Fundamentos de Aprendizagem Estatística: Regressão

Paulo C. Marques F. e Hedibert F. Lopes

Sexta-feira, 20 de Outubro de 2017

Insper

Aprendizagem supervisionada (1)

Insper

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

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

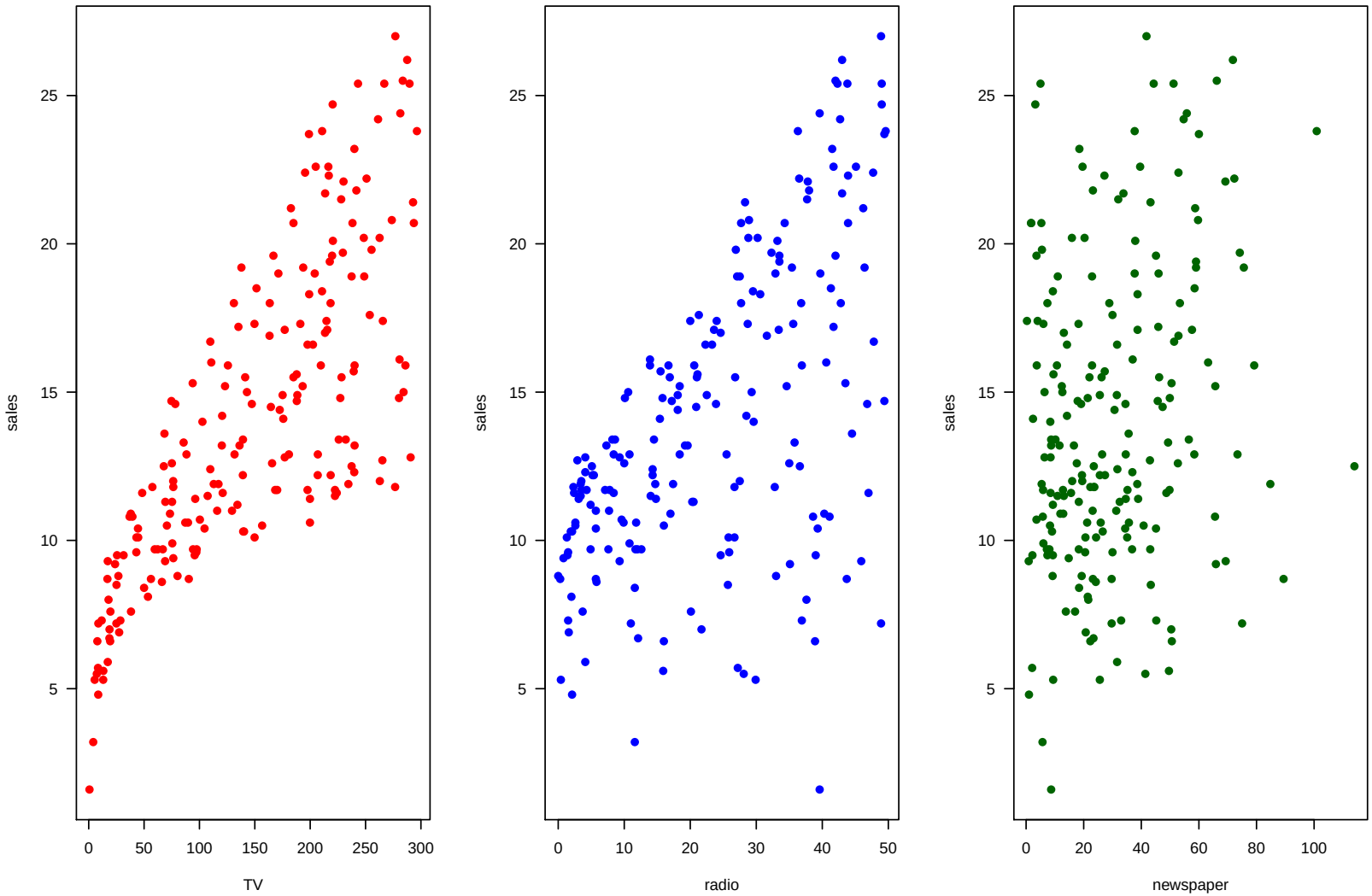
Exemplo: no conjunto de dados `Advertising` temos informações sobre as vendas de um produto em 200 mercados.

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.

