modelo induz uma distribuição conjunta  $F_{X,Y}$  para o par  $(X, Y)$  e que  $E[Y | X = x] = \psi(x)$ .

Neste modelo,  $\psi$  representa a *informação sistemática* do vetor de preditoras  $X$  sobre a variável resposta  $Y$ .

Intuitivamente, o erro aleatório representa fatores que afetam a resposta  $Y$ , mas não estão relacionados aos valores das preditoras que compõem  $X$ .

Ou seja, é como se tivéssemos outras variáveis preditoras que afetam o valor da resposta, mas que não são observadas, e o erro aleatório quantifica a incerteza devida a esta informação incompleta.

Incorporando os dados de treinamento  $(X_1, Y_1), \dots, (X_n, Y_n)$  e novos erros aleatórios  $\epsilon_1, \dots, \epsilon_n$ , o modelo aditivo assim completado fica descrito como:

- ▶  $Y = \psi(X) + \epsilon$ .
- ▶  $Y_i = \psi(X_i) + \epsilon_i$ .
- ▶  $X$  e os  $X_i$ 's têm a mesma distribuição  $F_X$ .
- ▶  $\epsilon$  e os  $\epsilon_i$ 's têm a mesma distribuição, com esperança 0 e variância  $\sigma^2$  (desconhecida).
- ▶  $X$ , os  $X_i$ 's,  $\epsilon$  e os  $\epsilon_i$ 's são todos independentes.

Notacionalmente, iremos interpretar  $X$  como um novo vetor de preditoras a partir do qual queremos prever a resposta  $Y$  utilizando a estimativa  $\hat{\psi}$  obtida a partir dos dados de treinamento  $(X_1, Y_1), \dots, (X_n, Y_n)$ .

Ou seja, nossa predição para a resposta  $Y$  será  $\hat{Y} = \hat{\psi}(X)$ .

Este modelo aditivo nos permite decompor o erro de predição esperado  $E[(Y - \hat{\psi}(X))^2]$  em fatores interpretáveis.

Pelas propriedades da esperança condicional, o erro de predição esperado de uma  $\hat{\psi}$  pode ser expresso como

$$\begin{aligned} E[(Y - \hat{\psi}(X))^2] &= E[E[(Y - \hat{\psi}(X))^2 \mid X]] \\ &= \int E[(Y - \hat{\psi}(x))^2 \mid X = x] dF_X(x). \end{aligned}$$

Portanto, basta examinar a decomposição do erro de predição condicionado em  $X = x$ .

## Resultado

$$E[(Y - \hat{\psi}(x))^2 \mid X = x] = \underbrace{\text{EQM}[\hat{\psi}(x)]}_{\text{redutível}} + \underbrace{\sigma^2}_{\text{irreduzível}},$$

em que  $\text{EQM}[\hat{\psi}(x)] = E[(\hat{\psi}(x) - \psi(x))^2]$  é o *erro quadrático médio* condicionado em  $X = x$ .

A interpretação deste resultado é que podemos reduzir  $\text{EQM}[\hat{\psi}(x)]$  escolhendo o método de aprendizagem que produz  $\hat{\psi}$ , mas não controlamos o tamanho de  $\sigma^2$ .

### Demonstração

$$\begin{aligned} \text{E}[(Y - \hat{\psi}(x))^2 \mid X = x] &= \text{E}[(\psi(x) + \epsilon - \hat{\psi}(x))^2 \mid X = x] \\ &= \text{E}[(\hat{\psi}(x) - \psi(x))^2] + 2 \text{E}[(\hat{\psi}(x) - \psi(x)) \cdot \epsilon] + \text{E}[\epsilon^2] \\ &= \text{E}[(\hat{\psi}(x) - \psi(x))^2] + 2 (\text{E}[\hat{\psi}(x)] - \psi(x)) \cdot \underbrace{\text{E}[\epsilon]}_{=0} + \sigma^2 \\ &= \text{E}[(\hat{\psi}(x) - \psi(x))^2] + \sigma^2 \\ &= \text{EQM}[\hat{\psi}(x)] + \sigma^2. \end{aligned}$$

Portanto, fatoramos o erro de predição esperado como

$$\text{E}[(Y - \hat{\psi}(X))^2] = \int \text{EQM}[\hat{\psi}(x)] dF_X(x) + \sigma^2.$$

A parcela redutível do erro de predição admite uma decomposição que possui um papel central em Aprendizagem Estatística.

## Resultado

$$\text{EQM}[\hat{\psi}(x)] = \text{Viés}^2[\hat{\psi}(x)] + \text{Var}[\hat{\psi}(x)],$$

em que o *viés* e a *variância* condicionados em  $X = x$  são dados por

$$\text{Viés}[\hat{\psi}(x)] = \text{E}[\hat{\psi}(x)] - \psi(x) \quad \text{e} \quad \text{Var}[\hat{\psi}(x)] = \text{E}[(\hat{\psi}(x) - \text{E}[\hat{\psi}(x)])^2].$$

## Demonstração

$$\begin{aligned}\text{EQM}[\hat{\psi}(x)] &= \mathbb{E}[(\hat{\psi}(x) - \psi(x))^2] \\ &= \mathbb{E}[(\hat{\psi}(x) - \mathbb{E}[\hat{\psi}(x)]) - (\psi(x) - \mathbb{E}[\hat{\psi}(x)])]^2] \\ &= \mathbb{E}[(\hat{\psi}(x) - \mathbb{E}[\hat{\psi}(x)])^2] \\ &\quad - 2(\mathbb{E}[\hat{\psi}(x)] - \mathbb{E}[\hat{\psi}(x)]) \cdot (\psi(x) - \mathbb{E}[\hat{\psi}(x)]) \\ &\quad + (\mathbb{E}[\hat{\psi}(x)] - \psi(x))^2 \\ &= \text{Viés}^2[\hat{\psi}(x)] + \text{Var}[\hat{\psi}(x)].\end{aligned}$$

Portanto, a fatoração completa do erro de predição esperado é dada por

$$\mathbb{E}[(Y - \hat{\psi}(X))^2] = \int \text{Viés}^2[\hat{\psi}(x)] dF_X(x) + \int \text{Var}[\hat{\psi}(x)] dF_X(x) + \sigma^2,$$

em que o primeiro e o segundo termo da fatoração são respectivamente o *viés ao quadrado esperado* e a *variância esperada* do método de aprendizagem.

