Minicurso: Aprendizado Estatístico de Máquina

Rafael Izbicki

http://www.rizbicki.ufscar.br/

http://www.small.ufscar.br/

Support: **CNPq, FAPESP, NSF**

Material

- "Machine Learning sob a ótica estatística" (com Tiago Mendonça) (`http://www.rizbicki.ufscar.br/sml`)
- James, G., Witten, D., Hastie, T. e Tibshirani, R. *An Introduction to Statistical Learning, with Applications in R*, Springer 2013.
- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning. Vol. 1. New York: Springer series in statistics*, 2001.

Material

- "Machine Learning sob a ótica estatística" (com Tiago Mendonça) (`http://www.rizbicki.ufscar.br/sml`)
- James, G., Witten, D., Hastie, T. e Tibshirani, R. *An Introduction to Statistical Learning, with Applications in R*, Springer 2013.
- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning. Vol. 1. New York: Springer series in statistics*, 2001.

Material

- "Machine Learning sob a ótica estatística" (com Tiago Mendonça) (`http://www.rizbicki.ufscar.br/sml`)
- James, G., Witten, D., Hastie, T. e Tibshirani, R. *An Introduction to Statistical Learning, with Applications in R*, Springer 2013.
- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning. Vol. 1. New York: Springer series in statistics*, 2001.

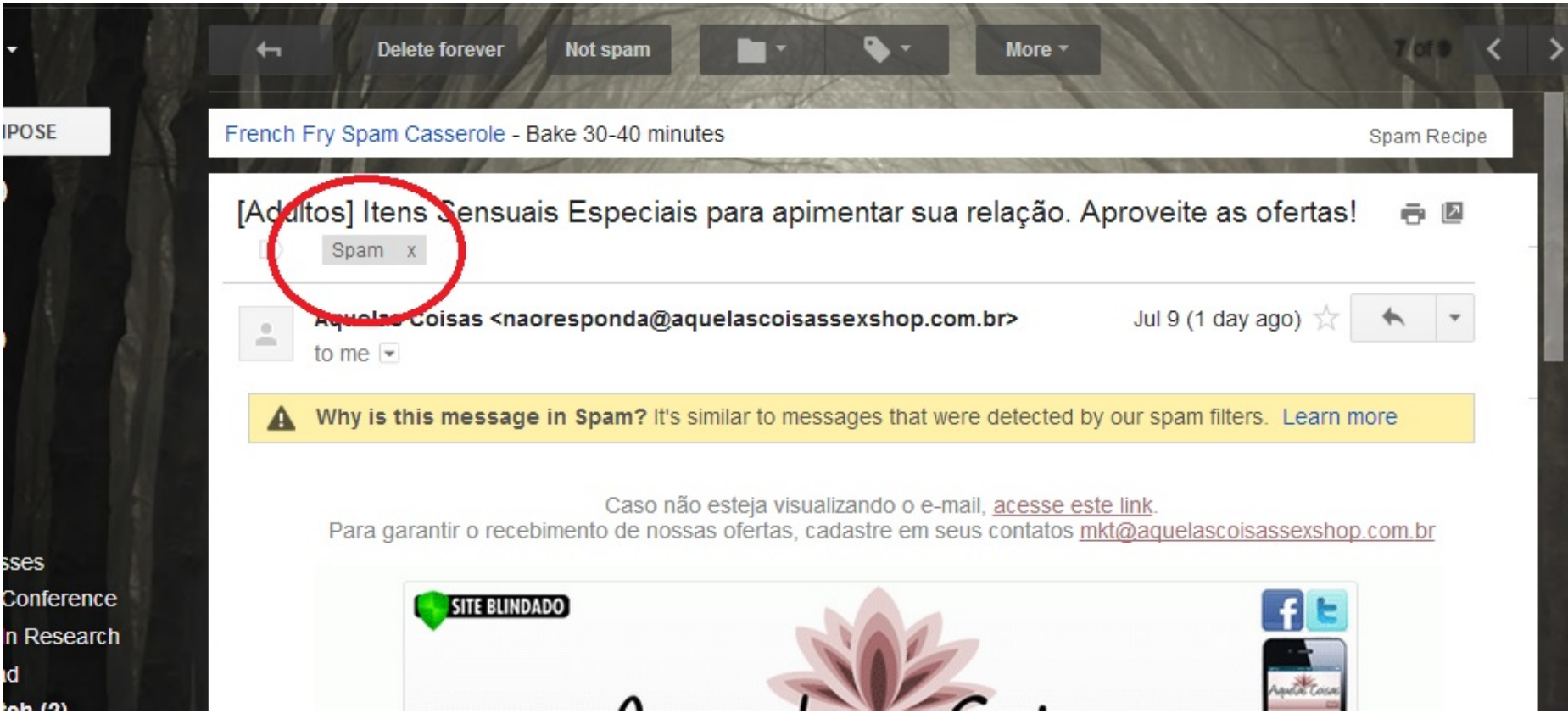
O que é o Machine Learning?

- Aprendizado supervisionado: Dadas medições $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, aprender um modelo para prever Y_i baseado em $\mathbf{X}_i$
- Aprendizado não supervisionado: Dadas medições $\mathbf{X}_1, \ldots, \mathbf{X}_n$, descobrir alguma estrutura com base em similaridade

O que é o Machine Learning?

- Aprendizado supervisionado: Dadas medições $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, aprender um modelo para prever Y_i baseado em $\mathbf{X}_i$
- Aprendizado não supervisionado: Dadas medições $\mathbf{X}_1, \ldots, \mathbf{X}_n$, descobrir alguma estrutura com base em similaridade