Aprendizagem supervisionada (2)

Insper



Sinonímia e problemas não supervisionados

Insper

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* (tema do nosso último módulo) não temos uma variável resposta.

O problema não supervisionado exemplar é a análise de conglomerados (*clusters*).

Regressão e classificação

Insper

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

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

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

Na aula de hoje trataremos dos fundamentos dos problemas de regressão.

No final das aulas de amanhã examinaremos os fundamentos dos problemas de classificação.

Nota: na literatura há um pequeno "bug" terminológico, pois *regressão logística* se refere a um método de classificação.

Função de regressão (1)

Insper

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

Problema

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

Observações:

1. Em Aprendizagem Estatística, o erro de predição esperado também é denominado *erro de teste esperado*. Nas aulas, usaremos os dois 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.

Função de regressão (2)

O erro de predição esperado

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

pode ser reescrito utilizando as propriedades da esperança condicional:

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

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

Função de regressão (3)

Insper

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

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

Zerando a derivada de g em relação a t obtemos

$$g'(t) = -2\,\mathrm{E}[Y \mid X = x] + 2t = 0$$

se e somente se $t = \mathrm{E}[Y \mid X = x]$. Note também que $g''(t) = 2 > 0$.

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

Função de regressão (4)

Insper

Uma segunda maneira de solucionar o problema é observar que

$$
\begin{aligned}
&\mathrm{E}[(Y-t)^2 \mid X=x] \\
&\quad = \mathrm{E}[((Y-\mathrm{E}[Y \mid X=x]) - (t-\mathrm{E}[Y \mid X=x]))^2 \mid X=x] \\
&\quad = \mathrm{E}[(Y-\mathrm{E}[Y \mid X=x])^2 \mid X=x] \\
&\qquad - 2\underbrace{(\mathrm{E}[Y \mid X=x]-\mathrm{E}[Y \mid X=x])}_{=0}(t-\mathrm{E}[Y \mid X=x]) \\
&\qquad + (t-\mathrm{E}[Y \mid X=x])^2 \\
&\quad = \mathrm{E}[(Y-\mathrm{E}[Y \mid X=x])^2 \mid X=x] + (t-\mathrm{E}[Y \mid X=x])^2.
\end{aligned}
$$

Portanto, novamente minimizamos o erro de predição esperado escolhendo $t = \psi(x) = \mathrm{E}[Y \mid X=x]$.

Aprendendo sobre ψ

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 $\mathrm{E}[Y \mid X = x]$.

Suponha que temos *dados de treinamento* $(X_1, Y_1), \ldots (X_n, Y_n)$ 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 $\hat{\psi} : \mathbb{R}^p \to \mathbb{R}$ que será nossa estimativa (nosso "chute") para ψ.

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

Predição e inferência (1)

Insper

Há dois propósitos, em geral antagônicos, quando se trata de aprender sobre ψ.

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 ψ como uma "caixa-preta" e queremos $\hat{\psi}$ que, para a variável resposta Y correspondente, minimize o erro de predição esperado $\mathrm{E}[(Y - \hat{\psi}(X))^2]$.

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.

Predição e inferência (2)

Insper

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:

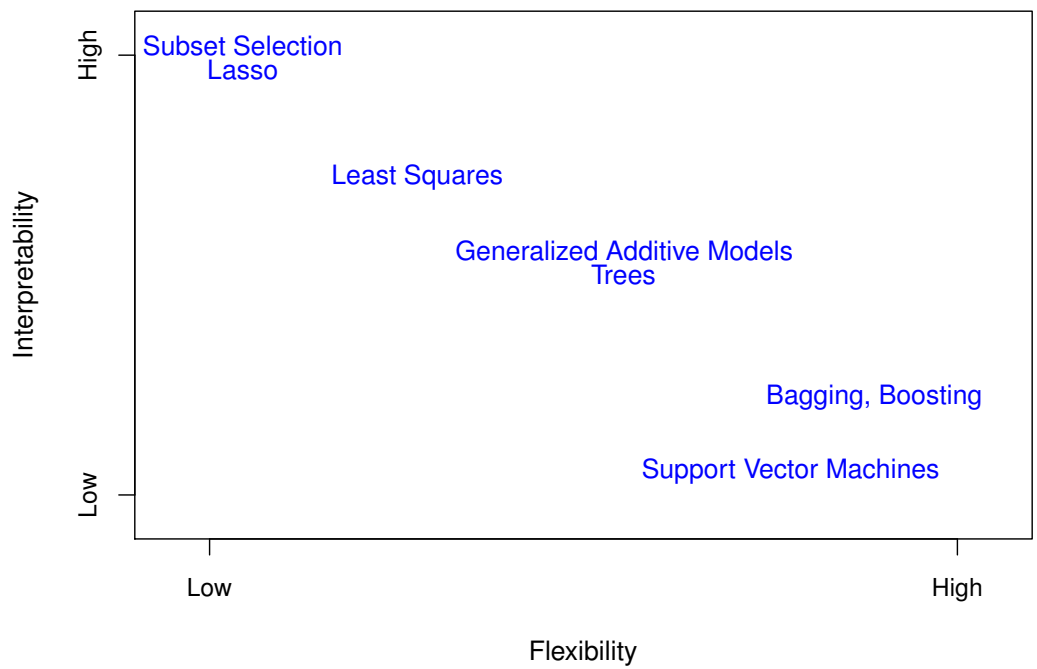
- Há uma associação entre `TV` e `sales`?
- Um aumento de 25% em `newspaper` levaria em média a que mudança em `sales`?