Informalmente, a variância de um método de aprendizagem estatística mede o quanto a estimativa  $\hat{\psi}$  muda conforme retreinamos o modelo com novos dados de treinamento.

Para um método com grande variância, uma pequena mudança nos dados de treinamento leva a uma grande mudança em  $\hat{\psi}$  (e vice-versa).

O viés de um método de aprendizagem mede o quanto  $\hat{\psi}$  difere em média de  $\psi$  sob replicação do processo de aprendizagem.

Por exemplo, ao usarmos um modelo linear para aprender sobre uma  $\psi$  fortemente não linear teremos um grande viés.

Em geral, métodos mais flexíveis apresentam um menor viés, mas, por outro lado, tais métodos possuem uma variância maior.

Empiricamente, nota-se em geral que nos métodos de aprendizagem estatística há um *perde-ganha viés-variância* (*bias-variance trade-off*): é fácil diminuir o viés do aprendizado aumentando a sua variância (e vice-versa).

A melhor maneira de concretizar os conceitos que discutimos até agora é examinar um processo de aprendizagem específico utilizando dados simulados.

No mundo artificial da simulação saberemos coisas que não saberíamos ao analisar dados reais: conheceremos a função de regressão  $\psi$  e as distribuições das preditoras e dos erros aleatórios.

Nesta simulação iremos associar uma única variável preditora  $X$ , que determina o número de anos de estudo de um indivíduo, com a variável resposta  $Y$ , definida pela renda anual do indivíduo, medida em milhares de reais.

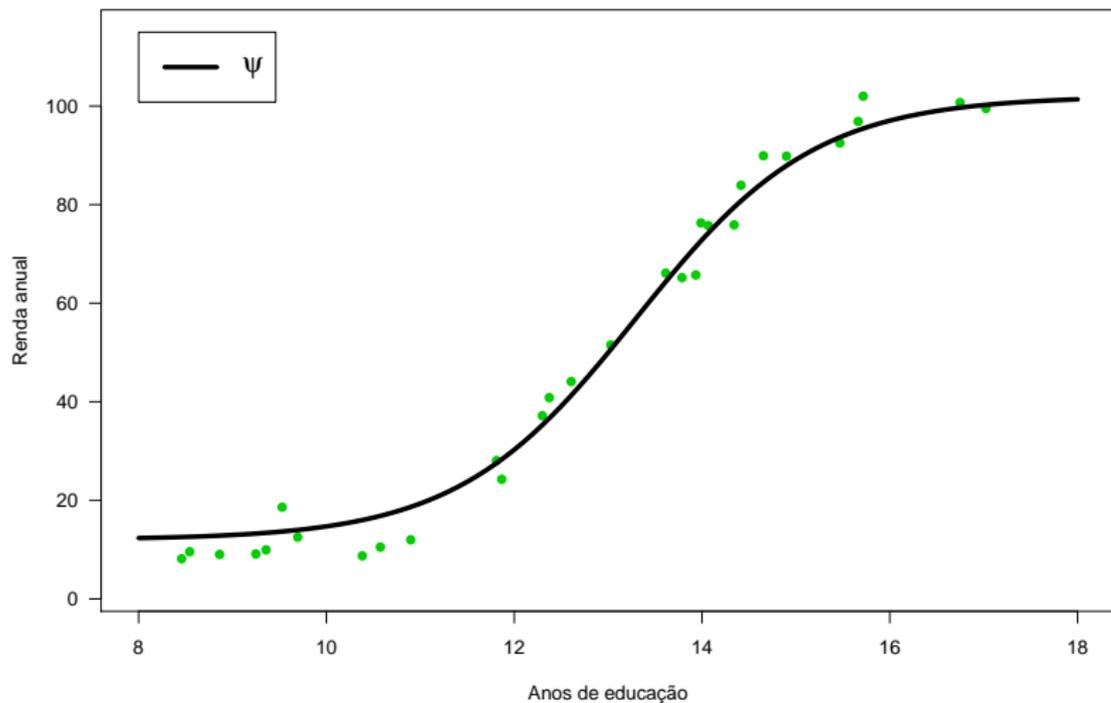
A informação sistemática de  $X$  sobre  $Y$  será dada pela função de regressão não linear definida por

$$\psi(x) = 45 \cdot \tanh\left(\frac{x}{1,9} - 7\right) + 57.$$

Para a preditora  $X$  utilizaremos uma distribuição uniforme entre 8 e 18 anos de estudo.

O erro aleatório terá distribuição normal com esperança 0 e desvio padrão igual a 4.

Na figura a seguir temos o gráfico da função de regressão  $\psi$  e de 30 pares  $(x_i, y_i)$  de dados de treinamento gerados pelo modelo aditivo especificado.



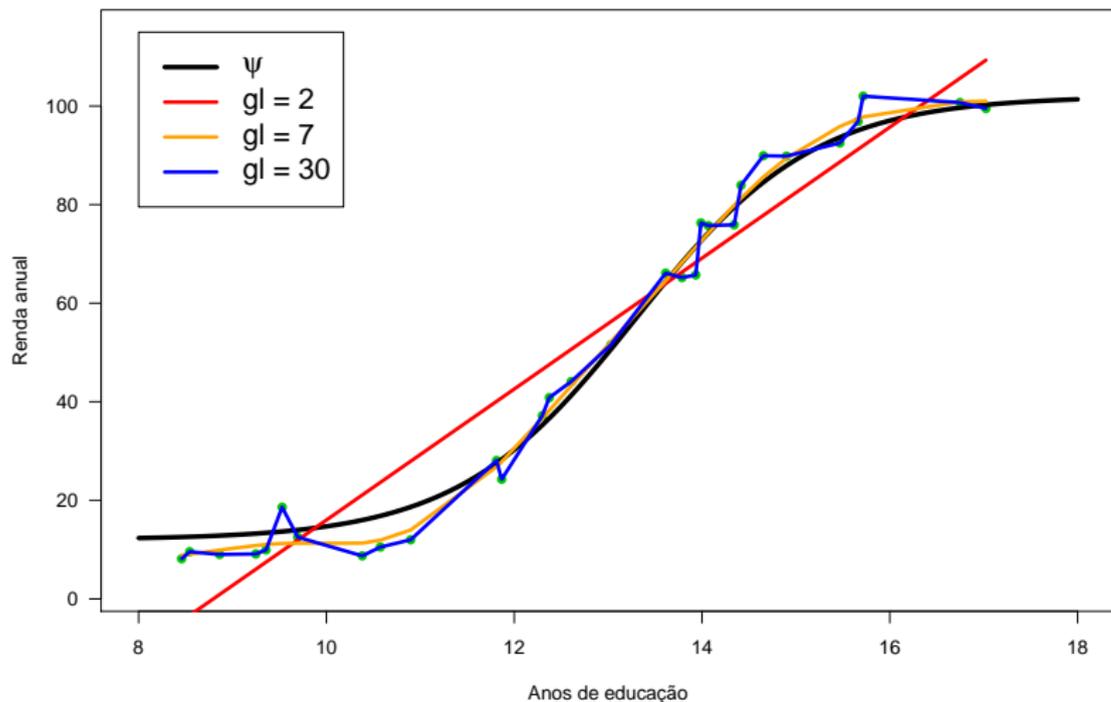
Nesta simulação, o método de aprendizagem será uma *smoothing spline*, implementada pela função `smooth.spline()` do R.