Exemplo: Detecção de Spams

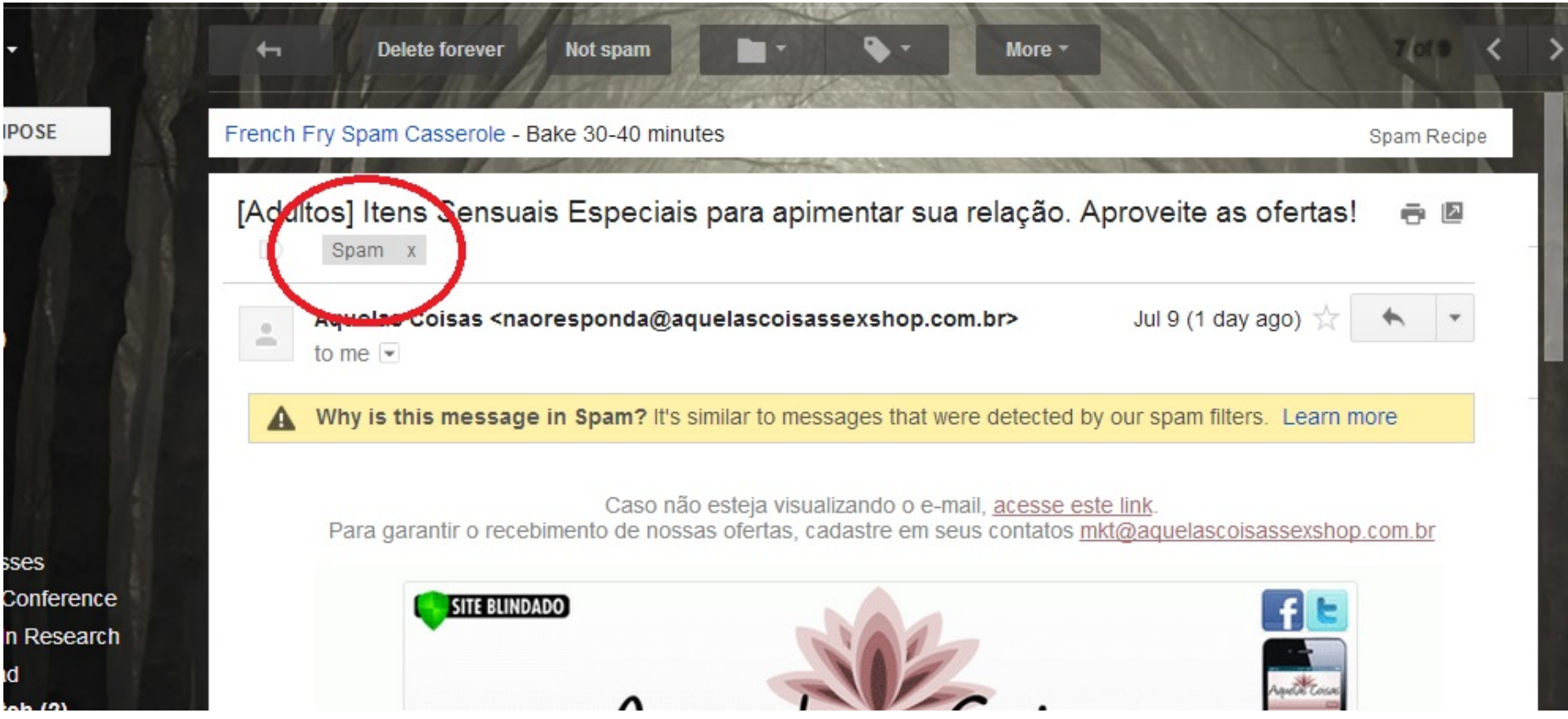


$\mathbf{X}_i \longrightarrow$ email

$Y_i \longrightarrow$ spam/não spam

Objetivo: prever Y_i com base em $\mathbf{X}_i$

Exemplo: Detecção de Spams

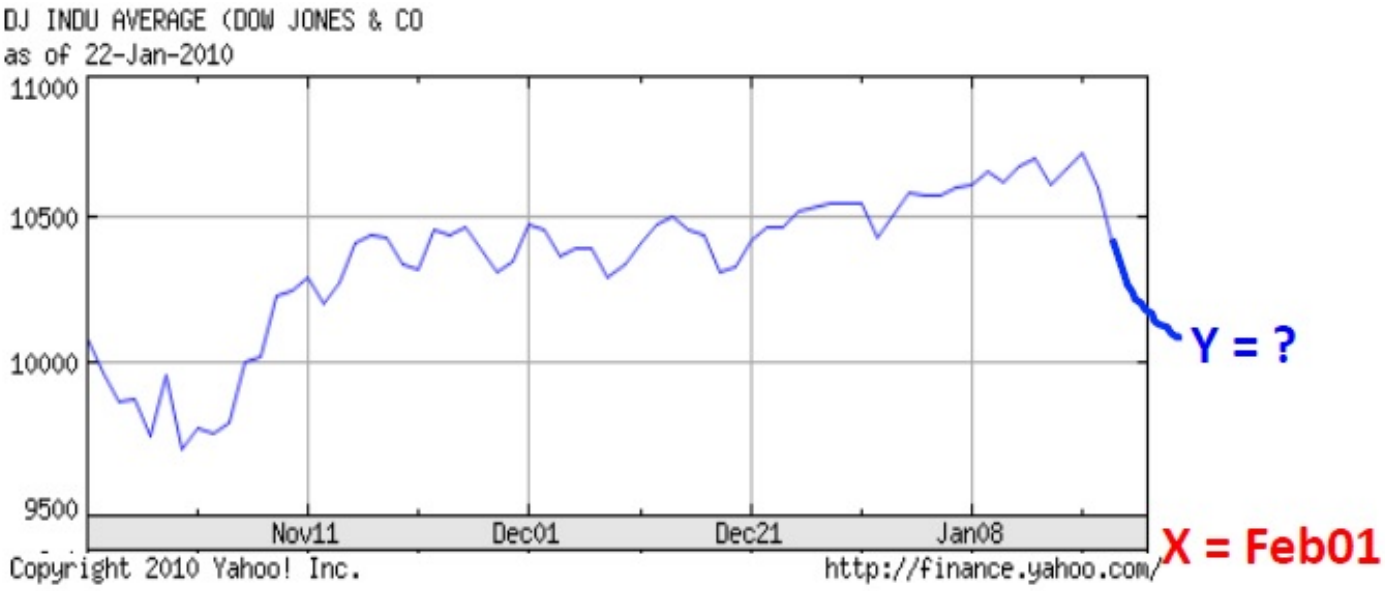


$\mathbf{X}_i \longrightarrow$ email

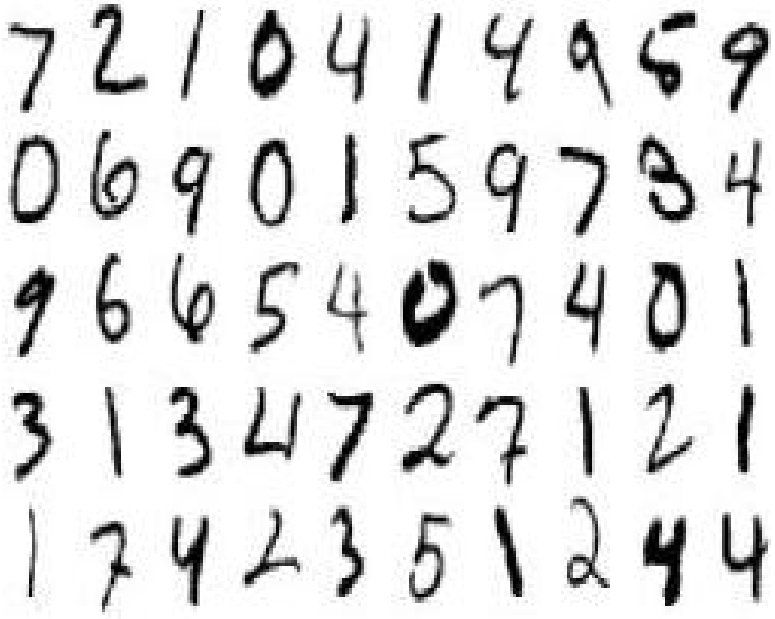
$Y_i \longrightarrow$ spam/não spam

Objetivo: prever Y_i com base em $\mathbf{X}_i$

Exemplo: Predição da Bolsa



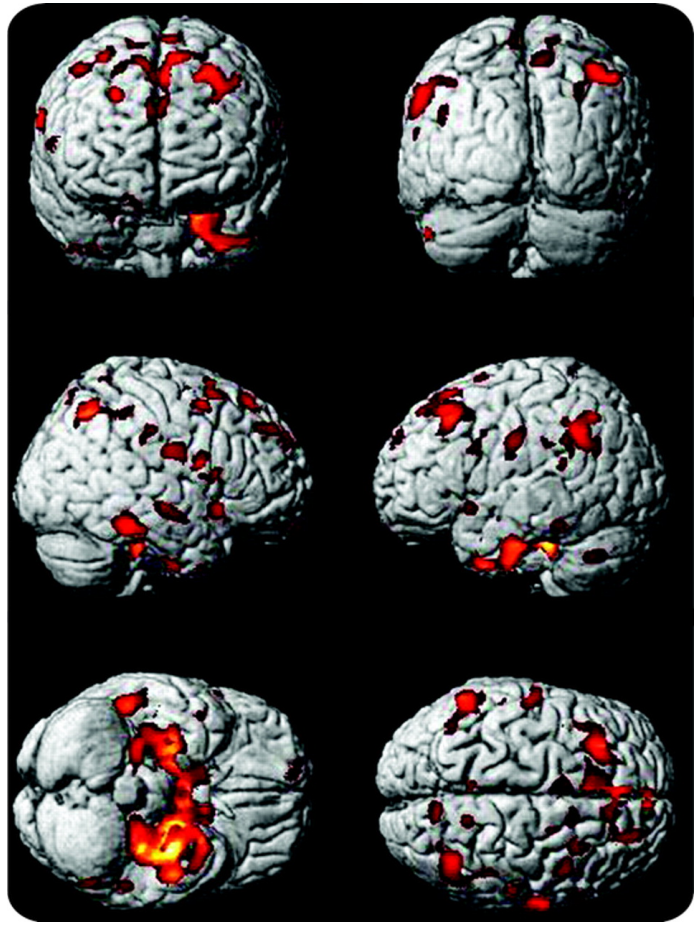
Exemplo: Reconhecimento de Dígitos



$\mathbf{X}_i \longrightarrow$ imagem de um dígito

$Y_i \longrightarrow$ dígito correspondente

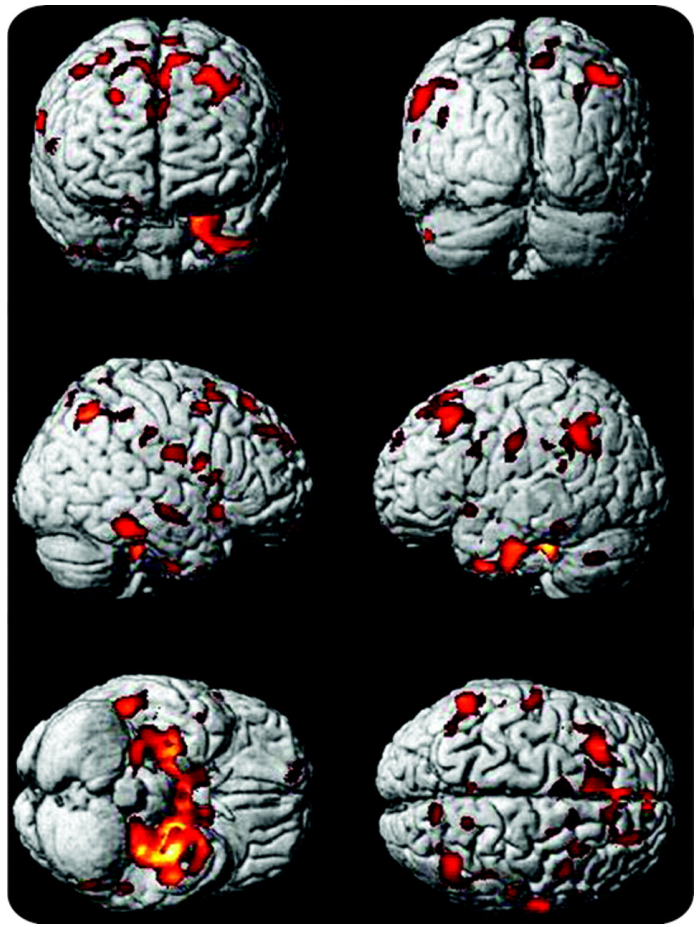
Exemplo: Predição de Alzheimer



$\mathbf{X}_i \longrightarrow$ imagem da ressonância magnética

$Y_i \longrightarrow$ Paciente com/sem Alzheimer

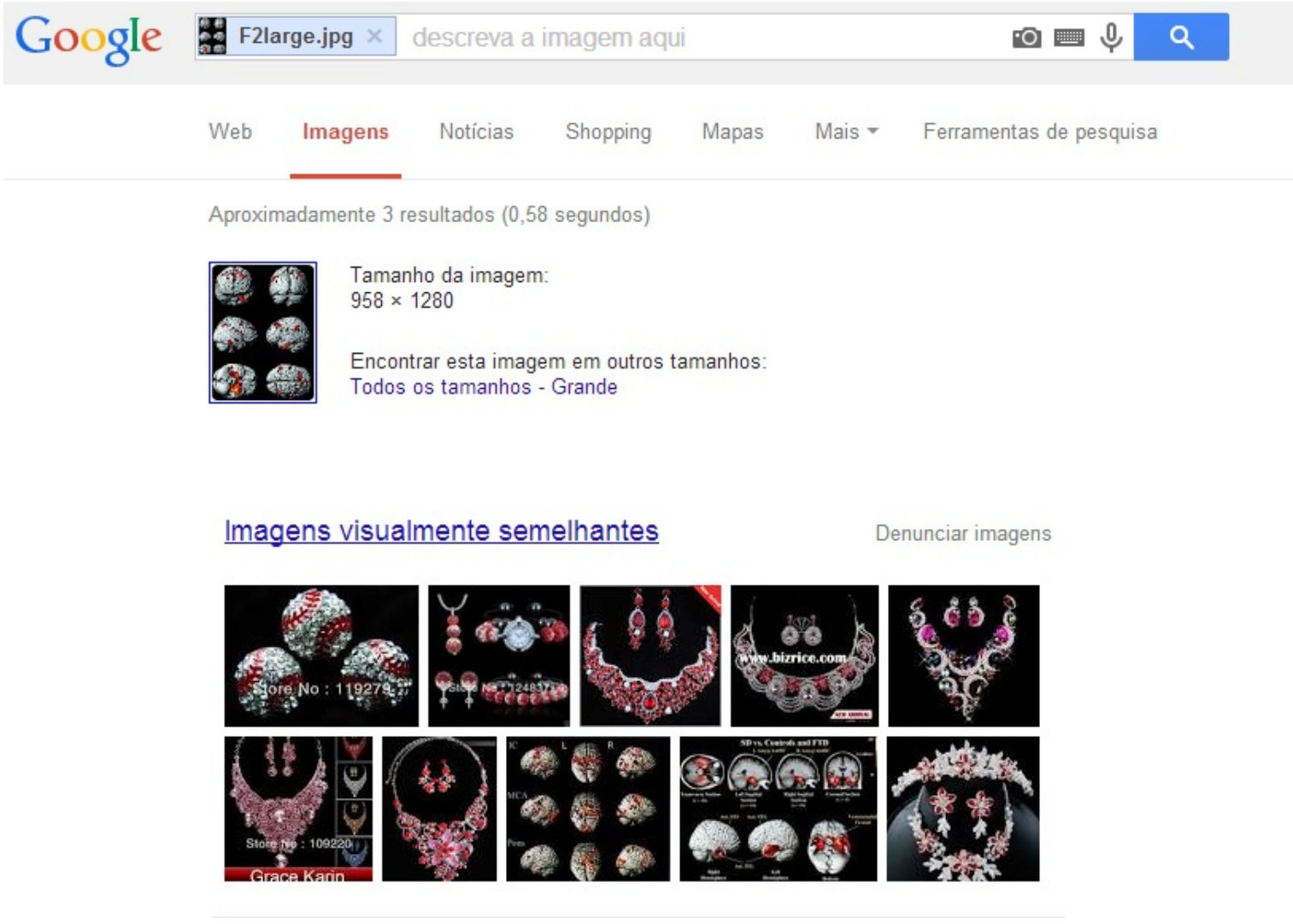
Exemplo: Predição de Alzheimer



$\mathbf{X}_i \longrightarrow$ imagem da ressonância magnética

$Y_i \longrightarrow$ Paciente com/sem Alzheimer

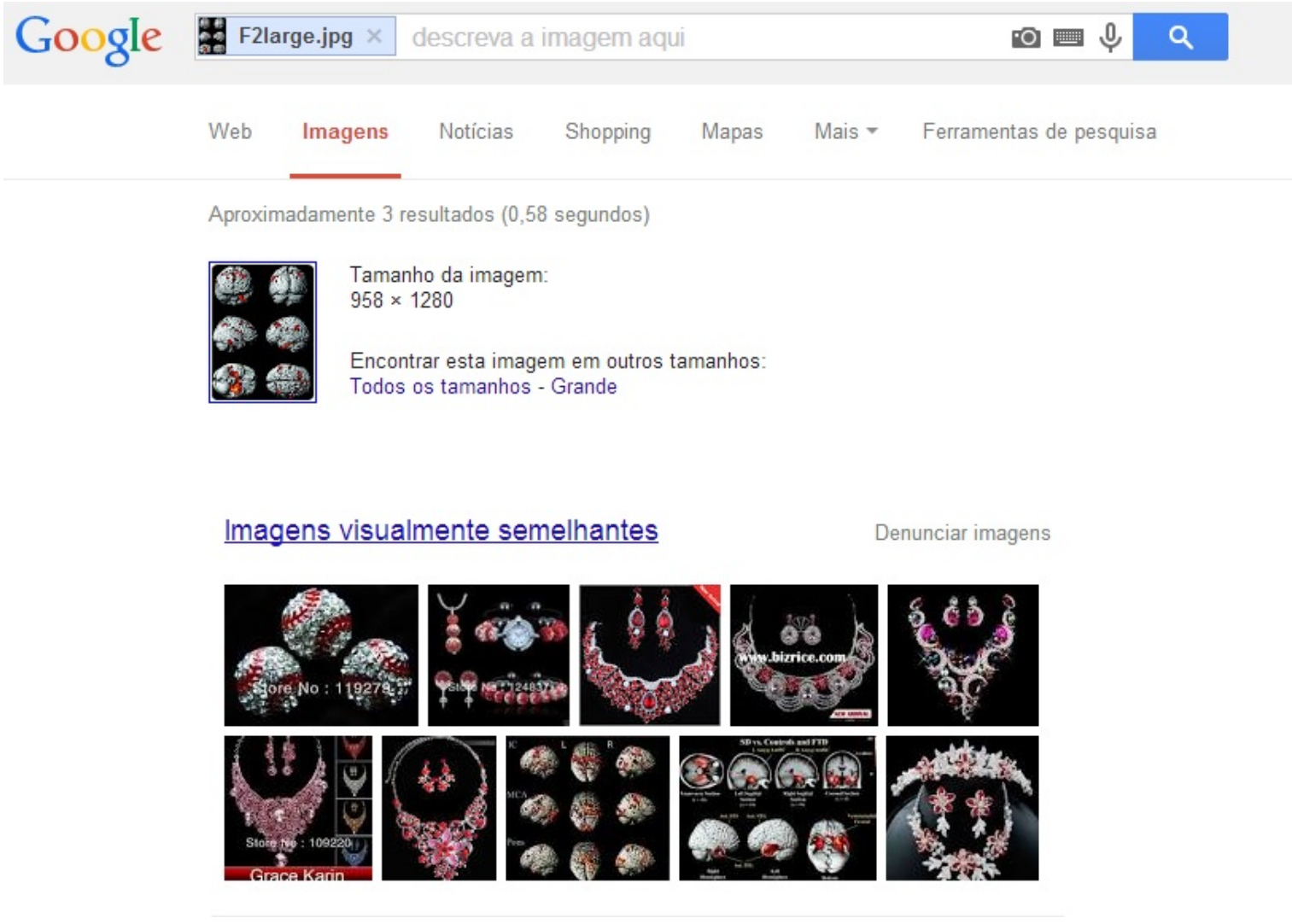
Exemplo: Busca por Imagens Semelhantes



$\mathbf{X}_i \longrightarrow$ imagens na internet

Busca por estrutura

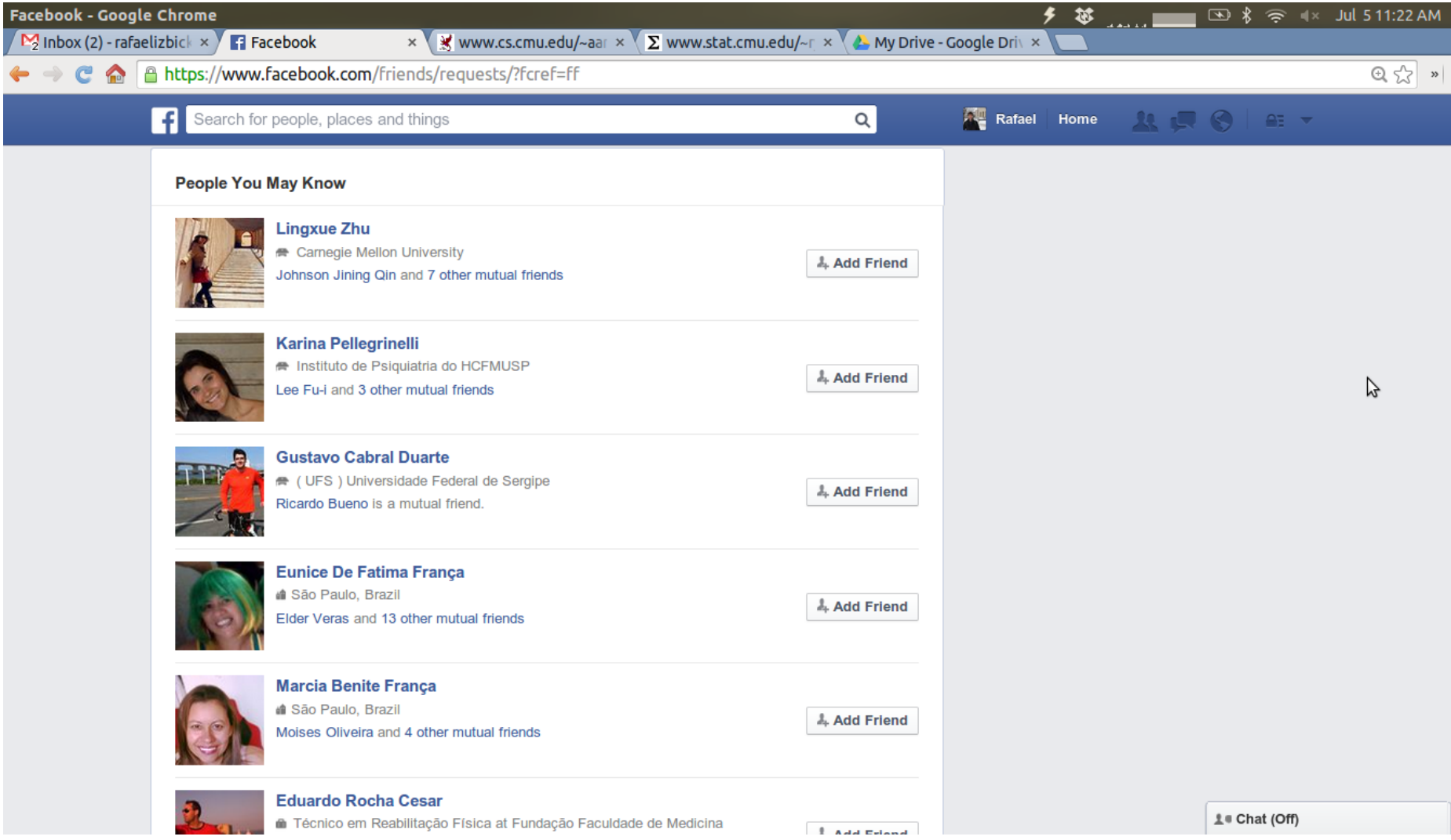
Exemplo: Busca por Imagens Semelhantes



$\mathbf{X}_i \longrightarrow$ imagens na internet

Busca por estrutura

Exemplo: Recomendação de Amizades



$\mathbf{X}_i \longrightarrow$ existe um link entre dois usuários

Exemplo: Sistemas de Recomendação

Linha: ida de uma pessoa a um supermercado
Coluna: comprou determinado produto?

Objetivo: descobrir regras do tipo

"Quem compra leite em geral também compra pão",

"Quem compra cerveja e refrigerante em geral também compra carne".

"Quem compra fralda em geral também compra cerveja".

Exemplo: Sistemas de Recomendação

Linha: ida de uma pessoa a um supermercado
Coluna: comprou determinado produto?

Objetivo: descobrir regras do tipo

"Quem compra leite em geral também compra pão",

"Quem compra cerveja e refrigerante em geral também compra carne".

"Quem compra fralda em geral também compra cerveja".

Exemplo: Sistemas de Recomendação

Linha: ida de uma pessoa a um supermercado
Coluna: comprou determinado produto?

Objetivo: descobrir regras do tipo

"Quem compra leite em geral também compra pão",

"Quem compra cerveja e refrigerante em geral também compra carne".

"Quem compra fralda em geral também compra cerveja".

Exemplo: Sistemas de Recomendação

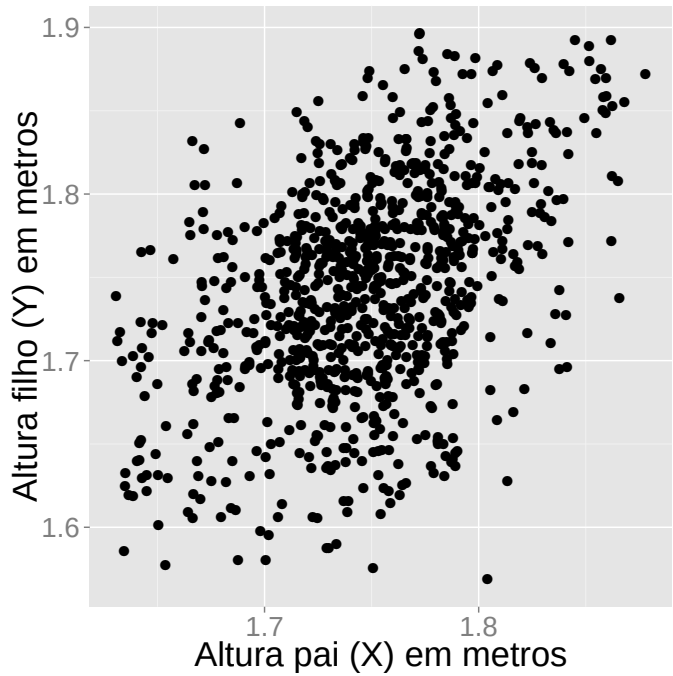
O que vamos ver nesse minicurso?