Predição e inferência (3)

Insper

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.



Modelo aditivo (1)

Insper

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* ϵ é uma variável aleatória, independente de X, cuja distribuição é tal que $\mathrm{E}[\epsilon] = 0$ e $\mathrm{Var}[\epsilon] = \sigma^2$ (desconhecida).

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

Neste modelo, ψ 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.

Modelo aditivo (2)

Insper

Incorporando os dados de treinamento $(X_1, Y_1), \ldots (X_n, Y_n)$ e novos erros aleatórios $\epsilon_1, \ldots, \epsilon_n$, 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.
- ϵ e os ϵ_i's têm a mesma distribuição, com esperança 0 e variância σ^2 (desconhecida).
- X, os X_i's, ϵ e os ϵ_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 $(X_1, Y_1), \ldots (X_n, Y_n)$.

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

Erro redutível e irredutível (1)

Insper

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

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

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

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

Resultado

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

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

Erro redutível e irredutível (2)

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 σ^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

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

Decomposição do erro redutível (1)

Insper

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

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

Decomposição do erro redutível (2)

Insper

Demonstração

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

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

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

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.

Perde-ganha viés-variância

Insper

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 ψ sob replicação do processo de aprendizagem.

Por exemplo, ao usarmos um modelo linear para aprender sobre uma ψ 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).

Concretizando os conceitos (1)

Insper

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 ψ 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.

Concretizando os conceitos (2)