Propositalmente, não estamos interessados nos detalhes internos deste modelo. Vamos tratá-lo como uma “caixa-preta”. Só precisamos saber que neste modelo há um parâmetro denominado *graus de liberdade* (`gl`) que regula a flexibilidade do aprendizado.

Quando temos `gl = 2`, ao treinar o modelo obtemos uma  $\hat{\psi}$  que é a reta de mínimos quadrados ordinários.

Conforme crescemos o número de graus de liberdade, por exemplo utilizando `gl = 7`, o modelo fornece curvas  $\hat{\psi}$  mais adaptadas aos dados.

No extremo, com `gl = 30` graus de liberdade, a  $\hat{\psi}$  fornecida pelo modelo é uma linha poligonal que interpola perfeitamente os dados de treinamento.



Para estes modelos, podemos obter aproximações de Monte Carlo para os fatores do erro de predição esperado utilizando o seguinte algoritmo.

Para  $g_l = 2, \dots, n$ , repetimos o seguinte procedimento.

Para  $i = 1, \dots, M$ , geramos  $x^{(i)}$  de  $F_X$  e repetimos o seguinte procedimento.

Para  $j = 1, \dots, N$ , geramos  $(x_1^{(j)}, y_1^{(j)}), \dots, (x_n^{(j)}, y_n^{(j)})$  e treinamos  $\hat{\psi}_j^{(g_l)}$ .

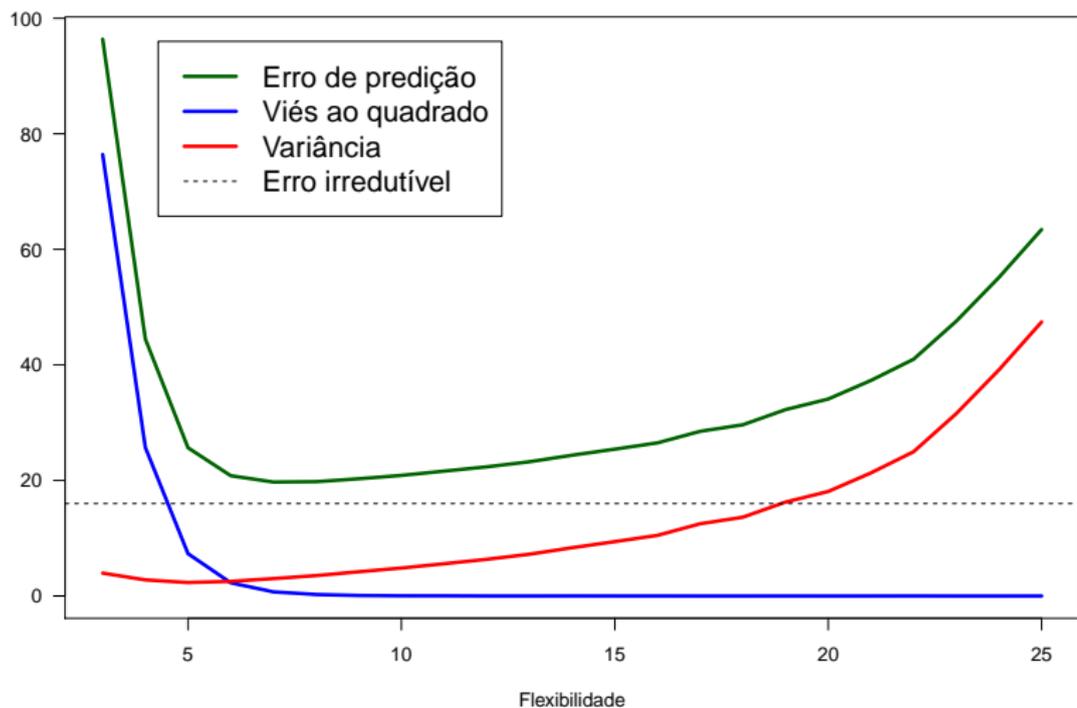
Aproximamos  $\text{Viés}^2[\hat{\psi}^{(g_l)}(x^{(i)})]$  por  $\left( \left( \frac{1}{N} \sum_{j=1}^N \hat{\psi}_j^{(g_l)}(x^{(i)}) \right) - \psi(x^{(i)}) \right)^2$ .

Aproximamos  $\text{Var}[\hat{\psi}^{(g_l)}(x^{(i)})]$  pela variância dos  $N$  valores dos  $\hat{\psi}_j^{(g_l)}(x^{(i)})$ 's.

Aproximamos o viés ao quadrado esperado por  $\frac{1}{M} \sum_{i=1}^M \text{Viés}^2[\hat{\psi}^{(g_l)}(x^{(i)})]$ .

Aproximamos a variância esperada por  $\frac{1}{M} \sum_{i=1}^M \text{Var}[\hat{\psi}^{(g_l)}(x^{(i)})]$ .

# Simulação do perde-ganha viés-variância (1)



A simulação mostra que o viés do aprendizado cai conforme aumentamos sua flexibilidade utilizando valores maiores para o número de graus de liberdade, ao mesmo tempo em que a variância do aprendizado cresce.

O uso de um modelo flexível demais leva ao *sobreajuste* (*overfitting*) dos dados de treinamento: tal modelo “ajusta” perfeitamente os dados de treinamento, mas possui um erro de predição elevado.

O equilíbrio entre viés e variância nos levaria a escolher, neste caso, aproximadamente 7 graus de liberdade.