- Elementos do aprendizado supervisionado
  - Risco
  - Overfitting/Underfitting
  - Data Splitting/Validação Cruzada
- Regressão
  - Lasso
  - KNN
  - Árvores
  - Bagging/Florestas Aleatórias
- Classificação
  - Classificadores plugin: logística, naive Bayes

Foco inicial: Y quantitativo

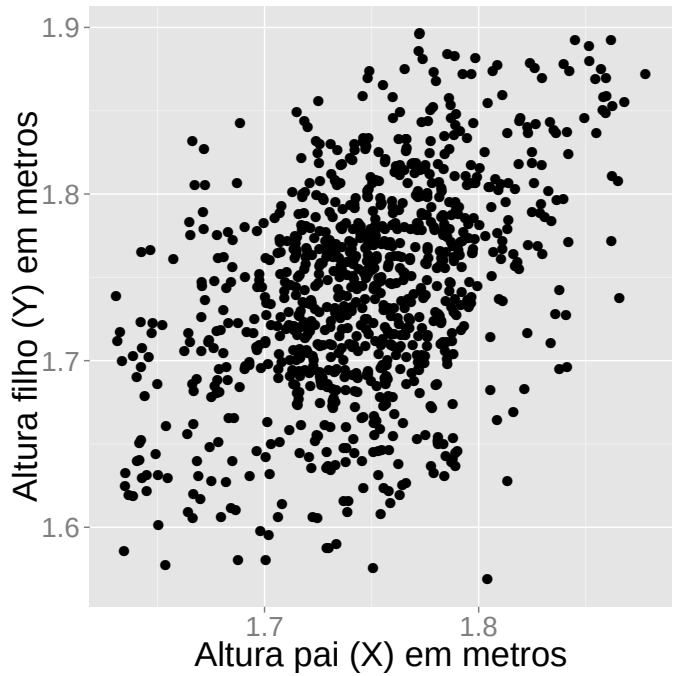
Regressão

Prever a altura de um filho (Y) com base na altura de seu pai ($\mathbf{X}$).
Amostra $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$



Criar uma função de predição $g(\mathbf{x})$: dado que a altura do pai de um indivíduo é $\mathbf{X} = \mathbf{x}$, $g(\mathbf{x})$ é nossa predição sobre Y.

Prever a altura de um filho (Y) com base na altura de seu pai ($\mathbf{X}$).
Amostra $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$



Criar uma função de predição $g(\mathbf{x})$: dado que a altura do pai de um indivíduo é $\mathbf{X} = \mathbf{x}$, $g(\mathbf{x})$ é nossa predição sobre Y.

Objetivos:

(i) construir g de modo a se obter boas predições

$$g(\mathbf{X}_{n+1}) \approx Y_{n+1}, \ldots, g(\mathbf{X}_{n+m}) \approx Y_{n+m}$$

(ii) saber quantificar o quão boa uma (função de) predição é.

Objetivos:

(i) construir g de modo a se obter boas predições

$$g(\mathbf{X}_{n+1}) \approx Y_{n+1}, \ldots, g(\mathbf{X}_{n+m}) \approx Y_{n+m}$$

(ii) saber quantificar o quão boa uma (função de) predição é.

Objetivos:

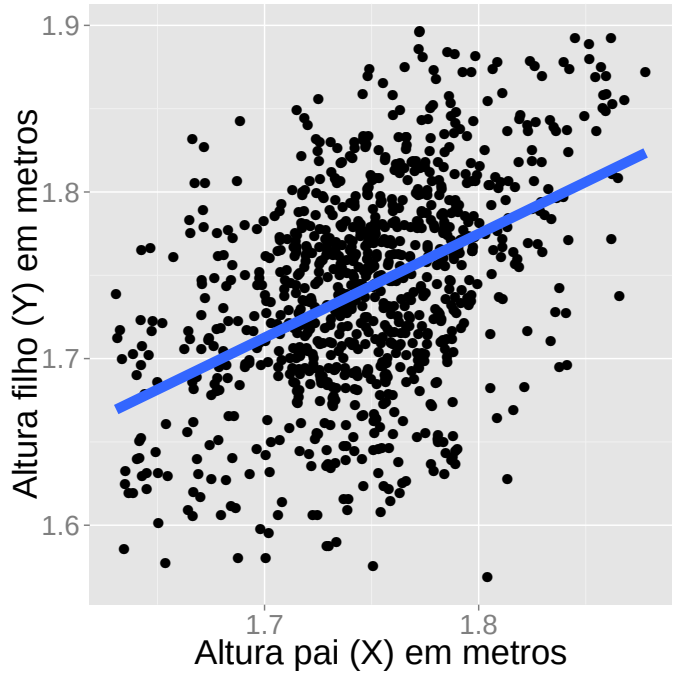
(i) construir g de modo a se obter boas predições

$$g(\mathbf{X}_{n+1}) \approx Y_{n+1}, \ldots, g(\mathbf{X}_{n+m}) \approx Y_{n+m}$$

(ii) saber quantificar o quão boa uma (função de) predição é.

Uma forma de criar g:

Uma forma de criar g: regressão linear



Predição da altura de um filho cujo pai tem $\mathbf{x} = 1,80m$:

$$g(1,80m) = 1,77m.$$

Erro quadrático: se $Y = 1,76m$, temos um erro de

$$(g(\mathbf{x}) - y)^2 = (1,76 - 1,77)^2 = 0,0001.$$

Predição da altura de um filho cujo pai tem $\mathbf{x} = 1,80m$:

$$g(1,80m) = 1,77m.$$

Erro quadrático: se $Y = 1,76m$, temos um erro de

$$(g(\mathbf{x}) - y)^2 = (1,76 - 1,77)^2 = 0,0001.$$

Assim quantificamos quão boa g é para um dado par $(\mathbf{x}, y)$.

Mais geral – função de risco:

$$R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$$

$L(g; (\mathbf{X}, Y)) = (Y - g(\mathbf{X}))^2$ é uma função de perda.

Assim quantificamos quão boa g é para um dado par $(\mathbf{x}, y)$.

Mais geral – função de risco:

$$R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$$

$L(g; (\mathbf{X}, Y)) = (Y - g(\mathbf{X}))^2$ é uma função de perda.

Assim quantificamos quão boa g é para um dado par $(\mathbf{x}, y)$.

Mais geral – função de risco:

$$R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$$

$L(g; (\mathbf{X}, Y)) = (Y - g(\mathbf{X}))^2$ é uma função de perda.

Assim quantificamos quão boa g é para um dado par $(\mathbf{x}, y)$.

Mais geral – função de risco:

$$R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$$

$L(g; (\mathbf{X}, Y)) = (Y - g(\mathbf{X}))^2$ é uma função de perda.

Suposição: Dados iid

Pela lei dos grandes números,

$$\frac{1}{m}\sum_{i=1}^{m}(Y_{n+i} - g(\mathbf{X}_{n+i}))^2 \approx \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right] := R(g)$$

Suposição: Dados iid

Pela lei dos grandes números,

$$\frac{1}{m}\sum_{i=1}^{m}(Y_{n+i} - g(\mathbf{X}_{n+i}))^2 \approx \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right] := R(g)$$

Resumindo até agora

- Observamos um conjunto de treinamento $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$.

  $\mathbf{X} \in \Re^d$ são chamados de preditores, variáveis explicativas, variáveis independentes, covariáveis ou *features*.

  Y é chamado de resposta, variável dependente ou *labels*

- Desejamos criar uma função de predição $g(\mathbf{x})$ para prever novas observações $\mathbf{X}_{n+1}, \ldots, \mathbf{X}_{n+m}$ bem

- Prever novas observações bem = criar g tal que $R(g)$ seja baixo

Resumindo até agora

- Observamos um conjunto de treinamento $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$.

  $\mathbf{X} \in \Re^d$ são chamados de preditores, variáveis explicativas, variáveis independentes, covariáveis ou *features*.

  Y é chamado de resposta, variável dependente ou *labels*

- Desejamos criar uma função de predição $g(\mathbf{x})$ para prever novas observações $\mathbf{X}_{n+1}, \ldots, \mathbf{X}_{n+m}$ bem

- Prever novas observações bem = criar g tal que $R(g)$ seja baixo

Resumindo até agora

- Observamos um conjunto de treinamento $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$.

  $\mathbf{X} \in \Re^d$ são chamados de preditores, variáveis explicativas, variáveis independentes, covariáveis ou *features*.

  Y é chamado de resposta, variável dependente ou *labels*

- Desejamos criar uma função de predição $g(\mathbf{x})$ para prever novas observações $\mathbf{X}_{n+1}, \ldots, \mathbf{X}_{n+m}$ bem

- Prever novas observações bem = criar g tal que $R(g)$ seja baixo

Qual a melhor função $g(\mathbf{x})$?

$r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$: função de regressão (não assumimos linearidade)

$R(r) \leq R(g)$ para toda função $g(\mathbf{x})$

Qual a melhor função $g(\mathbf{x})$?

$r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$: função de regressão (não assumimos linearidade)

$R(r) \leq R(g)$ para toda função $g(\mathbf{x})$

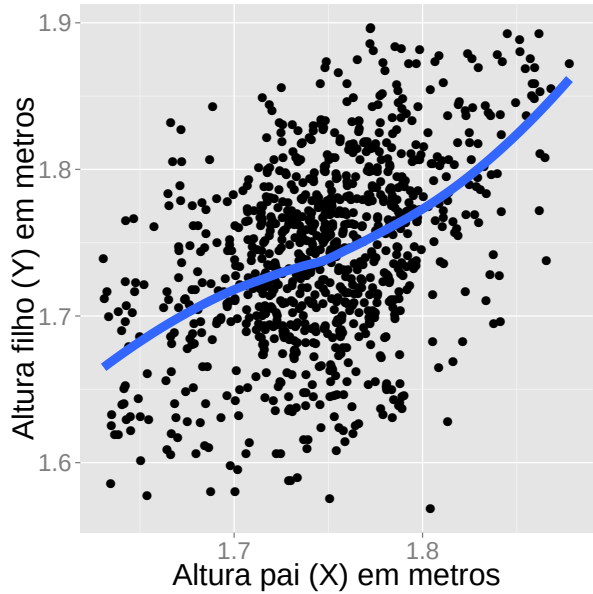
O problema está resolvido?

$r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ não conhecido!

O problema está resolvido?

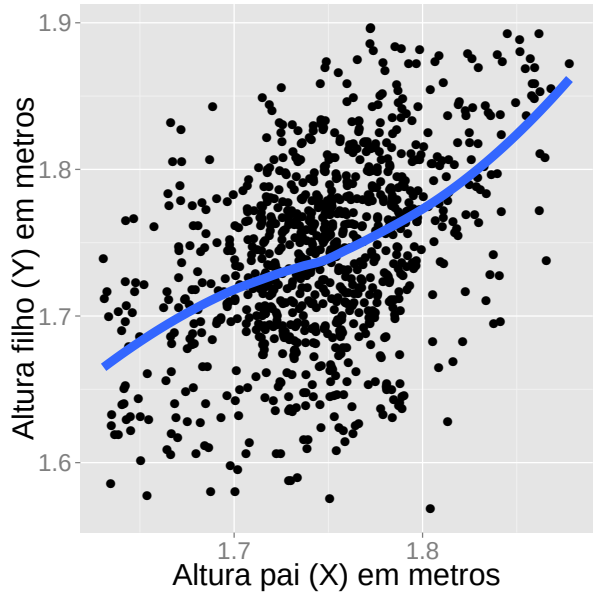
$r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ não conhecido!

$\mathbb{E}[Y|\mathbf{x}]$ não precisa ser linear



Vamos entender mais a fundo os elementos de um problema de predição.

$\mathbb{E}[Y|\mathbf{x}]$ não precisa ser linear



Vamos entender mais a fundo os elementos de um problema de predição.

Notação

Resposta	Covariáveis			
Y_1	$X_{1,1}$	$\dots$	$X_{1,d}$	$(= \mathbf{X}_1)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	
Y_n	$X_{n,1}$	$\dots$	$X_{n,d}$	$(= \mathbf{X}_n)$

Objetivo: estimar $r(\mathbf{x}) := \mathbb{E}[Y|\mathbf{x}]$

$x_{i,j}$: valor da j-ésima covariável no i-ésimo indivíduo.

Notação

Resposta	Covariáveis			
Y_1	$X_{1,1}$	$\dots$	$X_{1,d}$	$(= \mathbf{X}_1)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	
Y_n	$X_{n,1}$	$\dots$	$X_{n,d}$	$(= \mathbf{X}_n)$

Objetivo: estimar $r(\mathbf{x}) := \mathbb{E}[Y|\mathbf{x}]$

$x_{i,j}$: valor da j-ésima covariável no i-ésimo indivíduo.

Regressão linear sob um ponto de vista preditivo

Assume que $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ possui uma forma linear:

$$r(\mathbf{x}) = \beta^t \mathbf{x},$$

com $\beta = (\beta_0, \ldots, \beta_d)$ e $\mathbf{x} = (1, x_1, \ldots, x_d)$

Regressão linear sob um ponto de vista preditivo

Assume que $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ possui uma forma linear:

$$r(\mathbf{x}) = \beta^t \mathbf{x},$$

com $\beta = (\beta_0, \ldots, \beta_d)$ e $\mathbf{x} = (1, x_1, \ldots, x_d)$

Regressão linear sob um ponto de vista preditivo

Estimador usal de β: estimador de mínimos quadrados

Minimiza

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta^t \mathbf{x}_i)^2$$

$$\widehat{\beta} = (X^t X)^{-1} X^t Y.$$

Uma estimativa para $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ é

$$\widehat{r}(\mathbf{x}) = \widehat{\beta}^t \mathbf{x}.$$

Regressão linear sob um ponto de vista preditivo

Estimador usal de β: estimador de mínimos quadrados

Minimiza

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta^t \mathbf{x}_i)^2$$

$$\widehat{\beta} = (X^t X)^{-1} X^t Y.$$

Uma estimativa para $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ é

$$\widehat{r}(\mathbf{x}) = \widehat{\beta}^t \mathbf{x}.$$

Regressão linear sob um ponto de vista preditivo

Estimador usal de β: estimador de mínimos quadrados

Minimiza

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta^t \mathbf{x}_i)^2$$

$$\widehat{\beta} = (X^t X)^{-1} X^t Y.$$

Uma estimativa para $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ é

$$\widehat{r}(\mathbf{x}) = \widehat{\beta}^t \mathbf{x}.$$

Predição versus Inferência:

Inferência: assume que o modelo é correto. Principal objetivo: interpretação dos parâmetros.

- Quais parâmetros são significantes?
- Qual o efeito do aumento da dose do remédio no medicamento?

Predição: criar $g(\mathbf{x})$ com bom poder preditivo. Não assume que a verdadeira regressão é linear! Interpretação não é o foco.

Predição versus Inferência:

Inferência: assume que o modelo é correto. Principal objetivo: interpretação dos parâmetros.

- Quais parâmetros são significantes?
- Qual o efeito do aumento da dose do remédio no medicamento?

Predição: criar $g(\mathbf{x})$ com bom poder preditivo. Não assume que a verdadeira regressão é linear! Interpretação não é o foco.

Predição versus Inferência:

Inferência: assume que o modelo é correto. Principal objetivo: interpretação dos parâmetros.

- Quais parâmetros são significantes?
- Qual o efeito do aumento da dose do remédio no medicamento?

Predição: criar $g(\mathbf{x})$ com bom poder preditivo. Não assume que a verdadeira regressão é linear! Interpretação não é o foco.

Predição versus Inferência:

Inferência: assume que o modelo é correto. Principal objetivo: interpretação dos parâmetros.

- Quais parâmetros são significantes?
- Qual o efeito do aumento da dose do remédio no medicamento?

Predição: criar $g(\mathbf{x})$ com bom poder preditivo. Não assume que a verdadeira regressão é linear! Interpretação não é o foco.