Insper

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

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

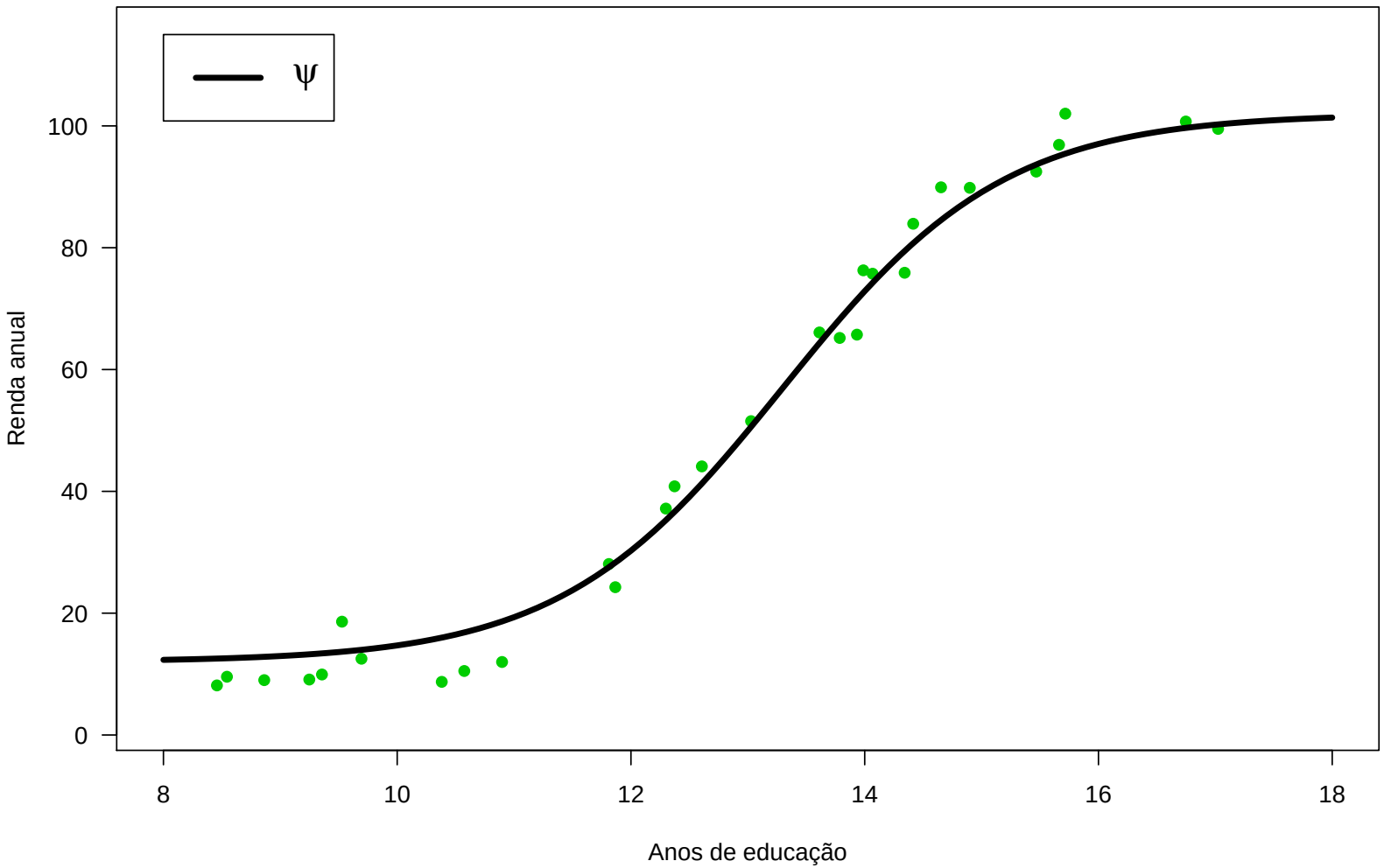
Para a preditora X utilizaremos uma distribuição uniforme entre 8 e 18 anos de estudo.

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 ψ e de 30 pares (x_i, y_i) de dados de treinamento gerados pelo modelo aditivo especificado.

Função de regressão e dados de treinamento

Insper



Métodos de aprendizagem (1)

Insper

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.