Vemos que aprender sobre o erro de predição tem um duplo propósito:

1. Escolher, dentro de uma classe de modelos alternativos, aquele que melhor equilibra viés e variância (*model selection* ou *seleção do modelo*).
2. Medir a performance preditiva do modelo escolhido (*model assesment* ou *apreciação do modelo*).

Ao trabalhar com dados reais, como poderíamos estimar o erro de predição esperado?

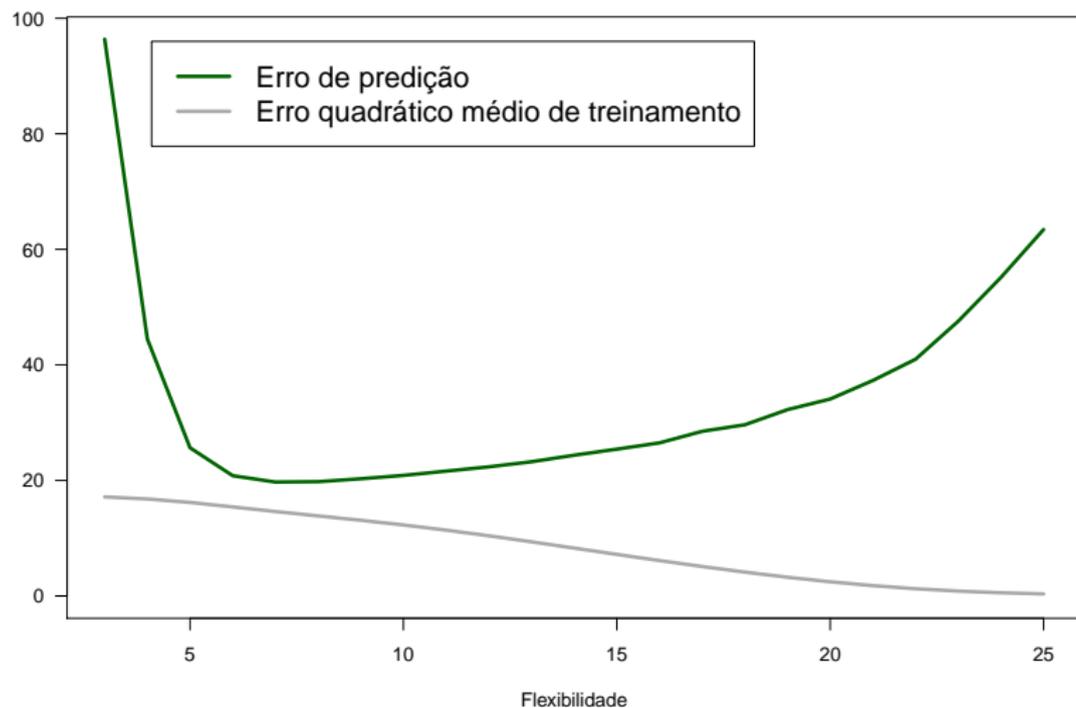
Para uma  $\hat{\psi}$  treinada com dados  $(x_1, y_1) \dots, (x_n, y_n)$ , defina o *erro quadrático médio de treinamento* por

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{\psi}(x_i))^2.$$

Seria possível estimar o erro de predição esperado utilizando o erro quadrático médio de treinamento?

A figura a seguir compara o erro quadrático médio de treinamento e o erro de predição esperado da nossa simulação.

## Erro quadrático médio de treinamento (2)



Vemos que não é possível estimar o erro de predição esperado utilizando diretamente o erro quadrático médio de treinamento

Conforme aumentamos o número de graus de liberdade, o erro quadrático médio de treinamento cai monotonicamente.

De fato, para o modelo com 30 graus de liberdade, que interpola perfeitamente os dados de treinamento, o erro quadrático médio de treinamento é igual a zero.

Portanto, se usássemos o erro quadrático médio de treinamento para escolher o número de graus de liberdade do modelo, sempre selecionaríamos um modelo sobreajustado.

As técnicas de *validação cruzada* são procedimentos de reamostragem utilizados para estimar o erro de predição esperado de um método de aprendizagem.

Suponha que temos em mãos algum método de aprendizagem supervisionada que fornece, a partir de dados de treinamento  $(x_1, y_1), \dots, (x_n, y_n)$ , uma estimativa  $\hat{\psi}$  para a função de regressão  $\psi$ .

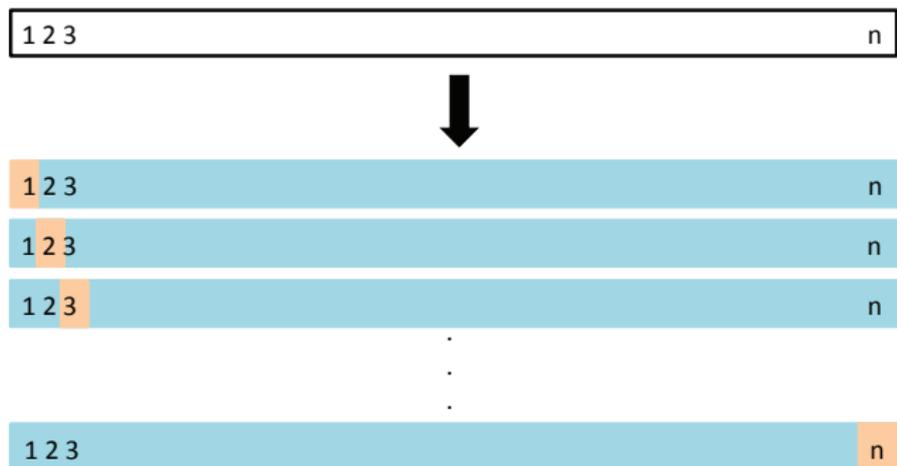
A *validação cruzada tirando-um-fora* (*leave-one-out cross validation*, que abreviaremos por LOOCV) permite estimar o erro de predição esperado do método de aprendizado em questão da seguinte maneira.

Para  $i = 1, \dots, n$ , remova o par  $(x_i, y_i)$  dos dados de treinamento e calcule a estimativa  $\hat{\psi}_{(\setminus i)}$  com os  $n - 1$  dados de treinamento remanescentes, obtendo  $\text{eqm}_i = (y_i - \hat{\psi}_{(\setminus i)}(x_i))^2$ . A estimativa LOOCV para o erro de predição esperado é

$$\text{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^n \text{eqm}_i.$$

## Validação cruzada tirando-um-fora (2)

A figura abaixo ilustra o procedimento LOOCV.