As duas culturas

L. Breiman: Statistical modeling: The two cultures. Statistical Science, 16(3):199–231, 2001.

Duas culturas no uso de modelos estatísticos:

- Data Modeling Culture:

- Algorithmic Modeling Culture:

As duas culturas

L. Breiman: Statistical modeling: The two cultures. Statistical Science, 16(3):199–231, 2001.

Duas culturas no uso de modelos estatísticos:

- Data Modeling Culture:

- Algorithmic Modeling Culture:

As duas culturas

L. Breiman: Statistical modeling: The two cultures. Statistical Science, 16(3):199–231, 2001.

Duas culturas no uso de modelos estatísticos:

- Data Modeling Culture: Domina a comunidade estatística. Se assume que o modelo é correto. Testar suposições é fundamental. Foco em inferência.
- Algorithmic Modeling Culture:

As duas culturas

L. Breiman: Statistical modeling: The two cultures. Statistical Science, 16(3):199–231, 2001.

Duas culturas no uso de modelos estatísticos:

- Data Modeling Culture: Domina a comunidade estatística. Se assume que o modelo é correto. Testar suposições é fundamental. Foco em inferência.
- Algorithmic Modeling Culture: Domina a comunidade de machine learning. **Não** se assume que o modelo utilizado é correto; o modelo é utilizado para criar bons algoritmos preditivos.

"Oddly, we are in a period where there has never been such a wealth of new statistical problems and sources of data. The danger is that if we define the boundaries of our field in terms of familar tools and familar problems, we will fail to grasp the new opportunities". (Breiman, 2001)

O Método de Mínimos Quadrados é bom?

✓Mínimos Quadrados=Máxima Verossimilhança sob normalidade, linearidade e homoscedasticidade, logo **consistente sob essas suposições**.

✓**BLUE** sob linearidade e homoscedasticidade.

× Não queremos assumir tudo isso!!

✓ É possível dar garantias sem essas suposições!

O Método de Mínimos Quadrados é bom?

✓Mínimos Quadrados=Máxima Verossimilhança sob normalidade, linearidade e homoscedasticidade, logo **consistente sob essas suposições**.

✓**BLUE** sob linearidade e homoscedasticidade.

× Não queremos assumir tudo isso!!

✓ É possível dar garantias sem essas suposições!

O Método de Mínimos Quadrados é bom?

✓Mínimos Quadrados=Máxima Verossimilhança sob normalidade, linearidade e homoscedasticidade, logo **consistente sob essas suposições**.

✓**BLUE** sob linearidade e homoscedasticidade.

× Não queremos assumir tudo isso!!

✓ É possível dar garantias sem essas suposições!

O Método de Mínimos Quadrados é bom?

✓Mínimos Quadrados=Máxima Verossimilhança sob normalidade, linearidade e homoscedasticidade, logo **consistente sob essas suposições**.

✓**BLUE** sob linearidade e homoscedasticidade.

× Não queremos assumir tudo isso!!

✓ É possível dar garantias sem essas suposições!

O Método de Mínimos Quadrados é bom?

✓Mínimos Quadrados=Máxima Verossimilhança sob normalidade, linearidade e homoscedasticidade, logo **consistente sob essas suposições**.

✓**BLUE** sob linearidade e homoscedasticidade.

× Não queremos assumir tudo isso!!

✓ É possível dar garantias sem essas suposições!

Melhor preditor linear (oráculo):

$$\boldsymbol{\beta}_* = \arg\min_{\boldsymbol{\beta}} R(g_{\boldsymbol{\beta}}),$$

onde $g_{\boldsymbol{\beta}}(\mathbf{x}) = \boldsymbol{\beta}^t \mathbf{x}$.

Teorema:

$$\widehat{\boldsymbol{\beta}} \xrightarrow[n \longrightarrow \infty]{P} \boldsymbol{\beta}_* \text{ e } R(g_{\widehat{\boldsymbol{\beta}}}) \xrightarrow[n \longrightarrow \infty]{P} R(g_{\boldsymbol{\beta}_*})$$

Melhor preditor linear (oráculo):

$$\boldsymbol{\beta}_* = \arg\min_{\boldsymbol{\beta}} R(g_{\boldsymbol{\beta}}),$$

onde $g_{\boldsymbol{\beta}}(\mathbf{x}) = \boldsymbol{\beta}^t \mathbf{x}$.

Teorema:

$$\widehat{\boldsymbol{\beta}} \xrightarrow[n \longrightarrow \infty]{P} \boldsymbol{\beta}_* \text{ e } R(g_{\widehat{\boldsymbol{\beta}}}) \xrightarrow[n \longrightarrow \infty]{P} R(g_{\boldsymbol{\beta}_*})$$

Decomposição Viés-Variância; Overfitting e Underfitting

Decomposição Viés-Variância do risco esperado

$$\mathbb{E}\left[(Y - g(\mathbf{X}))^2 \middle| \mathbf{X} = \mathbf{x}\right]$$

- $\mathbb{V}[Y|\mathbf{X} = \mathbf{x}]$ é a variância intrínseca da variável resposta
- $(r(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})])^2$ é o viés ao quadrado de g
- $\mathbb{V}[g(\mathbf{x})]$ é sua variância

Decomposição Viés-Variância do risco esperado

$$\mathbb{E}\left[(Y - g(\mathbf{X}))^2 \middle| \mathbf{X} = \mathbf{x}\right] = \mathbb{V}[Y|\mathbf{X} = \mathbf{x}] + (r(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})])^2 + \mathbb{V}[g(\mathbf{x})]$$

- $\mathbb{V}[Y|\mathbf{X} = \mathbf{x}]$ é a variância intrínseca da variável resposta
- $(r(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})])^2$ é o viés ao quadrado de g
- $\mathbb{V}[g(\mathbf{x})]$ é sua variância

Decomposição Viés-Variância do risco esperado

$$\mathbb{E}\left[(Y - g(\mathbf{X}))^2 \middle| \mathbf{X} = \mathbf{x}\right] = \mathbb{V}[Y|\mathbf{X} = \mathbf{x}] + (r(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})])^2 + \mathbb{V}[g(\mathbf{x})]$$

- $\mathbb{V}[Y|\mathbf{X} = \mathbf{x}]$ é a variância intrínseca da variável resposta
- $(r(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})])^2$ é o viés ao quadrado de g
- $\mathbb{V}[g(\mathbf{x})]$ é sua variância

Decomposição Viés-Variância do risco esperado

$$\mathbb{E}\left[(Y - g(\mathbf{X}))^2 \middle| \mathbf{X} = \mathbf{x}\right] = \mathbb{V}[Y|\mathbf{X} = \mathbf{x}] + (r(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})])^2 + \mathbb{V}[g(\mathbf{x})]$$

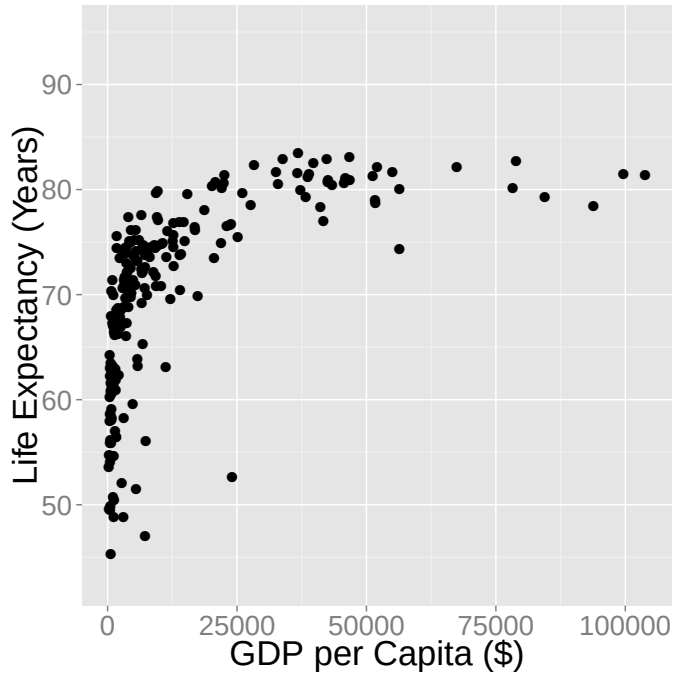
- $\mathbb{V}[Y|\mathbf{X} = \mathbf{x}]$ é a variância intrínseca da variável resposta
- $(r(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})])^2$ é o viés ao quadrado de g
- $\mathbb{V}[g(\mathbf{x})]$ é sua variância

Decomposição Viés-Variância do risco esperado

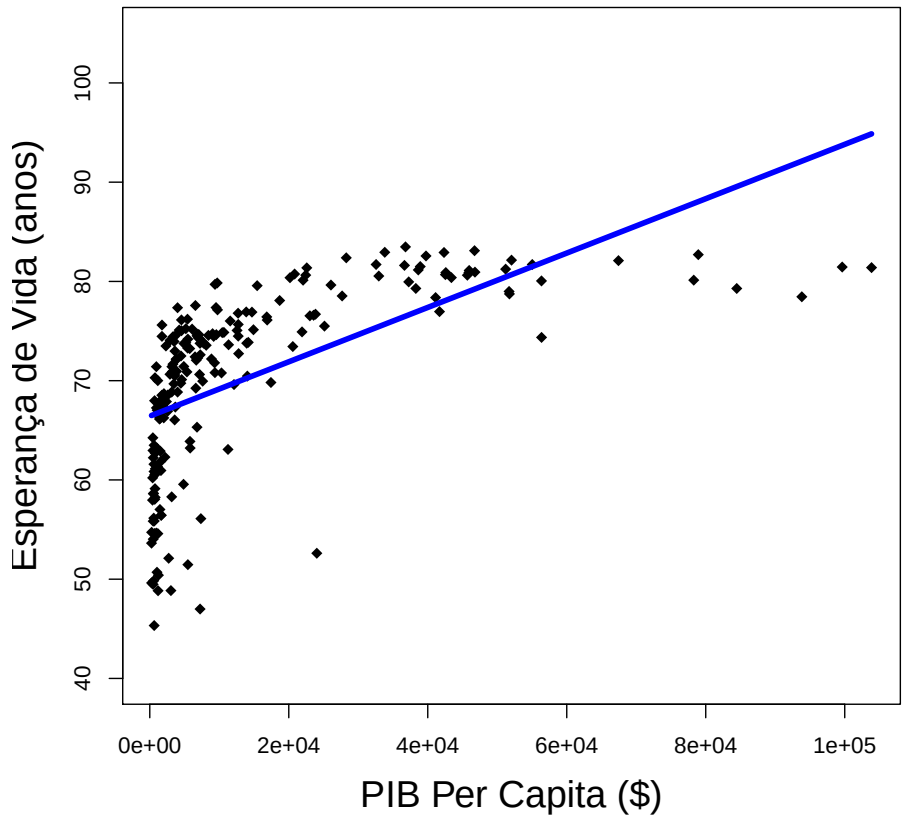
$$\mathbb{E}\left[(Y - g(\mathbf{X}))^2 \middle| \mathbf{X} = \mathbf{x}\right] = \mathbb{V}[Y|\mathbf{X} = \mathbf{x}] + (r(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})])^2 + \mathbb{V}[g(\mathbf{x})]$$

- $\mathbb{V}[Y|\mathbf{X} = \mathbf{x}]$ é a variância intrínseca da variável resposta
- $(r(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})])^2$ é o viés ao quadrado de g
- $\mathbb{V}[g(\mathbf{x})]$ é sua variância

Exemplo de Regressão Linear



Exemplo de Regressão Linear



$$g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

Viés alto.

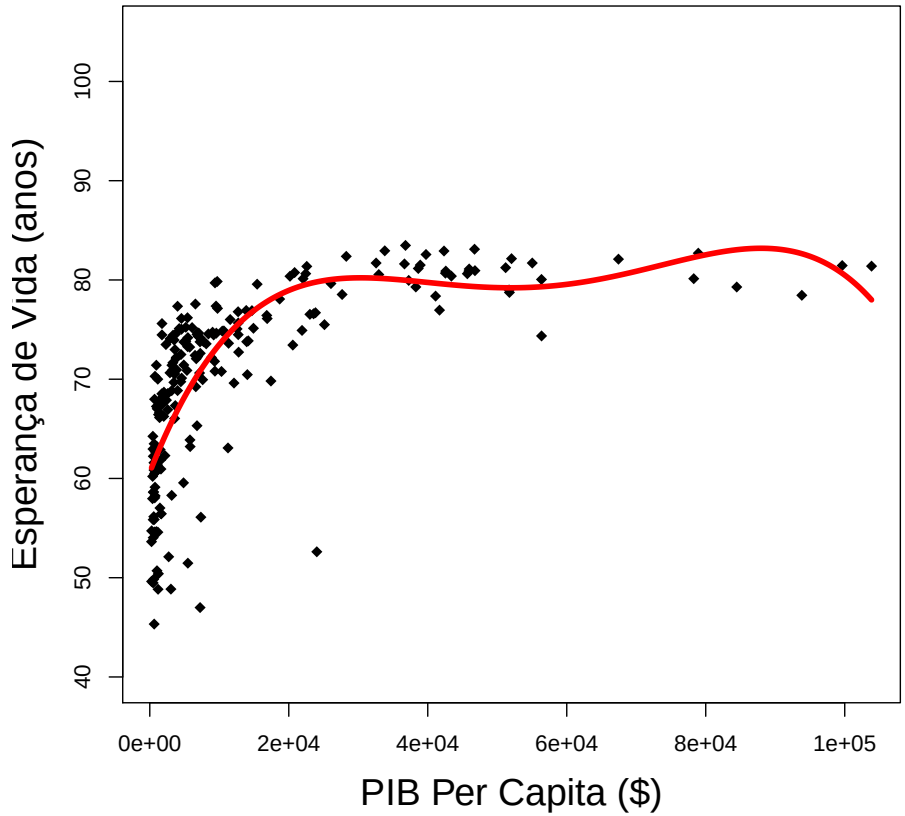
Podemos usar a metodologia de regressão linear para ajustar polinômios:

$$g(x) = \overbrace{\widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2 + \widehat{\beta}_3 x^3 + \widehat{\beta}_4 x^4}^{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \widehat{\beta}_4 x_4}$$

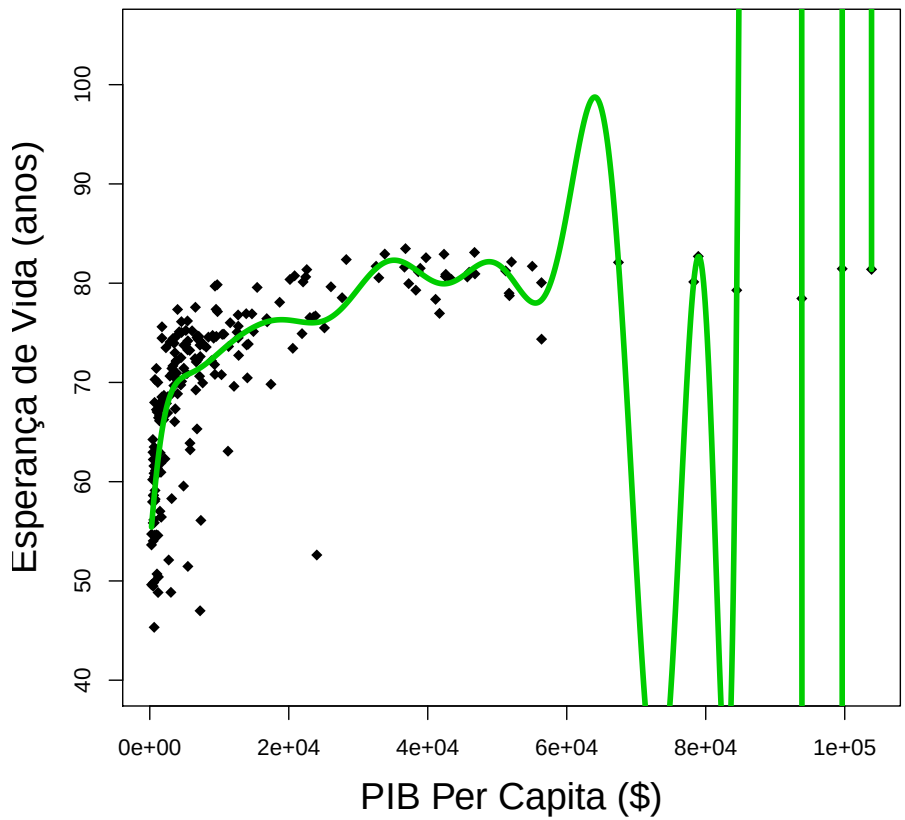
Viés alto.