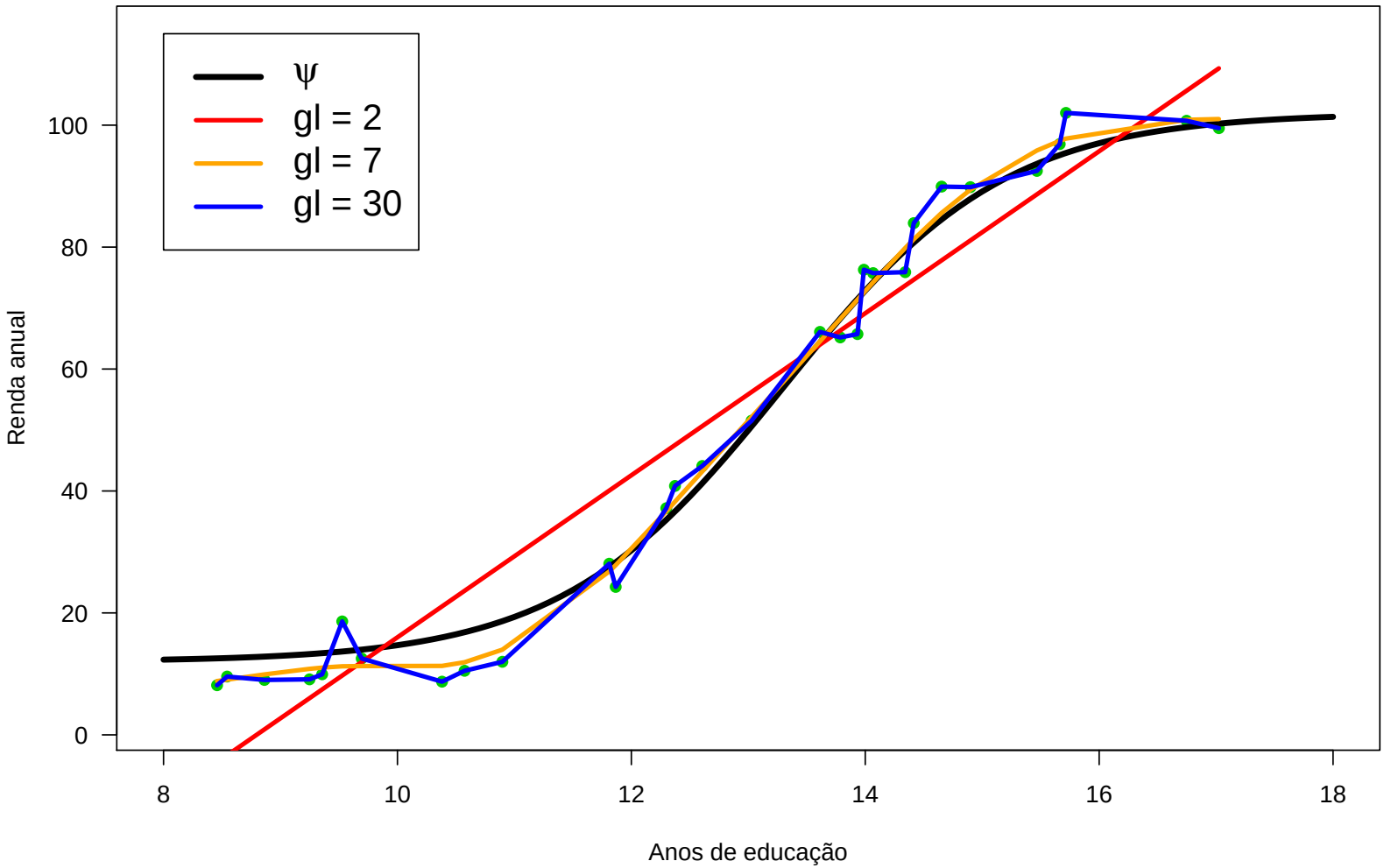
Quando temos gl $= 2$, ao treinar o modelo obtemos uma $\hat{\psi}$ que é a reta de mínimos quadrados ordinários.

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.

Métodos de aprendizagem (2)

Insper



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

Para gl $= 2, \ldots, n$, repetimos o seguinte procedimento.

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

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

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

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

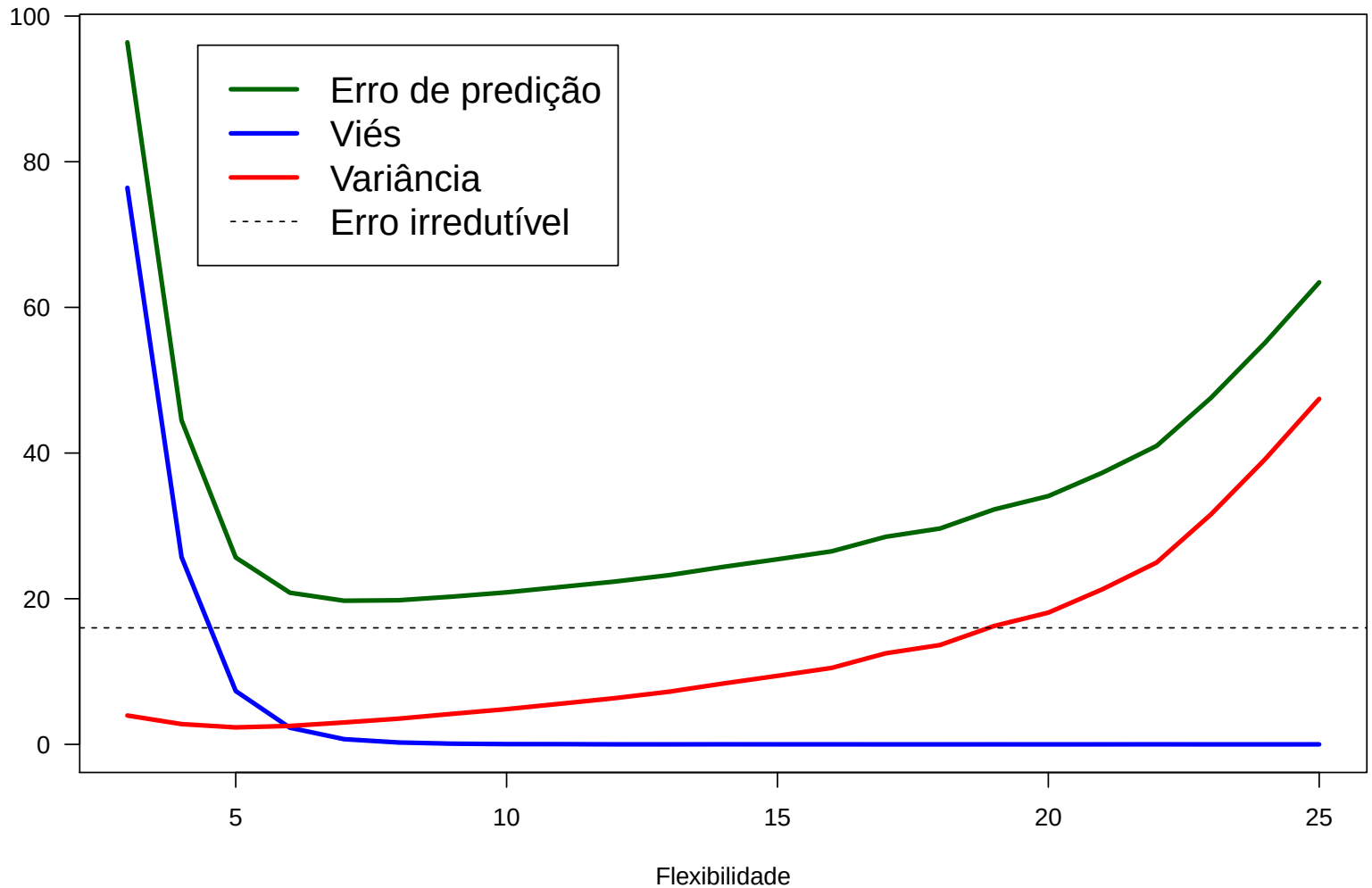
Aproximamos o viés ao quadrado esperado por $\frac{1}{M}\sum_{i=1}^{M}\text{Viés}^2[\hat{\psi}^{(\text{gl})}(x^{(i)})]$.