A *validação cruzada em  $k$  lotes* ( *$k$ -fold cross validation*) permite estimar o erro de predição esperado de um método de aprendizado da seguinte maneira.

Quebre aleatoriamente os dados de treinamento em  $k$  lotes com aproximadamente o mesmo tamanho.

Para  $i = 1, \dots, k$ , remova o  $i$ -ésimo lote dos dados de treinamento e calcule a estimativa  $\hat{\psi}_{(\setminus i)}$  a partir dos dados de treinamento remanescentes, obtendo

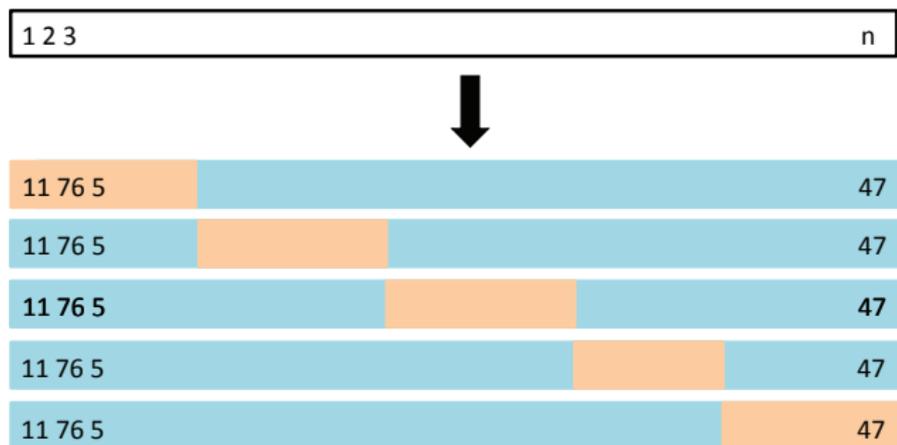
$$\text{eqm}_i = \frac{1}{(\text{tamanho do } i\text{-ésimo lote})} \sum_{(x_j, y_j) \in i\text{-ésimo lote}} (y_j - \hat{\psi}_{(\setminus j)}(x_j))^2.$$

A estimativa da validação cruzada em  $k$  lotes para o erro de predição esperado é

$$\text{CV}_{(k)} = \frac{1}{k} \sum_{i=1}^k \text{eqm}_i.$$

## Validação cruzada em $k$ lotes (2)

A figura abaixo ilustra o procedimento de validação cruzada em  $k$  lotes.



Note que LOOCV torna-se um caso particular da validação cruzada em  $k$  lotes se fizermos  $k = n$ .

LOOCV é um procedimento com maior custo computacional, uma vez que estamos treinando o modelo  $n$  vezes. No entanto, a validação cruzada em  $k$  lotes tem outra vantagem, menos óbvia.

No procedimento LOOCV tiramos a média dos  $eqm_i$ 's obtidos pelo ajuste de conjuntos de dados muito similares, o que induz uma forte correlação positiva entre tais  $eqm_i$ 's. Uma vez que

$$\text{Var}[\text{CV}_{(n)}] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}[\text{EQM}_i] + \frac{2}{n^2} \sum_{1 \leq i < j \leq n} \text{Cov}[\text{EQM}_i, \text{EQM}_j],$$

o procedimento LOOCV, em geral, apresenta uma variância maior do que a validação cruzada em  $k$  lotes. No fundo, o que está em jogo é mais uma encarnação do perde-ganha viés-variância.

A **prática** indica que as escolhas  $k = 5$  e  $k = 10$  são um bom compromisso entre viés, variância e custo computacional.

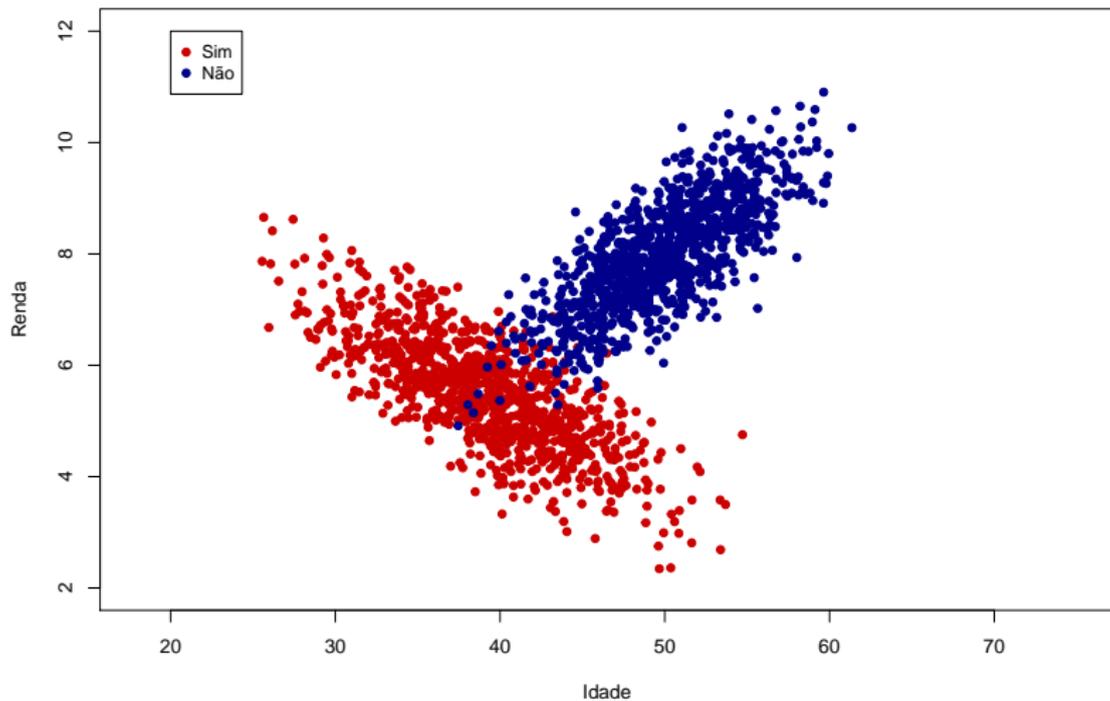
Imagine que temos dados de 2000 clientes de um banco. Para cada cliente, observamos duas *variáveis preditoras*: **Idade** e **Renda** média mensal, medida em salários mínimos (em um problema mais realista teríamos muito mais variáveis preditoras disponíveis).