Podemos usar a metodologia de regressão linear para ajustar polinômios:

$$g(x) = \overbrace{\widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2 + \widehat{\beta}_3 x^3 + \widehat{\beta}_4 x^4}^{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \widehat{\beta}_4 x_4}$$

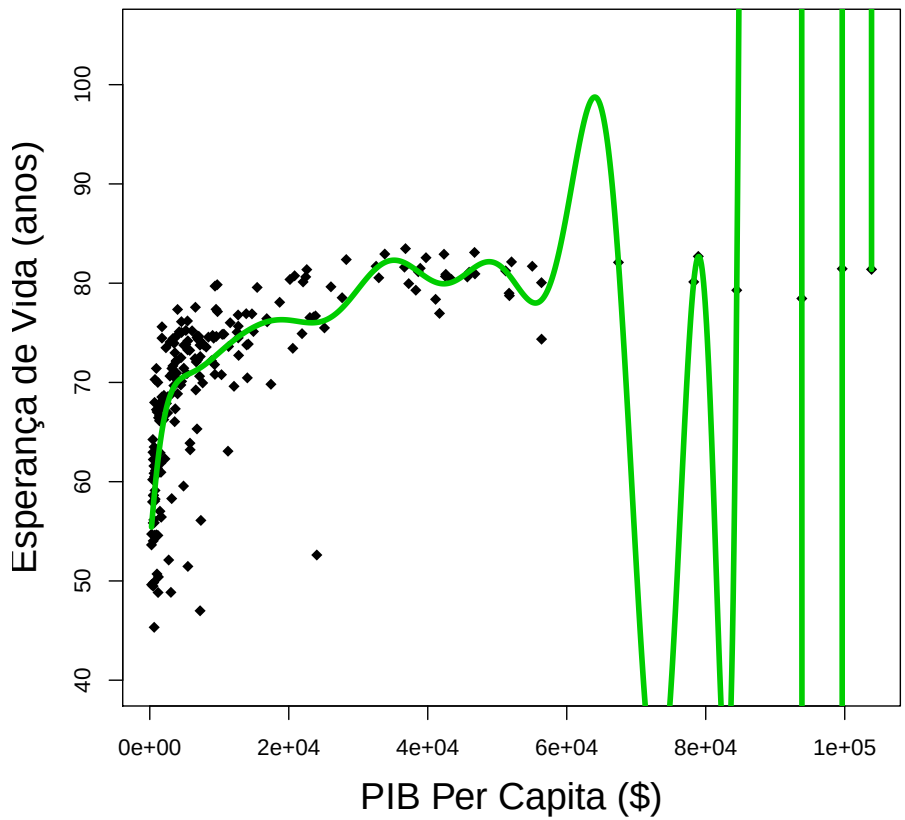


$$g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2 + \widehat{\beta}_3 x^3 + \widehat{\beta}_4 x^4$$



$$g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2 + \ldots + \widehat{\beta}_{50} x^{50}$$

Variância alta.



$$g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2 + \ldots + \widehat{\beta}_{50} x^{50}$$

Variância alta.

Seleção de Modelos: Super e Sub-Ajuste

Seleção de Modelos: Super e Sub-Ajuste

Encontrar a melhor função de predição em

$$\mathbb{G} = \left\{ g(x) : g(x) = \beta_0 + \sum_{i=1}^{p} \beta_i x^i, \text{ para } p \in \{1, 2, \ldots, 50\} \right\}$$

Qual o melhor p?

$p = 50$: super-ajuste $\Rightarrow$ baixo poder preditivo.

$p = 1$: sub-ajuste $\Rightarrow$ baixo poder preditivo.

Como formalizar/resolver isso?

Seleção de Modelos: Super e Sub-Ajuste

Encontrar a melhor função de predição em

$$\mathbb{G} = \left\{ g(x) : g(x) = \beta_0 + \sum_{i=1}^{p} \beta_i x^i, \text{ para } p \in \{1, 2, \ldots, 50\} \right\}$$

Qual o melhor p?

$p = 50$: super-ajuste $\Rightarrow$ baixo poder preditivo.

$p = 1$: sub-ajuste $\Rightarrow$ baixo poder preditivo.

Como formalizar/resolver isso?

Seleção de Modelos: Super e Sub-Ajuste

Encontrar a melhor função de predição em

$$\mathbb{G} = \left\{ g(x) : g(x) = \beta_0 + \sum_{i=1}^{p} \beta_i x^i, \text{ para } p \in \{1, 2, \ldots, 50\} \right\}$$

Qual o melhor p?

$p = 50$: super-ajuste $\Rightarrow$ baixo poder preditivo.

$p = 1$: sub-ajuste $\Rightarrow$ baixo poder preditivo.

Como formalizar/resolver isso?

Seleção de Modelos: Super e Sub-Ajuste

Encontrar a melhor função de predição em

$$\mathbb{G} = \left\{ g(x) : g(x) = \beta_0 + \sum_{i=1}^{p} \beta_i x^i, \text{ para } p \in \{1, 2, \ldots, 50\} \right\}$$

Qual o melhor p?

$p = 50$: super-ajuste $\Rightarrow$ baixo poder preditivo.

$p = 1$: sub-ajuste $\Rightarrow$ baixo poder preditivo.

Como formalizar/resolver isso?

Seleção de Modelos: Super e Sub-Ajuste

Encontrar a melhor função de predição em

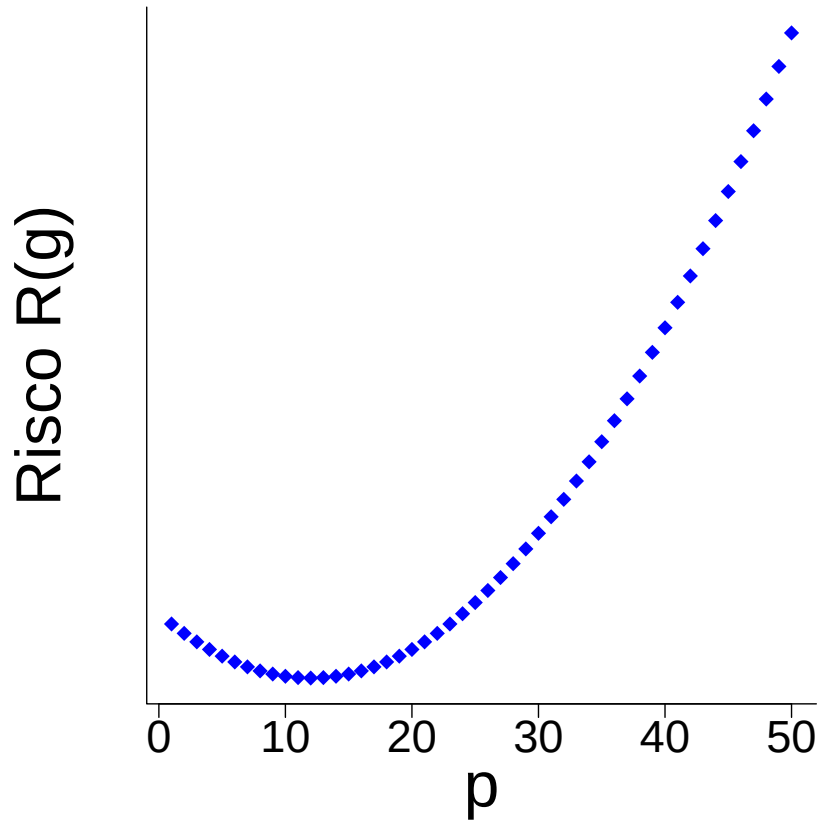
$$\mathbb{G} = \left\{ g(x) : g(x) = \beta_0 + \sum_{i=1}^{p} \beta_i x^i, \text{ para } p \in \{1, 2, \ldots, 50\} \right\}$$

Qual o melhor p?

$p = 50$: super-ajuste $\Rightarrow$ baixo poder preditivo.

$p = 1$: sub-ajuste $\Rightarrow$ baixo poder preditivo.

Como formalizar/resolver isso?


Risco R(g)
0
10
20
30
40
50
p

Como estimar $R(g)$?

Como estimar $R(g)$?

Estimar $R(g)$ é importante para comparar diferentes candidatos $g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots$.

Como estimar o risco $R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$?

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(\mathbf{X}_i))^2 := EQM(g) \ ?$$

Como estimar $R(g)$?

Estimar $R(g)$ é importante para comparar diferentes candidatos $g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots$.

Como estimar o risco $R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$?

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(\mathbf{X}_i))^2 := EQM(g) \ ?$$

Como estimar $R(g)$?

Estimar $R(g)$ é importante para comparar diferentes candidatos $g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots$.

Como estimar o risco $R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$?

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(\mathbf{X}_i))^2 := EQM(g) \ ?$$

Como estimar $R(g)$?

Estimar $R(g)$ é importante para comparar diferentes candidatos $g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots$.

Como estimar o risco $R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$?

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(\mathbf{X}_i))^2 := EQM(g) \ ?$$

Não!!!

Como estimar $R(g)$?

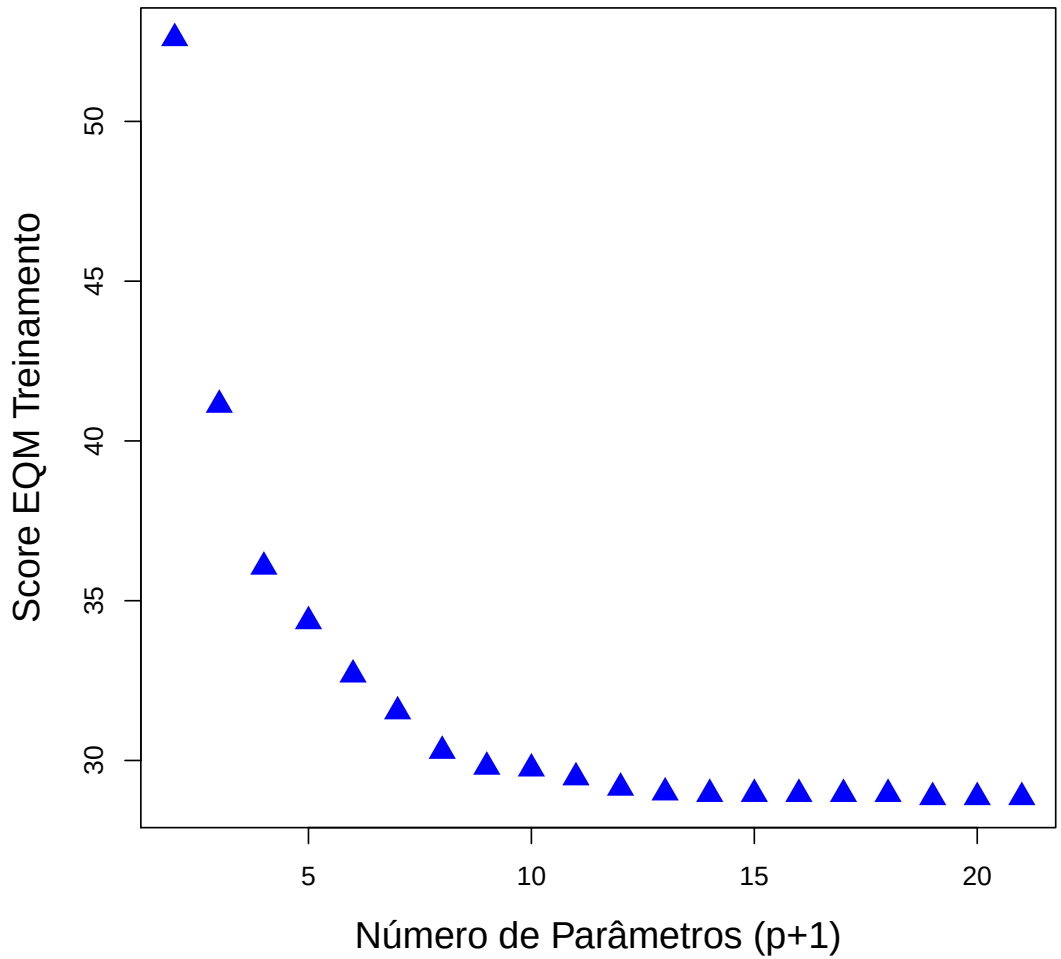
O erro quadrático médio avaliado no conjunto de treinamento em geral é um muito otimista

Leva ao super-ajuste

Como estimar $R(g)$?

O erro quadrático médio avaliado no conjunto de treinamento em geral é um muito otimista

Leva ao super-ajuste


Score EQM Treinamento
30
35
40
45
50
5
10
15
20
Número de Parâmetros (p+1)

Data-splitting

Data-splitting

$$\overbrace{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_s, Y_s)}^{\text{Treinamento (e.g., 70\%)}}, \overbrace{(\mathbf{X}_{s+1}, Y_{s+1}), \ldots, (\mathbf{X}_n, Y_n)}^{\text{Validação (e.g., 30\%)}}$$

Treinamento: estimar g

Validação: estimar $R(g)$

$$R(g) \approx \frac{1}{n-s} \sum_{i=s+1}^{n} (Y_i - g(\mathbf{X}_i))^2$$

Consistente pela lei dos grandes números.

Data-splitting

$$\overbrace{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_s, Y_s)}^{\text{Treinamento (e.g., 70\%)}}, \overbrace{(\mathbf{X}_{s+1}, Y_{s+1}), \ldots, (\mathbf{X}_n, Y_n)}^{\text{Validação (e.g., 30\%)}}$$

Treinamento: estimar g

Validação: estimar $R(g)$

$$R(g) \approx \frac{1}{n-s} \sum_{i=s+1}^{n} (Y_i - g(\mathbf{X}_i))^2$$

Consistente pela lei dos grandes números.

Data-splitting

$$\overbrace{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_s, Y_s)}^{\text{Treinamento (e.g., 70\%)}}, \overbrace{(\mathbf{X}_{s+1}, Y_{s+1}), \ldots, (\mathbf{X}_n, Y_n)}^{\text{Validação (e.g., 30\%)}}$$

Treinamento: estimar g

Validação: estimar $R(g)$

$$R(g) \approx \frac{1}{n-s} \sum_{i=s+1}^{n} (Y_i - g(\mathbf{X}_i))^2$$

Consistente pela lei dos grandes números.

Data-splitting

$$\overbrace{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_s, Y_s)}^{\text{Treinamento (e.g., 70\%)}}, \overbrace{(\mathbf{X}_{s+1}, Y_{s+1}), \ldots, (\mathbf{X}_n, Y_n)}^{\text{Validação (e.g., 30\%)}}$$

Treinamento: estimar g

Validação: estimar $R(g)$

$$R(g) \approx \frac{1}{n-s} \sum_{i=s+1}^{n} (Y_i - g(\mathbf{X}_i))^2$$

Consistente pela lei dos grandes números.

Em geral,

$$R(g) \approx \frac{1}{n-s} \sum_{i=s+1}^{n} L(g; (\mathbf{X}_i, Y_i)) := \widehat{R}(g).$$

Validação Cruzada

Quando o tamanho amostral é pequeno: validação cruzada

$$R(g) \approx \frac{1}{n}\sum_{i=1}^{n}(Y_i - g_{-i}(X_i))^2$$

onde g_{-i} é ajustado usando-se todas as observações exceto a i-ésima delas

Validação Cruzada

Quando o tamanho amostral é pequeno: validação cruzada

$$R(g) \approx \frac{1}{n} \sum_{i=1}^{n} (Y_i - g_{-i}(X_i))^2$$

onde g_{-i} é ajustado usando-se todas as observações exceto a i-ésima delas

Validação Cruzada

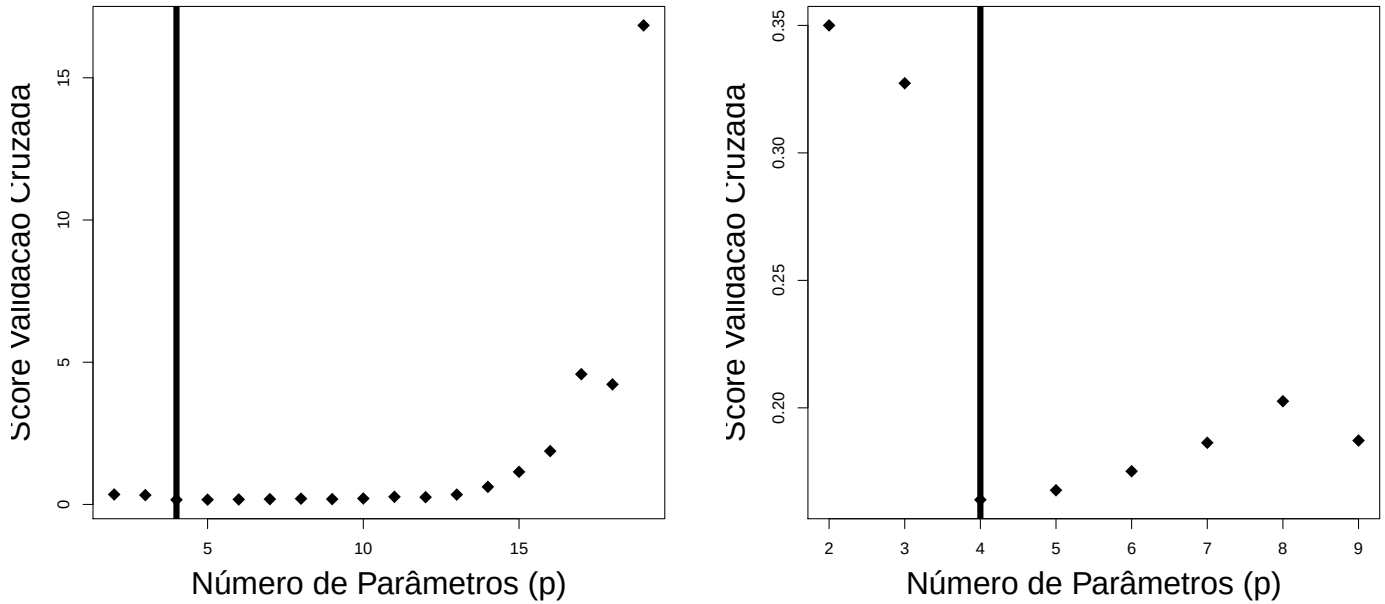
Quando o tamanho amostral é pequeno: validação cruzada

$$R(g) \approx \frac{1}{n} \sum_{i=1}^{n} (Y_i - g_{-i}(X_i))^2$$

onde g_{-i} é ajustado usando-se todas as observações exceto a i-ésima delas

Validação Cruzada

Exemplo.