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

Na figura a seguir, temos os resultados da simulação para os modelos considerados.

Simulação do perde-ganha viés-variância (1)

Insper



Simulação do perde-ganha viés-variância (2)

Insper

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*).

Erro quadrático médio de treinamento (1)

Insper

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

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

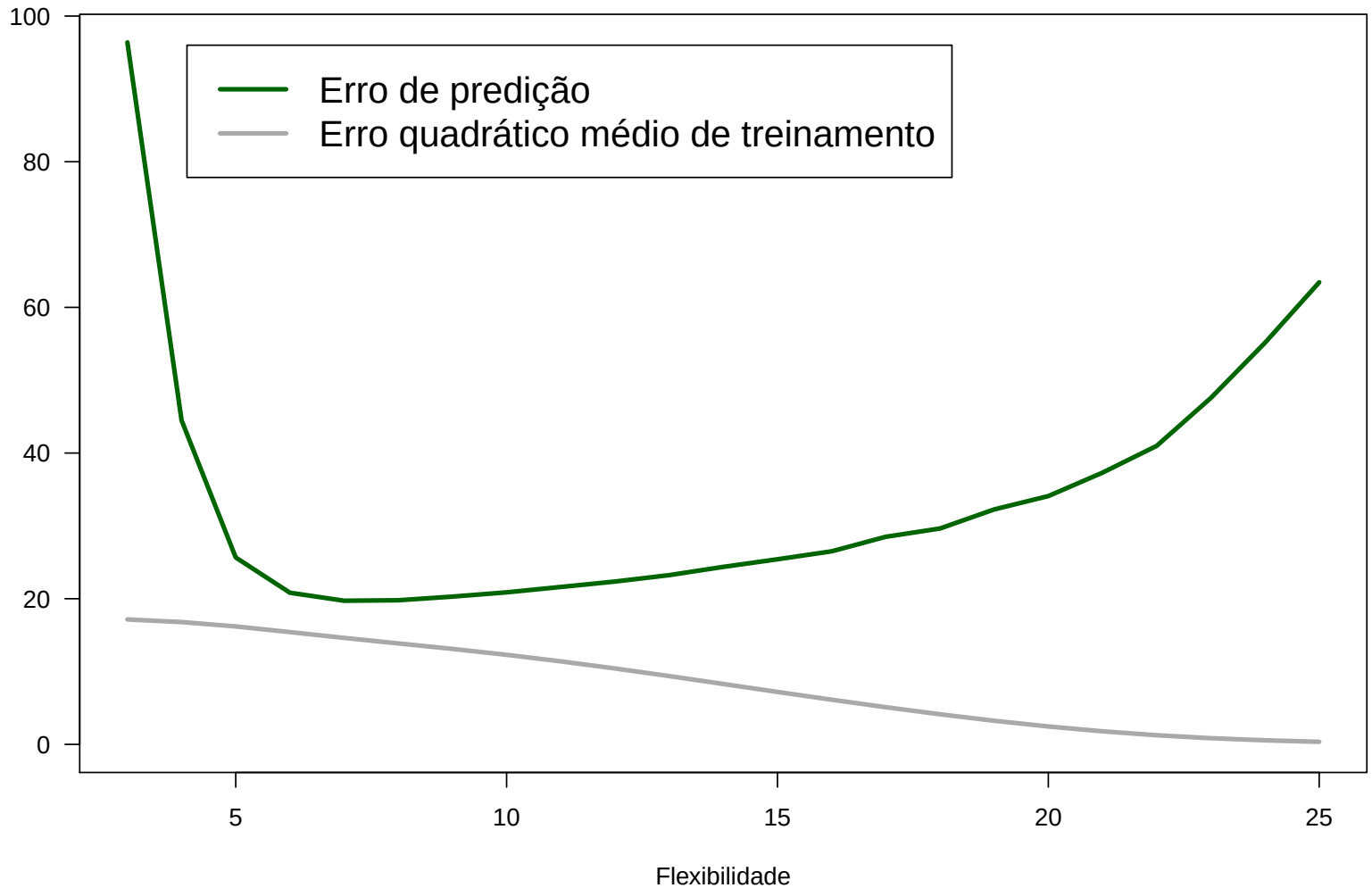
$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\psi}(x_i))^2.$$