Os 2000 clientes estão rotulados em duas classes, **Sim** ou **Não**, indicando se usaram ou não o limite do cheque especial por mais de 21 dias no último ano. Esta é nossa *variável reposta*.

Um *classificador* é uma regra que, para um novo cliente que acaba de abrir conta no banco, permite classificá-lo em uma das classes, **Sim** ou **Não**, a partir dos valores de suas preditoras **Idade** e **Renda**.

A utilidade prática de um classificador é que ele permite prever o comportamento de um futuro cliente. Baseado nesta predição, o banco pode tomar diversas decisões especificamente para aquele cliente, tais como valores das tarifas dos serviços etc.

Utilizou o cheque especial por mais de 21 dias no último ano?



No caso geral, temos um vetor de *variáveis preditoras*  $X \in \mathbb{R}^p$ , uma *variável resposta*  $Y \in C = \{0, \dots, c\}$ , e queremos construir um *classificador*  $\varphi : \mathbb{R}^p \rightarrow C$ .

Sem perda de generalidade, vamos supor que temos apenas duas classes:  $C = \{0, 1\}$  (correspondentes às classes **Sim** e **Não** dos dados bancários).

Os problemas em que há um classificador  $\varphi$  que nunca erra são de pouco interesse científico.

No banco, sempre haverá clientes com a mesma idade e renda, mas que possuem padrões distintos de utilização do cheque especial.

Introduzimos incertezas em nossa descrição tratando o par  $(X, Y) \in \mathbb{R}^p \times C$  como um vetor aleatório com função de distribuição conjunta  $F_{X,Y}$ .

Metáfora: um oráculo gera um  $X$  a partir da função de distribuição marginal  $F_X$  e depois disso gera um  $Y$  a partir da função de distribuição condicional  $F_{Y|X}(\cdot | x)$  (conhecida na literatura como “distribuição do supervisor”).

A função de distribuição conjunta fica determinada formalmente por

$$F_{X,Y}(x, y) = \int_{(-\infty, x]} F_{Y|X}(y | t) dF_X(t),$$

na qual usamos a notação  $(-\infty, x] := (-\infty, x_1] \times \cdots \times (-\infty, x_d]$ .

Estamos fazendo inferência: não conhecemos  $F_{X,Y}$ .

Nosso contexto é não paramétrico: a menos do suporte, não impomos quaisquer restrições a  $F_{X,Y}$ .

Definimos o *erro de classificação esperado* de um classificador  $\varphi$  pela probabilidade do classificador errar:  $\Pr\{\varphi(X) \neq Y\}$ , que será uma medida da “capacidade de generalização” do classificador.

Na literatura, o erro de classificação esperado também é denominado *erro de teste esperado*, ou apenas *erro de teste*. Novamente, utilizaremos estes termos sem diferenciação.

Para fazer a conexão com o *erro de predição esperado* definido no contexto de regressão note que  $\Pr\{\varphi(X) \neq Y\} = E[I_{\{\varphi(X) \neq Y\}}]$ .

Portanto, nos termos da Teoria da Decisão, o que está em jogo é a função de perda que estamos utilizando nos dois problemas (regressão e classificação).

No problema de regressão, da aula anterior, trabalhamos com uma função de perda quadrática e no problema de classificação utilizamos uma perda 0-1.

Defina o *classificador de Bayes* por

$$\varphi^*(x) = \begin{cases} 1 & \text{se } \Pr\{Y = 1 \mid X = x\} =: \eta(x) > 1/2; \\ 0 & \text{caso contrário.} \end{cases}$$

O classificador de Bayes tem um papel formal: em um problema real não conhecemos  $F_{X,Y}$ . Portanto, também não conhecemos  $\eta(x)$  e não conseguimos construir  $\varphi^*$ .

O erro de classificação esperado de  $\varphi^*$  é denominado *erro de Bayes*.

Quando temos dados simulados a partir de uma distribuição conhecida, podemos construir o classificador de Bayes  $\varphi^*$  explicitamente.

### Resultado

O classificador de Bayes é ótimo: para qualquer classificador  $\varphi$ , temos que  $\Pr\{\varphi^*(X) \neq Y\} \leq \Pr\{\varphi(X) \neq Y\}$ .

## Demonstração

Para qualquer classificador  $\varphi$  e um vetor de preditoras  $x \in \mathbb{R}^p$ , note que  $\Pr\{\varphi(X) = Y \mid X = x\} = \Pr\{\varphi(x) = Y \mid X = x\}$  é igual a  $\Pr\{Y = 0 \mid X = x\} = 1 - \eta(x)$ , quando  $\varphi(x) = 0$ , e é igual a  $\Pr\{Y = 1 \mid X = x\} = \eta(x)$ , quando  $\varphi(x) = 1$ . Assim,

$$\begin{aligned}\Pr\{\varphi(X) \neq Y \mid X = x\} &= 1 - \Pr\{\varphi(X) = Y \mid X = x\} \\ &= 1 - (I_{\{\varphi(x)=0\}}(1 - \eta(x)) + I_{\{\varphi(x)=1\}}\eta(x)) \\ &= \eta(x) - (2\eta(x) - 1)I_{\{\varphi(x)=1\}},\end{aligned}$$

uma vez que  $I_{\{\varphi(x)=0\}} = 1 - I_{\{\varphi(x)=1\}}$ . Portanto,

$$\begin{aligned}\Pr\{\varphi(X) \neq Y \mid X = x\} - \Pr\{\varphi^*(X) \neq Y \mid X = x\} \\ = (2\eta(x) - 1) (I_{\{\varphi^*(x)=1\}} - I_{\{\varphi(x)=1\}}).\end{aligned}$$

## Demonstração (continuação)

Temos dois casos: se  $\varphi^*(x) = 0$ , então, pela definição do classificador de Bayes, temos que  $2\eta(x) - 1 \leq 0$  e  $I_{\{\varphi^*(x)=1\}} - I_{\{\varphi(x)=1\}} \leq 0$ . Quando  $\varphi^*(x) = 1$ , temos que  $2\eta(x) - 1 \geq 0$  e  $I_{\{\varphi^*(x)=1\}} - I_{\{\varphi(x)=1\}} \geq 0$ . Assim, em ambos os casos, temos que

$$\Pr\{\varphi(X) \neq Y \mid X = x\} - \Pr\{\varphi^*(X) \neq Y \mid X = x\} \geq 0. \quad (*)$$

Pela definição de probabilidade condicional, para todo classificador  $\varphi$ , temos que

$$\Pr\{\varphi(X) \neq Y\} = \int \Pr\{\varphi(X) \neq Y \mid X = x\} dF_X(x).$$

Obtemos o resultado desejado integrando (\*) com respeito a  $dF_X(x)$ .