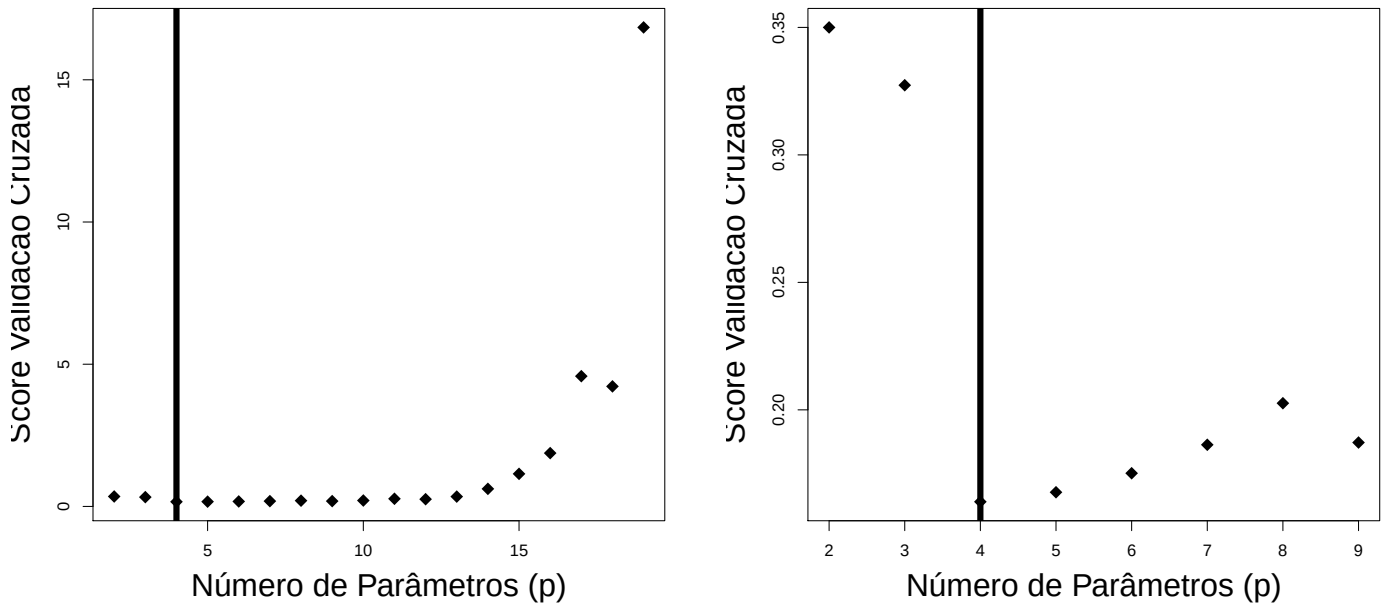


$$g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \ldots + \widehat{\beta}_{p-1} x^{p-1}$$

Melhor modelo: $p = 4$

Validação Cruzada

Exemplo.



$$g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \ldots + \widehat{\beta}_{p-1} x^{p-1}$$

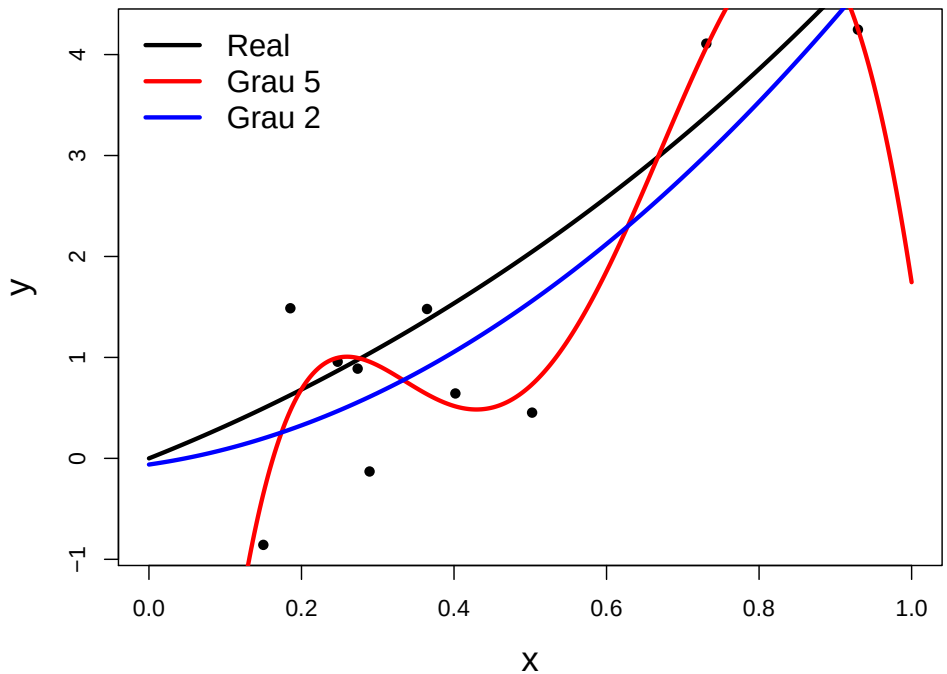
Melhor modelo: $p = 4$

Interlúdio: a especificação do modelo estar correta garante melhor poder preditivo?

Aplicativo Shiny

Interlúdio: a especificação do modelo estar correta garante melhor poder preditivo?

Aplicativo Shiny



$\beta = (0, 3, 2, 0.2, 0.1, 0.1)$

Resumindo até aqui

- Avaliação da qualidade preditiva de g:

$$R(g) = \mathbb{E}[(Y - g(\mathbf{X}))^2]$$

- Melhor função de predição: $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}]$

- g muito complexa: overfitting
  g muito simples: overfitting

- Na prática: construir g's via conjunto de treinamento;
  encontrar melhor estimando $R(g)$ via conjunto de validação;

Resumindo até aqui

- Avaliação da qualidade preditiva de g:

$$R(g) = \mathbb{E}[(Y - g(\mathbf{X}))^2]$$

- Melhor função de predição: $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}]$

- g muito complexa: overfitting
  g muito simples: overfitting

- Na prática: construir g's via conjunto de treinamento;
  encontrar melhor estimando $R(g)$ via conjunto de validação;

Resumindo até aqui

- Avaliação da qualidade preditiva de g:

$$R(g) = \mathbb{E}[(Y - g(\mathbf{X}))^2]$$

- Melhor função de predição: $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}]$

- g muito complexa: overfitting
  g muito simples: overfitting

- Na prática: construir g's via conjunto de treinamento;
  encontrar melhor estimando $R(g)$ via conjunto de validação;

Resumindo até aqui

- Avaliação da qualidade preditiva de g:

$$R(g) = \mathbb{E}[(Y - g(\mathbf{X}))^2]$$

- Melhor função de predição: $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}]$

- g muito complexa: overfitting
  g muito simples: overfitting

- Na prática: construir g's via conjunto de treinamento;
  encontrar melhor estimando $R(g)$ via conjunto de validação;

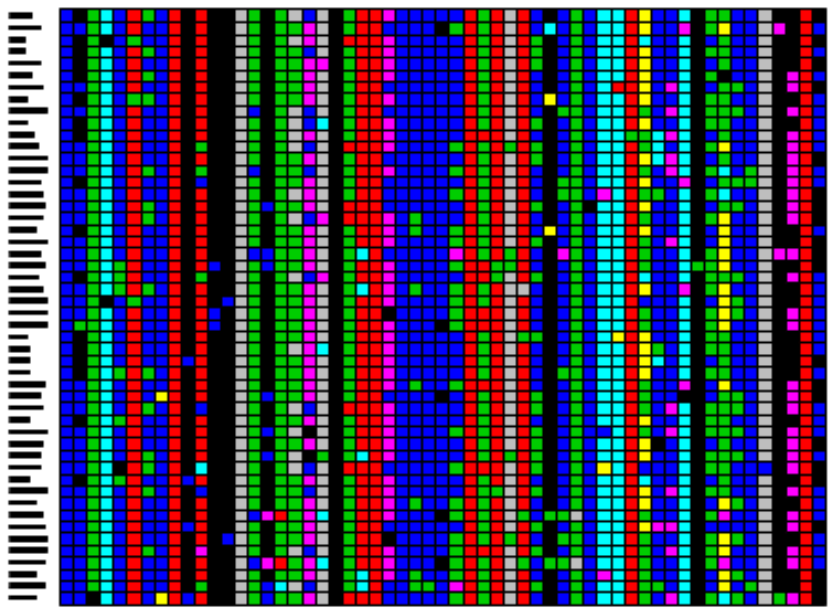
Seleção de variáveis: lasso

Exemplo

$Y =$ HIV resistance

X_j = amino acid in position j of the virus.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{100} X_{100} + \epsilon$$



Porque não usar todas as covariáveis?

(1) Várias variáveis importam pouco

(2) Muitos coeficientes para estimar

risco preditivo $=$ viés^2 $+$ variância $+$ erro inevitável

Viés: erro por omitir variáveis importantes

Variância: error por ter que estimar muitos parâmetros

Porque não usar todas as covariáveis?

(1) Várias variáveis importam pouco

(2) Muitos coeficientes para estimar

risco preditivo $=$ viés^2 $+$ variância $+$ erro inevitável

Viés: erro por omitir variáveis importantes

Variância: error por ter que estimar muitos parâmetros

Porque não usar todas as covariáveis?

(1) Várias variáveis importam pouco

(2) Muitos coeficientes para estimar

risco preditivo = viés^2 + variância + erro inevitável

Viés: erro por omitir variáveis importantes

Variância: error por ter que estimar muitos parâmetros

Porque não usar todas as covariáveis?

(1) Várias variáveis importam pouco

(2) Muitos coeficientes para estimar

risco preditivo = viés^2 + variância + erro inevitável

Viés: erro por omitir variáveis importantes

Variância: error por ter que estimar muitos parâmetros

Porque não usar todas as covariáveis?

(1) Várias variáveis importam pouco

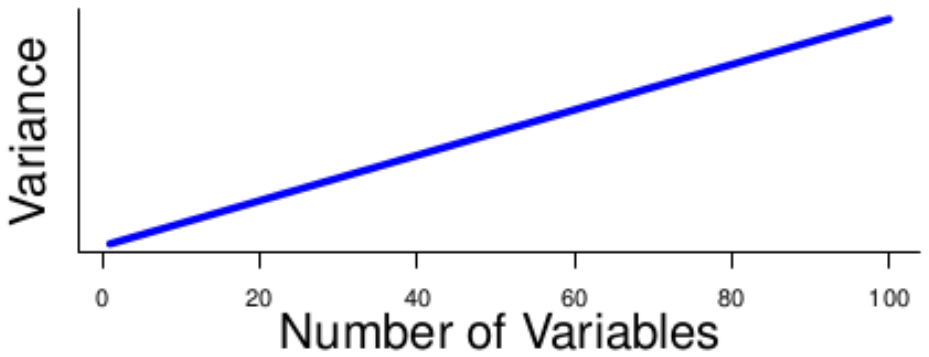
(2) Muitos coeficientes para estimar

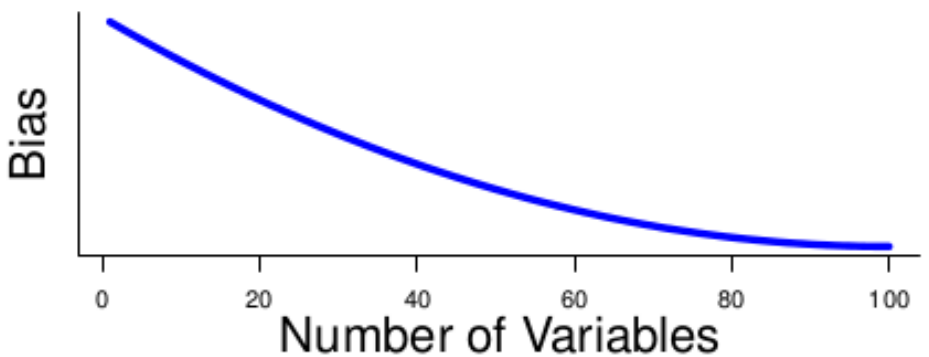
risco preditivo = viés^2 + variância + erro inevitável

Viés: erro por omitir variáveis importantes

Variância: error por ter que estimar muitos parâmetros

The Bias-Variance Tradeoff





Podemos buscar o melhor modelo dentre

$$
\begin{aligned}
\mathbb{G} = \{ & g(x) = \widehat{\beta}_0, \\
& g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1, \\
& g(x) = \widehat{\beta}_0 + \widehat{\beta}_2 x_2, \\
& \ldots \\
& g(x) = \widehat{\beta}_0 + \widehat{\beta}_d x_d, \\
& g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2, \\
& g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_3 x_3, \\
& \ldots \\
& g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \ldots + \widehat{\beta}_d x_d \}
\end{aligned}
$$

Queremos então selecionar um dentre todos os modelos disponíveis

Há 2^d modelos!!

Se $d = 30$, são 1.073.741.824 modelos!!

Se $d = 100$, são mais modelos que átomos no universo!!

Infactível.

Uma solução: stepwise. Reduz para $\approx d^2$

Queremos então selecionar um dentre todos os modelos disponíveis

Há 2^d modelos!!

Se $d = 30$, são 1.073.741.824 modelos!!

Se $d = 100$, são mais modelos que átomos no universo!!

Infactível.

Uma solução: stepwise. Reduz para $\approx d^2$

Queremos então selecionar um dentre todos os modelos disponíveis

Há 2^d modelos!!

Se $d = 30$, são 1.073.741.824 modelos!!

Se $d = 100$, são mais modelos que átomos no universo!!

Infactível.

Uma solução: stepwise. Reduz para $\approx d^2$

Queremos então selecionar um dentre todos os modelos disponíveis

Há 2^d modelos!!

Se $d = 30$, são 1.073.741.824 modelos!!

Se $d = 100$, são mais modelos que átomos no universo!!

Infactível.

Uma solução: stepwise. Reduz para $\approx d^2$

Queremos então selecionar um dentre todos os modelos disponíveis

Há 2^d modelos!!

Se $d = 30$, são 1.073.741.824 modelos!!

Se $d = 100$, são mais modelos que átomos no universo!!

Infactível.

Uma solução: stepwise. Reduz para $\approx d^2$

Queremos então selecionar um dentre todos os modelos disponíveis

Há 2^d modelos!!

Se $d = 30$, são 1.073.741.824 modelos!!

Se $d = 100$, são mais modelos que átomos no universo!!

Infactível.