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)

Insper



Erro quadrático médio de treinamento (3)

Insper

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.

Validação cruzada tirando-um-fora (1)

Insper

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 supervisonada que fornece, a partir de dados de treinamento $(x_1, y_1), \ldots, (x_n, y_n)$, uma estimativa $\hat{\psi}$ para a função de regressão ψ.

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.

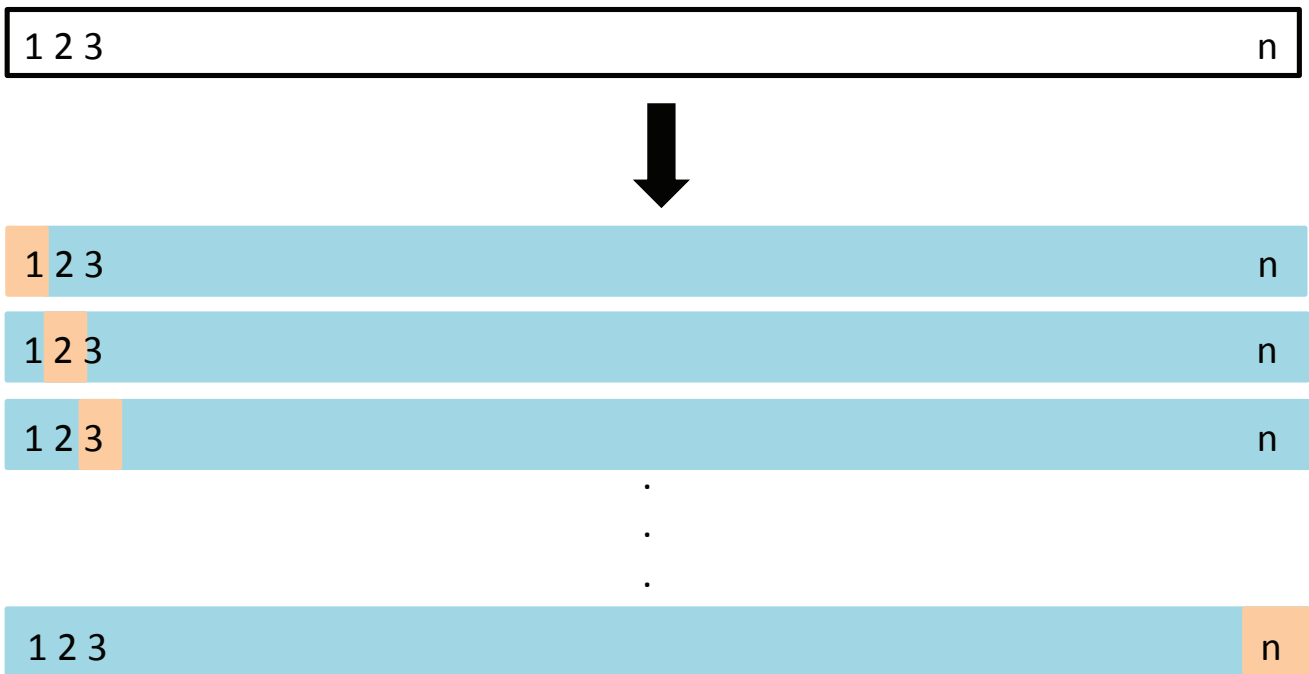
Para $i = 1, \ldots, n$, remova o par (x_i, y_i) dos dados de treinamento e calcule a estimativa $\hat{\psi}_{(\setminus i)}$ com os $n - 1$ dados de treinamento remanescentes, obtendo $\text{eqm}_i = (y_i - \hat{\psi}_{(\setminus i)}(x_i))^2$. A estimativa LOOCV para o erro de predição esperado é