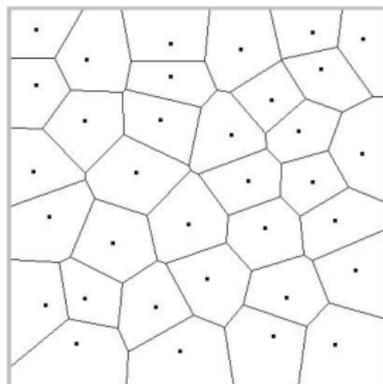
Suponha que temos uma certa função de distância definida no espaço das preditoras  $\mathbb{R}^p$ .

Para dados de treinamento  $(x_1, y_1), \dots, (x_n, y_n)$  e um novo  $x \in \mathbb{R}^p$ , o classificador dos  $k$  vizinhos mais próximos (*k-nearest neighbors*, que abreviaremos por *k-nn*) determina os  $k$  pontos em  $\{x_1, \dots, x_n\}$  mais próximos de  $x$  e classifica  $x$  como pertencente à classe mais frequente entre as respostas  $y_i$  destes  $k$  vizinhos mais próximos (voto da maioria).

Quando  $k$  é par, algum critério *ad-hoc* de desempate é adotado.

Diversas distâncias podem ser utilizadas: Euclidiana, Mahalanobis etc. A literatura sobre esta escolha é extensa, especialmente quando algumas das preditoras são categóricas.

No caso  $k = 1$ , as preditoras  $x_1, \dots, x_n$  dos dados de treinamento definem células de classificação que formam uma estrutura geométrica conhecida como tesselação (mosaico) de Voronoi.



Cover e Hart provaram em 1967 que, assintoticamente, o erro de classificação esperado da regra 1-NN nunca é maior do que o dobro do erro de Bayes, de maneira universal, ou seja, para qualquer  $F_{X,Y}$ .

De maneira análoga ao resultado discutido no contexto de regressão, não podemos utilizar a *taxa de erro de treinamento*

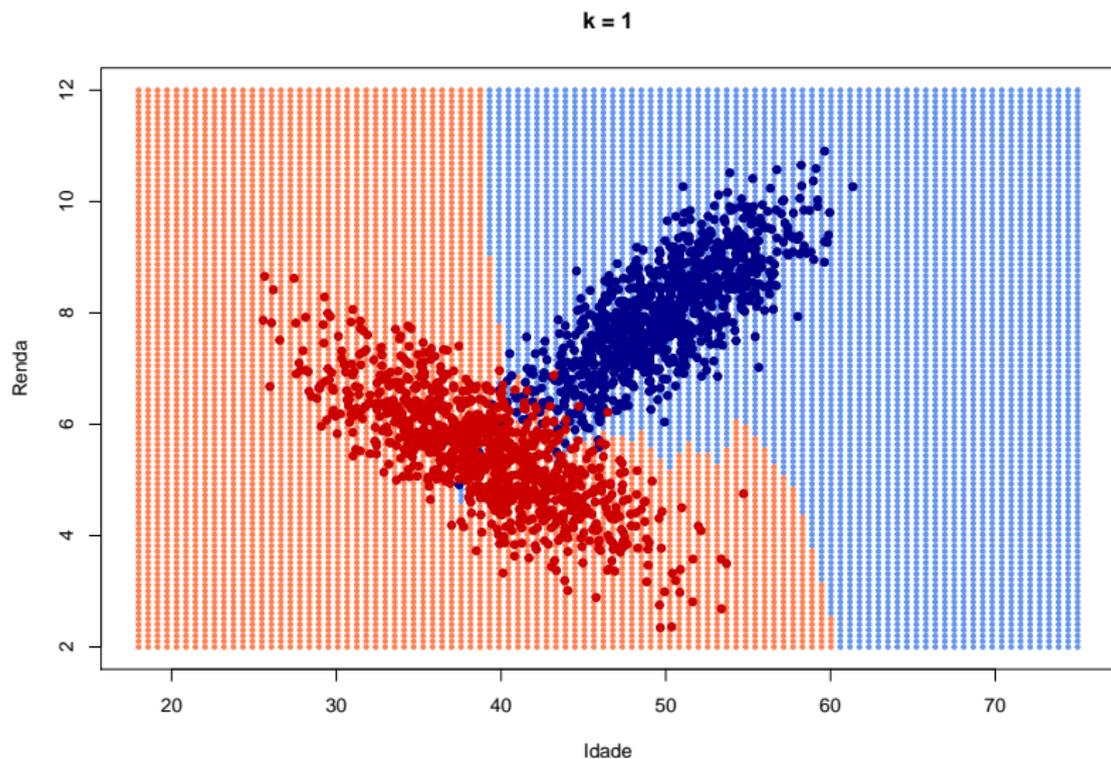
$$\frac{1}{n} \sum_{i=1}^n I_{\{\hat{\phi}(x_i)=y_i\}},$$

para escolher  $k$ , pois esta taxa é igual a 0 quando  $k = 1$  e escolheríamos sempre o classificador sobreajustado.

Escolhemos  $k$  minimizando a estimativa do erro de classificação esperado obtida por validação cruzada.