Uma solução: stepwise. Reduz para $\approx d^2$

Lasso: Penalização esperta

Ideia: β_i é pequeno: covariável não contribui muito para a variância

β_i não precisa ser exatamente zero

Lasso: Penalização esperta

Ideia: β_i é pequeno: covariável não contribui muito para a variância

β_i não precisa ser exatamente zero

Lasso:

$$\arg\min_{\beta} \text{EQM}(g_\beta) \text{ sujeito a } \sum_{j=1}^{d} |\beta_j| \leq B,$$

Observações: (i) a solução do lasso é fácil de ser encontrada, (ii) ela em geral possui muitos zeros.

Equivalente:

$$\arg\min_{\beta} \text{EQM}(g_\beta) + \lambda \sum_{j=1}^{d} |\beta_j|$$

Lasso:

$$\arg\min_{\beta} \text{EQM}(g_\beta) \text{ sujeito a } \sum_{j=1}^{d} |\beta_j| \leq B,$$

Observações: (i) a solução do lasso é fácil de ser encontrada, (ii) ela em geral possui muitos zeros.

Equivalente:

$$\arg\min_{\beta} \text{EQM}(g_\beta) + \lambda \sum_{j=1}^{d} |\beta_j|$$

Lasso:

$$\arg\min_{\beta} \text{EQM}(g_\beta) \text{ sujeito a } \sum_{j=1}^{d} |\beta_j| \leq B,$$

Observações: (i) a solução do lasso é fácil de ser encontrada, (ii) ela em geral possui muitos zeros.

Equivalente:

$$\arg\min_{\beta} \text{EQM}(g_\beta) + \lambda \sum_{j=1}^{d} |\beta_j|$$

Aplicativo Shiny

Como escolher λ?

Para cada λ, buscamos

$$\beta^{\lambda} \equiv \arg\min_{\beta} \text{EQM}(g_{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|$$

Em seguida, buscamos

$$\arg\min_{\beta^{\lambda}} \widehat{R}(g_{\beta^{\lambda}})$$

Como escolher λ?

Para cada λ, buscamos

$$\beta^{\lambda} \equiv \arg\min_{\beta} \mathrm{EQM}(g_{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|$$

Em seguida, buscamos

$$\arg\min_{\beta^{\lambda}} \widehat{R}(g_{\beta^{\lambda}})$$

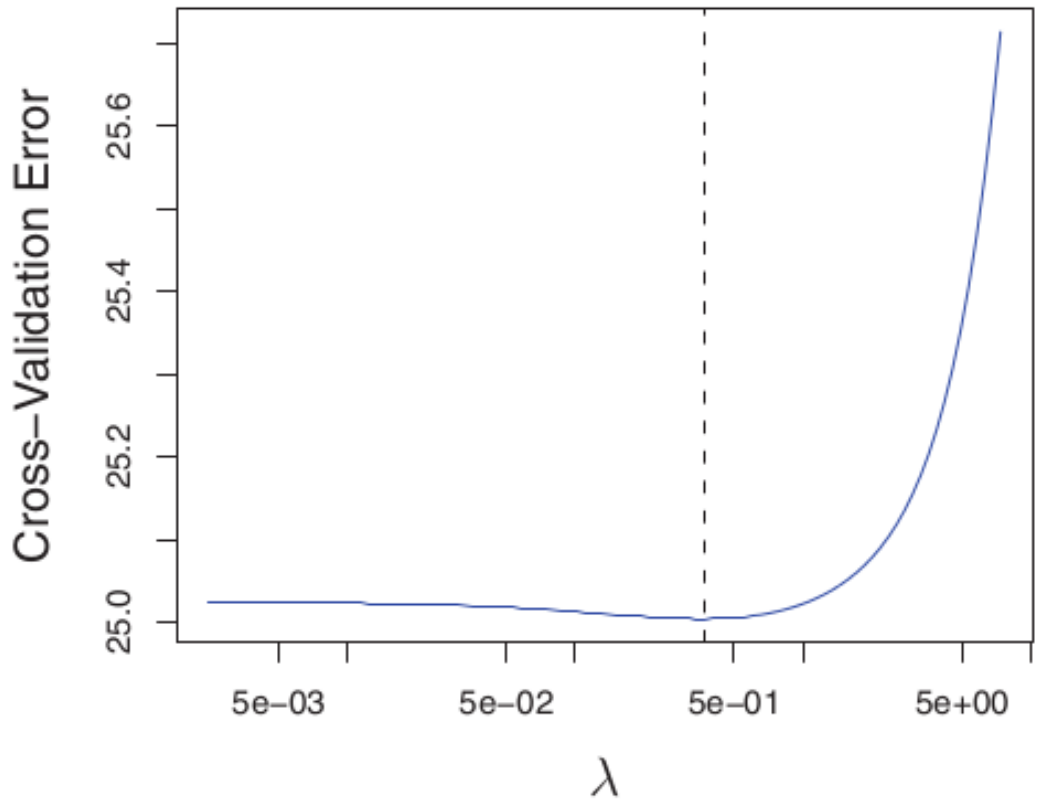
Como escolher λ?

Para cada λ, buscamos

$$\beta^{\lambda} \equiv \arg\min_{\beta} \text{EQM}(g_{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|$$

Em seguida, buscamos

$$\arg\min_{\beta^{\lambda}} \widehat{R}(g_{\beta^{\lambda}})$$


Cross−Validation Error
25.0
25.2
25.4
25.6
5e−03
5e−02
5e−01
5e+00
λ

Muitas vezes funciona melhor se as covariáveis são normalizadas

Lasso aumenta o viés e diminui a variância do EMQ.

risco preditivo $=$ viés^2 $+$ variância $+$ erro inevitável

Muitas vezes funciona melhor se as covariáveis são normalizadas

Lasso aumenta o viés e diminui a variância do EMQ.

risco preditivo $=$ viés^2 $+$ variância $+$ erro inevitável

Theorem

(Greenshtein e Ritov (2004)) Seja

$$\beta_* = \arg\min_{\beta} \mathbb{E}[(Y - \beta^t \mathbf{X})^2] \text{ sujeito a } ||\beta||_1 \leq L$$

o melhor preditor linear esparso. Se $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ *são i.i.d.'s e* $|Y|, |X_1|, \ldots, |X_n| \leq B$ *para algum* $B > 0$*, então*

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \beta^t \mathbf{X}_i)^2 \text{ sujeito a } ||\beta||_1 \leq L$$

é tal que, com probabilidade ao menos $1 - \delta$*,*

$$R(\widehat{\beta}) - R(\beta_*) = \sqrt{\frac{16(L+1)^4 B^2}{n} \log\left(\frac{\sqrt{2}d}{\sqrt{\delta}}\right)}.$$

Quanto menor o valor de L, mais próximo o risco do estimador do lasso fica do risco do oráculo. Ou seja, mais fácil é se recuperar o melhor β. Por outro lado, quanto menor o valor de L, pior é o oráculo.

Aplicação

Dados simulados: $Y = X_1 + X_2 + X_3 + X_4 + X_5 + \epsilon$.

Além de $x_1, \ldots, x_5$, observamos mais 15 variáveis não relacionadas a y.

Aplicação

Dados simulados: $Y = X_1 + X_2 + X_3 + X_4 + X_5 + \epsilon$.

Além de $x_1, \ldots, x_5$, observamos mais 15 variáveis não relacionadas a y.

Aplicação

Dados simulados: $Y = X_1 + X_2 + X_3 + X_4 + X_5 + \epsilon$.

Além de $x_1, \ldots, x_5$, observamos mais 15 variáveis não relacionadas a y.

Resultados:

Todos os subconjuntos:

Tempo: 1 hora e 20 minutos; $\widehat{R}(g^*) = 0.30$

Covariáveis selecionadas: $x_1, x_2, x_3, x_4, x_5, x_{10}, x_{12}, x_{13}, x_{19}, x_{20}$

Forward stepwise:

Tempo: 0.46 segundos; $\widehat{R}(g^*) = 0.30$

Covariáveis selecionadas: $x_1, x_2, x_3, x_4, x_5, x_{10}, x_{12}, x_{13}, x_{19}, x_{20}$

Lasso:

Tempo: 0.09 segundos; $\widehat{R}(g^*) = 0.25$

Covariáveis selecionadas: x_1, x_2, x_3, x_4, x_5

Resultados:

Todos os subconjuntos:

Tempo: 1 hora e 20 minutos; $\widehat{R}(g^*) = 0.30$

Covariáveis selecionadas: $x_1, x_2, x_3, x_4, x_5, x_{10}, x_{12}, x_{13}, x_{19}, x_{20}$

Forward stepwise:

Tempo: 0.46 segundos; $\widehat{R}(g^*) = 0.30$

Covariáveis selecionadas: $x_1, x_2, x_3, x_4, x_5, x_{10}, x_{12}, x_{13}, x_{19}, x_{20}$

Lasso:

Tempo: 0.09 segundos; $\widehat{R}(g^*) = 0.25$

Covariáveis selecionadas: x_1, x_2, x_3, x_4, x_5

Resultados:

Todos os subconjuntos:

Tempo: 1 hora e 20 minutos; $\widehat{R}(g^*) = 0.30$

Covariáveis selecionadas: $x_1, x_2, x_3, x_4, x_5, x_{10}, x_{12}, x_{13}, x_{19}, x_{20}$

Forward stepwise:

Tempo: 0.46 segundos; $\widehat{R}(g^*) = 0.30$

Covariáveis selecionadas: $x_1, x_2, x_3, x_4, x_5, x_{10}, x_{12}, x_{13}, x_{19}, x_{20}$

Lasso:

Tempo: 0.09 segundos; $\widehat{R}(g^*) = 0.25$

Covariáveis selecionadas: x_1, x_2, x_3, x_4, x_5

Resultados:

Todos os subconjuntos:

Tempo: 1 hora e 20 minutos; $\widehat{R}(g^*) = 0.30$

Covariáveis selecionadas: $x_1, x_2, x_3, x_4, x_5, x_{10}, x_{12}, x_{13}, x_{19}, x_{20}$

Forward stepwise:

Tempo: 0.46 segundos; $\widehat{R}(g^*) = 0.30$

Covariáveis selecionadas: $x_1, x_2, x_3, x_4, x_5, x_{10}, x_{12}, x_{13}, x_{19}, x_{20}$

Lasso:

Tempo: 0.09 segundos; $\widehat{R}(g^*) = 0.25$

Covariáveis selecionadas: x_1, x_2, x_3, x_4, x_5

Resultados:

Todos os subconjuntos:

Tempo: 1 hora e 20 minutos; $\widehat{R}(g^*) = 0.30$

Covariáveis selecionadas: $x_1, x_2, x_3, x_4, x_5, x_{10}, x_{12}, x_{13}, x_{19}, x_{20}$

Forward stepwise:

Tempo: 0.46 segundos; $\widehat{R}(g^*) = 0.30$

Covariáveis selecionadas: $x_1, x_2, x_3, x_4, x_5, x_{10}, x_{12}, x_{13}, x_{19}, x_{20}$

Lasso:

Tempo: 0.09 segundos; $\widehat{R}(g^*) = 0.25$

Covariáveis selecionadas: x_1, x_2, x_3, x_4, x_5

Resultados:

Todos os subconjuntos:

Tempo: 1 hora e 20 minutos; $\widehat{R}(g^*) = 0.30$

Covariáveis selecionadas: $x_1, x_2, x_3, x_4, x_5, x_{10}, x_{12}, x_{13}, x_{19}, x_{20}$

Forward stepwise:

Tempo: 0.46 segundos; $\widehat{R}(g^*) = 0.30$

Covariáveis selecionadas: $x_1, x_2, x_3, x_4, x_5, x_{10}, x_{12}, x_{13}, x_{19}, x_{20}$

Lasso:

Tempo: 0.09 segundos; $\widehat{R}(g^*) = 0.25$

Covariáveis selecionadas: x_1, x_2, x_3, x_4, x_5

Resumindo

Em uma regressão linear . . .

- Usar todas as covariáveis pode levar ao overfitting
- Lasso permite escolher variáveis de forma rápida

$\Rightarrow$ R.

Resumindo

Em uma regressão linear . . .

- Usar todas as covariáveis pode levar ao overfitting
- Lasso permite escolher variáveis de forma rápida

$\Rightarrow$ R.

Resumindo

Em uma regressão linear . . .

- Usar todas as covariáveis pode levar ao overfitting
- Lasso permite escolher variáveis de forma rápida

$\Rightarrow$ R.

Resumindo

Em uma regressão linear . . .

- Usar todas as covariáveis pode levar ao overfitting
- Lasso permite escolher variáveis de forma rápida

$\Rightarrow$ R.

Métodos não paramétricos

Até agora, nos restringimos a métodos *paramétricos*

Número finito de parâmetros

Ex:

$$r(x) = \beta_0 + \beta_1 x$$

$$r(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$r(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_1 x_2$$

Métodos paramétricos muitas vezes são muito restritivos e simplistas.

Viés alto

Até agora, nos restringimos a métodos *paramétricos*

Número finito de parâmetros

Ex:

$$r(x) = \beta_0 + \beta_1 x$$

$$r(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$r(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_1 x_2$$

Métodos paramétricos muitas vezes são muito restritivos e simplistas.

Viés alto

Até agora, nos restringimos a métodos *paramétricos*

Número finito de parâmetros

Ex:

$$r(x) = \beta_0 + \beta_1 x$$

$$r(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$r(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_1 x_2$$

Métodos paramétricos muitas vezes são muito restritivos e simplistas.

Viés alto

Até agora, nos restringimos a métodos *paramétricos*

Número finito de parâmetros

Ex:

$$r(x) = \beta_0 + \beta_1 x$$

$$r(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$r(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_1 x_2$$

Métodos paramétricos muitas vezes são muito restritivos e simplistas.

Viés alto

Até agora, nos restringimos a métodos *paramétricos*

Número finito de parâmetros

Ex:

$$r(x) = \beta_0 + \beta_1 x$$

$$r(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$r(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_1 x_2$$

Métodos paramétricos muitas vezes são muito restritivos e simplistas.

Viés alto

n é grande: modelos mais flexíveis

n grande o suficiente para a variância não aumentar muito

n é grande: modelos mais flexíveis

n grande o suficiente para a variância não aumentar muito

Método dos k Vizinhos Mais Próximos

Benedetti, 1977 e Stone, 1977; popular em aprendizado de máquina: KNN

Motivação: médias locais

Método dos k Vizinhos Mais Próximos

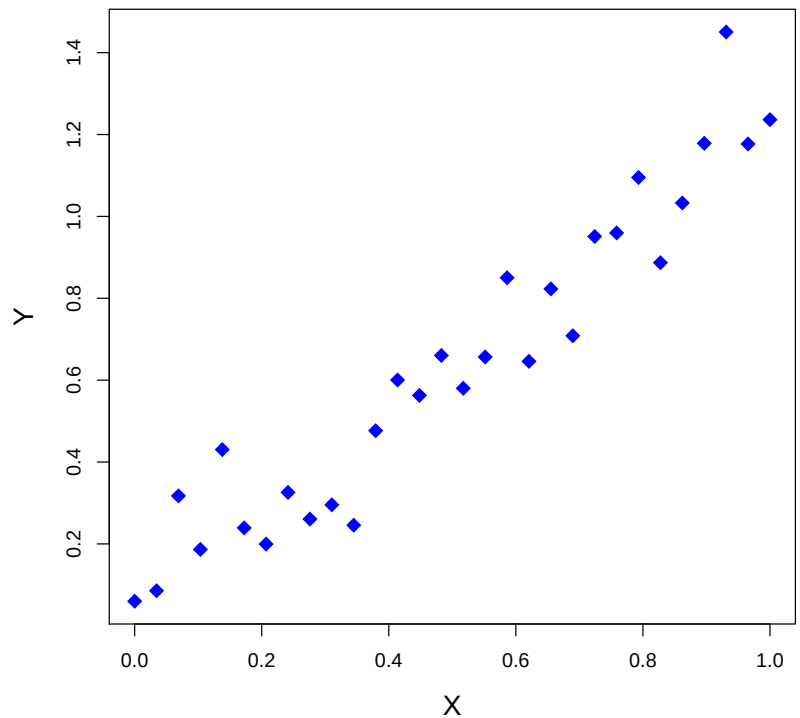
Benedetti, 1977 e Stone, 1977; popular em aprendizado de máquina: KNN