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

Validação cruzada tirando-um-fora (2)

Insper

A figura abaixo ilustra o procedimento LOOCV.



Validação cruzada em k lotes (1)

Insper

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 $i = 1, \ldots, k$, 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

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

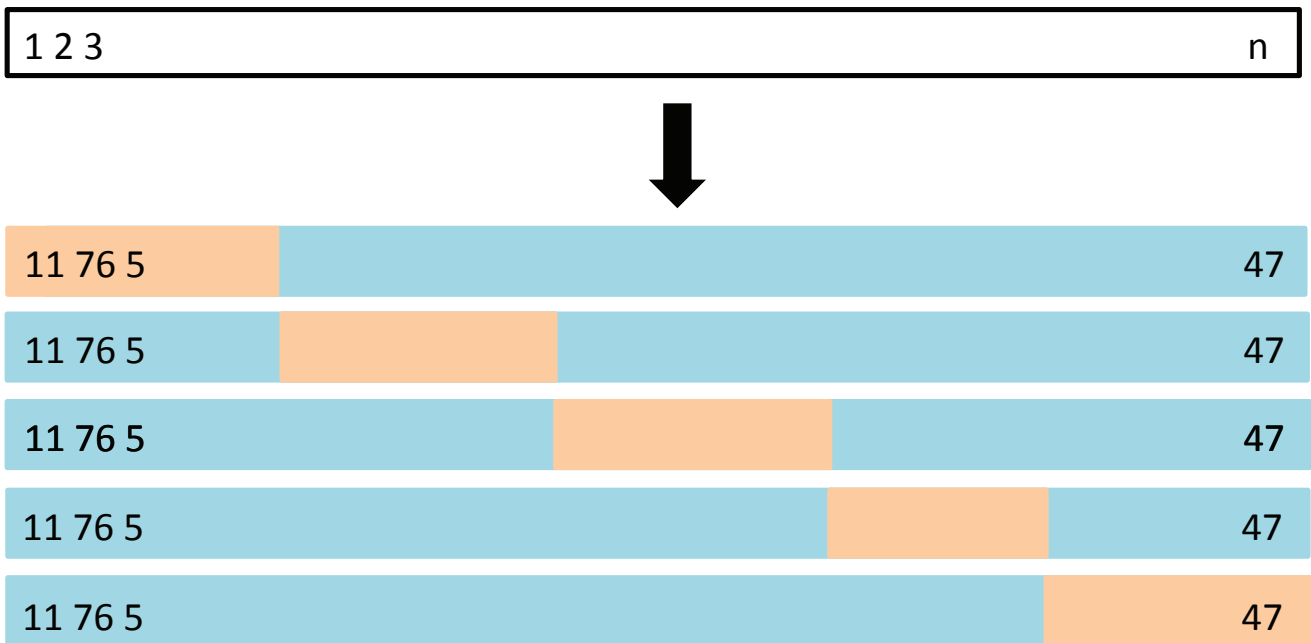
A estimativa da validação cruzada em k lotes para o erro de predição esperado é

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

Validação cruzada em k lotes (2)

Insper

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



Relações entre os dois procedimentos

Insper

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.

Créditos

O conjunto de dados `Advertising` e as figuras dos slides 13, 33 e 35 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/