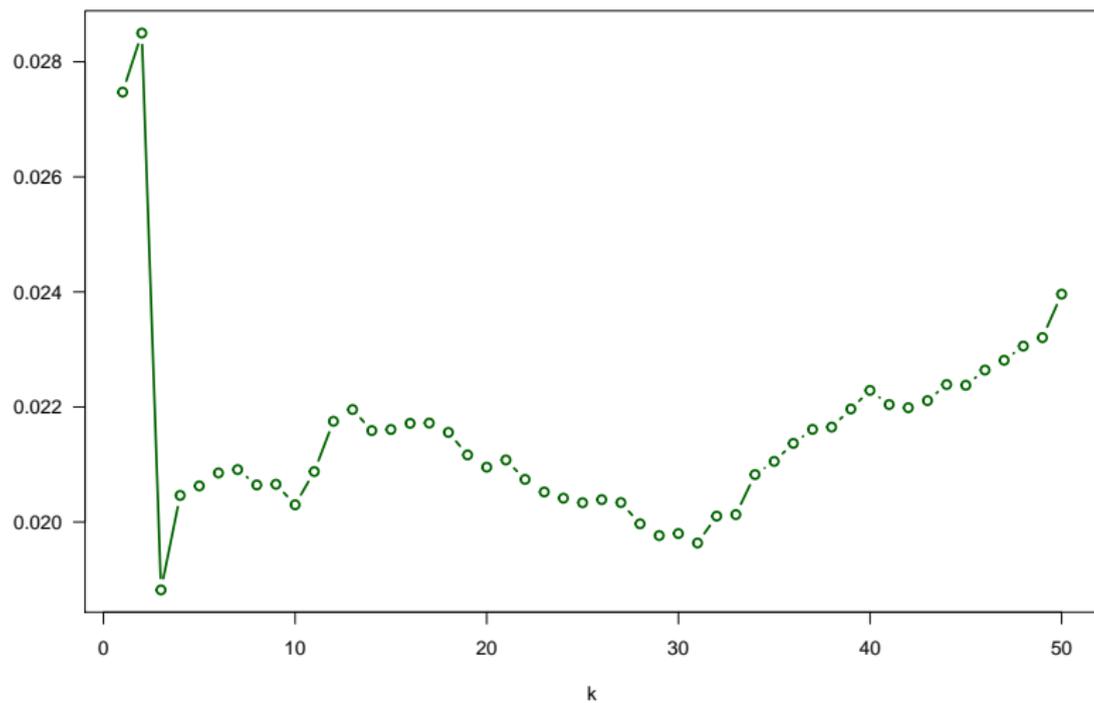
Quando temos muitas preditoras ( $p$  grande), o classificador  $k$ -nn sofre a “maldição da dimensionalidade”, pois, grosso modo, em um espaço euclidiano de dimensão muito alta todos os  $x_i$ 's estariam aproximadamente à mesma distância da origem.

Uma variante do  $k$ -nn é utilizada para regressão não paramétrica: ao invés do voto da maioria, toma-se a média das respostas dos  $k$  vizinhos mais próximos.

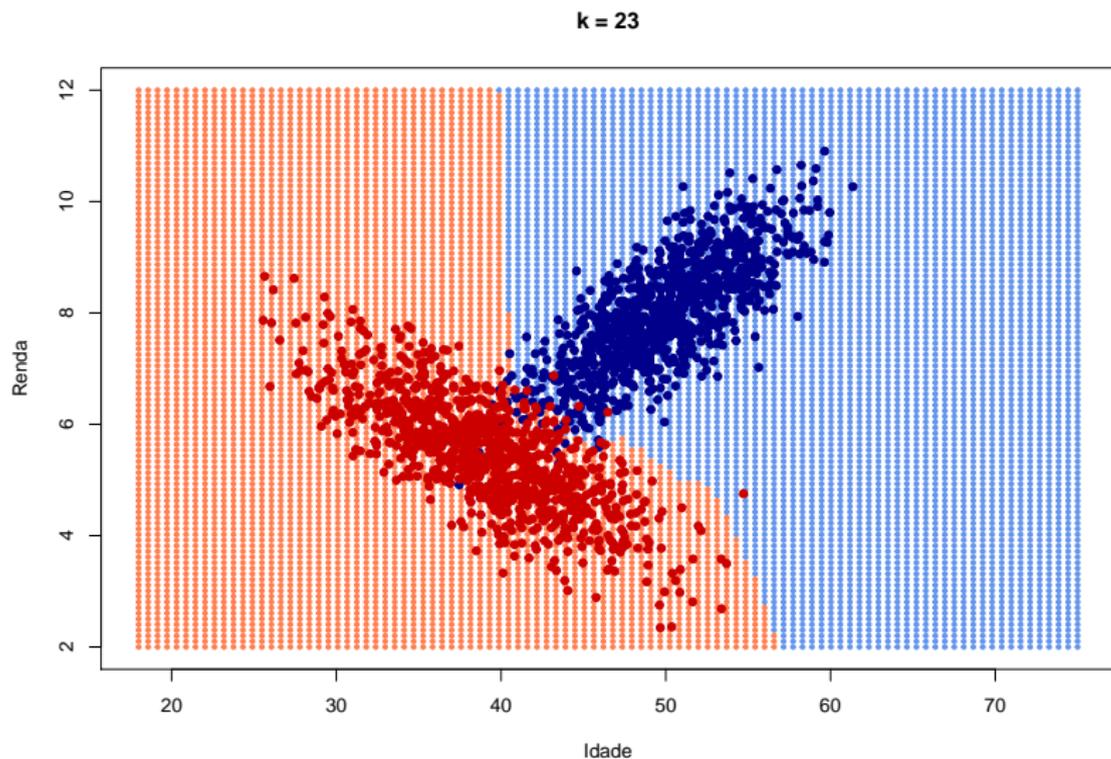


# Escolhendo $k$ via validação cruzada

Validação cruzada em 10 lotes



## [1] 3



# Aproximação de Monte Carlo para o erro de Bayes Insper

```
library(mvtnorm)

N <- 10^4
y_bayes <- numeric(N)

x <- matrix(nrow = N, ncol = 2)
y <- sample(c(0,1), N, replace = TRUE)

for (i in 1:N) {
  if (y[i] == 0) { x[i,] <- mvrnorm(1, mu = mu0, Sigma = Sigma0) }
  else          { x[i,] <- mvrnorm(1, mu = mu1, Sigma = Sigma1) }

  if (1 + exp( dmvnorm(x[i,], mean = mu0, sigma = Sigma0, log = TRUE)
    - dmvnorm(x[i,], mean = mu1, sigma = Sigma1, log = TRUE) ) < 2) {
    y_bayes[i] <- 1
  } else {
    y_bayes[i] <- 0
  }
}

bayes_error <- mean(y_bayes != y)
cat("Bayes error = ", sprintf("%.4f\n", bayes_error))
```

```
## Bayes error = 0.0183
```

O conjunto de dados **Advertising** e as figuras dos slides 12, 33 e 34 foram extraídos de

*An Introduction to Statistical Learning, with applications in R,*

com a permissão dos autores G. James, D. Witten, T. Hastie and R. Tibshirani.

<http://www-bcf.usc.edu/~gareth/ISL/>