Motivação: médias locais

Método dos k Vizinhos Mais Próximos

Benedetti, 1977 e Stone, 1977; popular em aprendizado de máquina: KNN

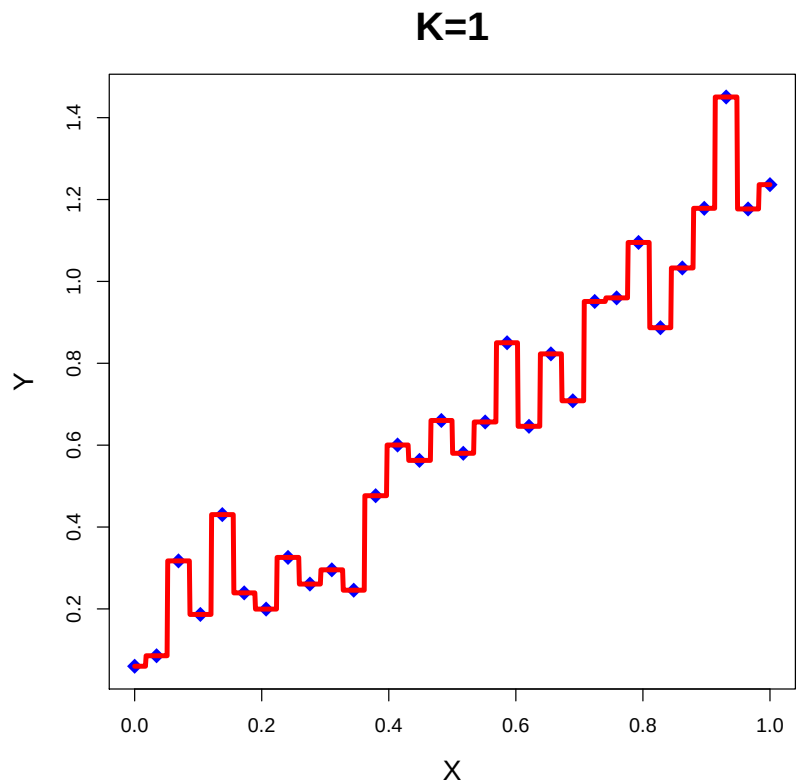
Motivação: médias locais



Método dos k Vizinhos Mais Próximos

Benedetti, 1977 e Stone, 1977; popular em aprendizado de máquina: KNN

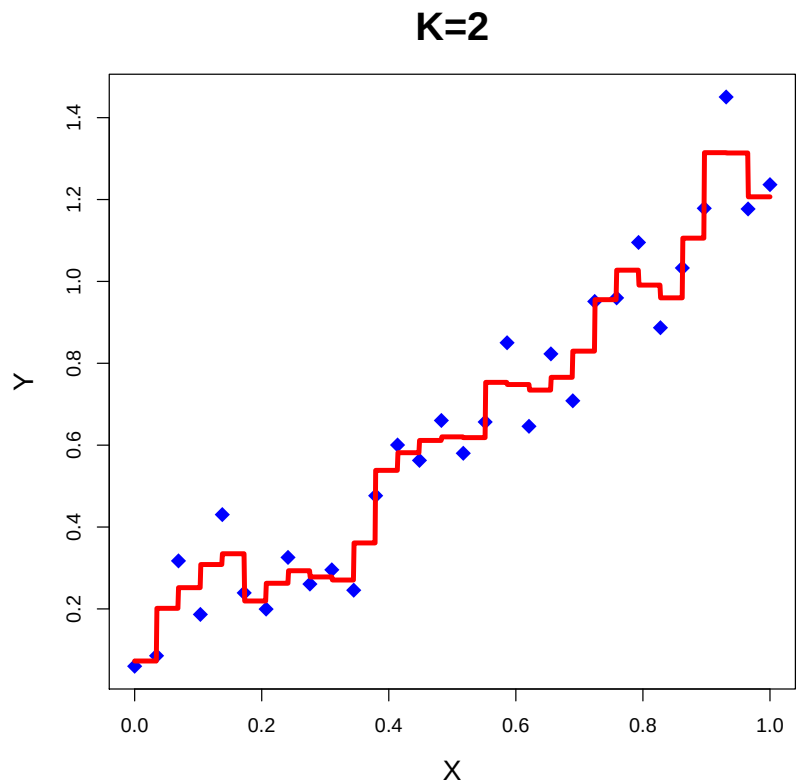
Motivação muito diferente: médias locais



Método dos k Vizinhos Mais Próximos

Benedetti, 1977 e Stone, 1977; popular em aprendizado de máquina: KNN

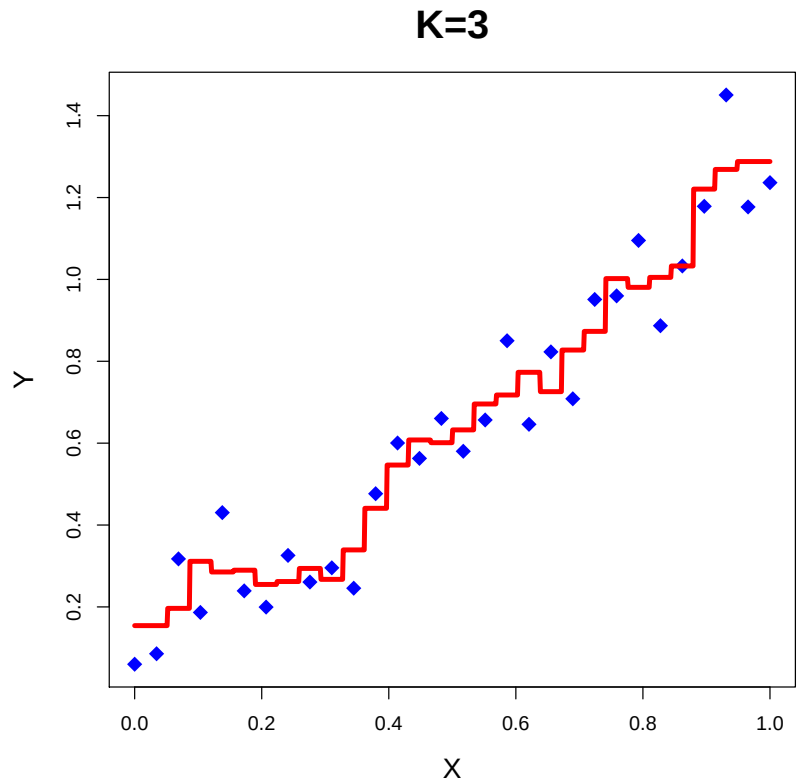
Motivação muito diferente: médias locais



Método dos k Vizinhos Mais Próximos

Benedetti, 1977 e Stone, 1977; popular em aprendizado de máquina: KNN

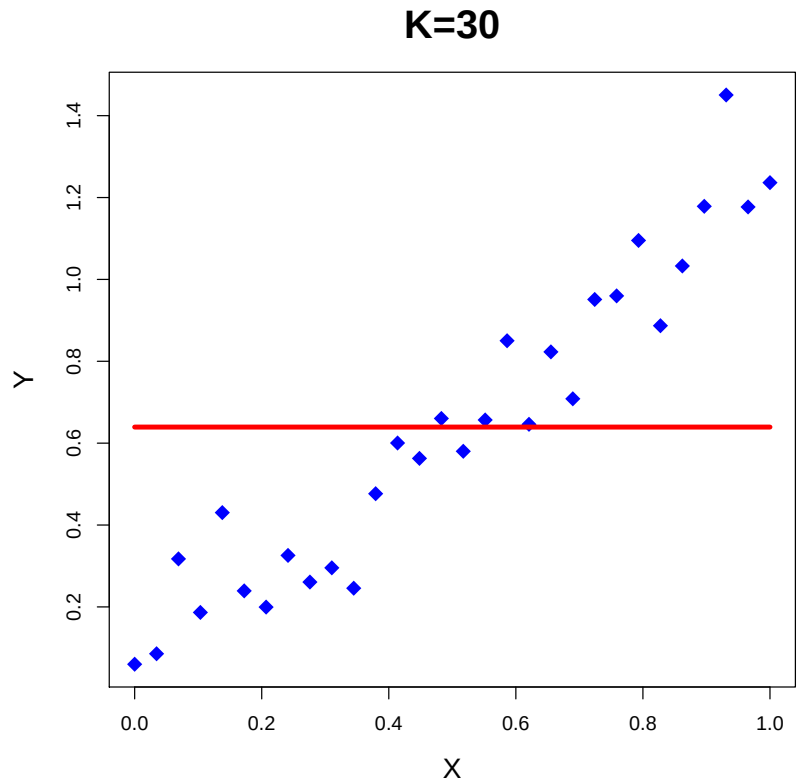
Motivação muito diferente: médias locais



Método dos k Vizinhos Mais Próximos

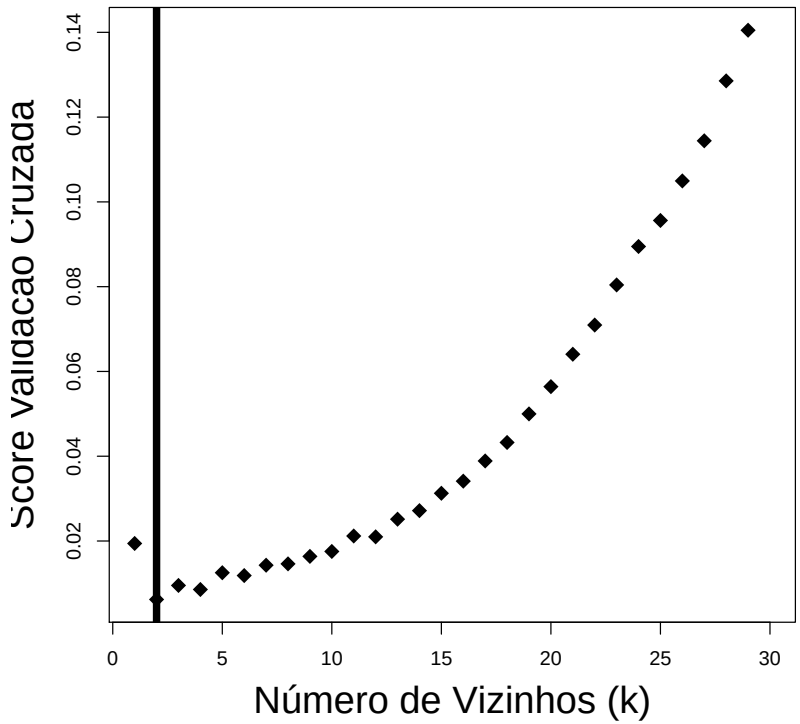
Benedetti, 1977 e Stone, 1977; popular em aprendizado de máquina: KNN

Motivação muito diferente: médias locais



Como escolher k? Validação cruzada!

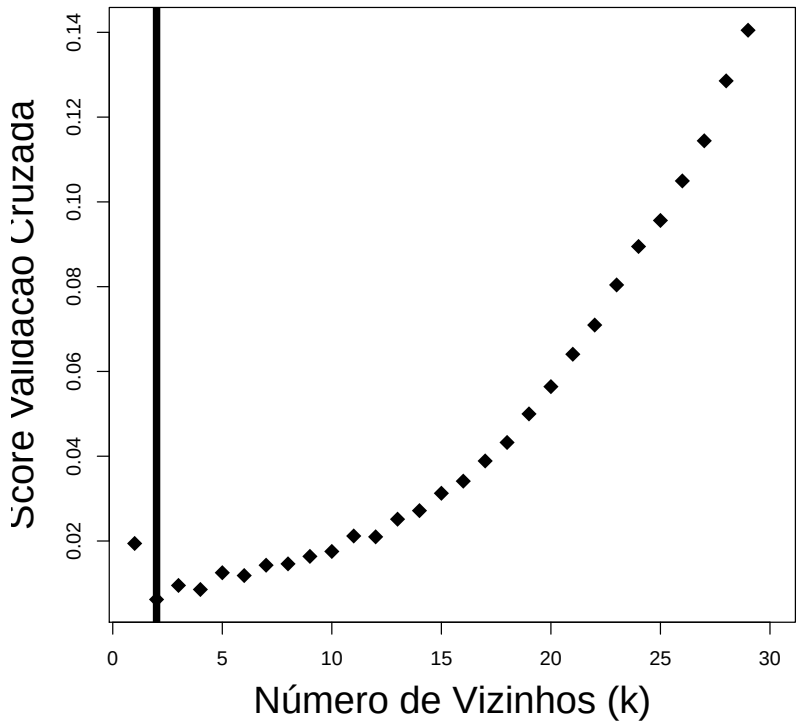
Papel no balanço viés-variância?



k=2

Como escolher k? Validação cruzada!

Papel no balanço viés-variância?



k=2

É necessário guardar todas as observações para se fazer predições

Taxa de Convergência

Teorema: Se r é L-Lipschitz,

$$\mathbb{E}[(\widehat{r}(\mathbf{X}) - r(\mathbf{X}))^2] \leq Kn^{-\frac{2}{2+d}}$$

Se covariáveis têm redundância, taxas melhores

$\Rightarrow$ R.

Taxa de Convergência

Teorema: Se r é L-Lipschitz,

$$\mathbb{E}[(\widehat{r}(\mathbf{X}) - r(\mathbf{X}))^2] \leq Kn^{-\frac{2}{2+d}}$$

Se covariáveis têm redundância, taxas melhores

$\Rightarrow$ R.

Taxa de Convergência

Teorema: Se r é L-Lipschitz,

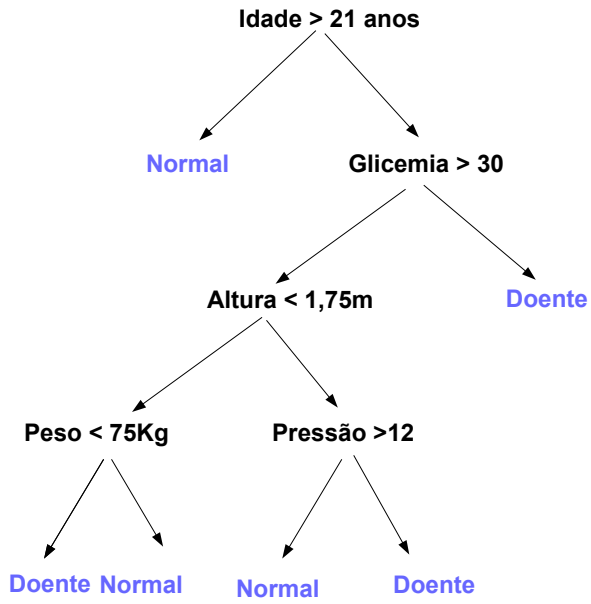
$$\mathbb{E}[(\widehat{r}(\mathbf{X}) - r(\mathbf{X}))^2] \leq Kn^{-\frac{2}{2+d}}$$

Se covariáveis têm redundância, taxas melhores

$\Rightarrow$ R.

Árvore de Regressão

Árvores de Regressão



O que é uma árvore?

Nós, folhas

Ideia: dividir o espaço das covariáveis em uma partição $R_1, \ldots, R_J$

Se $x \in R_k$,

$$g(\mathbf{x}) = \frac{1}{|\{i : \mathbf{x}_i \in R_k\}|} \sum_{i:\mathbf{x}_i \in R_k} y_i,$$

Ideia: dividir o espaço das covariáveis em uma partição $R_1, \ldots, R_J$

Se $x \in R_k$,

$$g(\mathbf{x}) = \frac{1}{|\{i : \mathbf{x}_i \in R_k\}|} \sum_{i:\mathbf{x}_i \in R_k} y_i,$$

Como determinar as regiões $R_1, \ldots, R_J$?

1. Criamos uma árvore "grande"

2. Podamos esta árvore

Como determinar as regiões $R_1, \ldots, R_J$?

1. Criamos uma árvore "grande"

2. Podamos esta árvore

Etapa 1:

Medida de quão pura uma árvore T é:

$$\mathcal{P}(T) = \sum_{R} \sum_{\mathbf{x}_k \in R} (y_k - \widehat{y}_R)^2,$$

$\widehat{y}_R$: valor predito para uma observação pertencente à região R

Etapa 1:

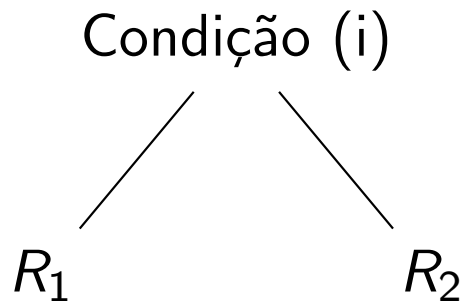
Medida de quão pura uma árvore T é:

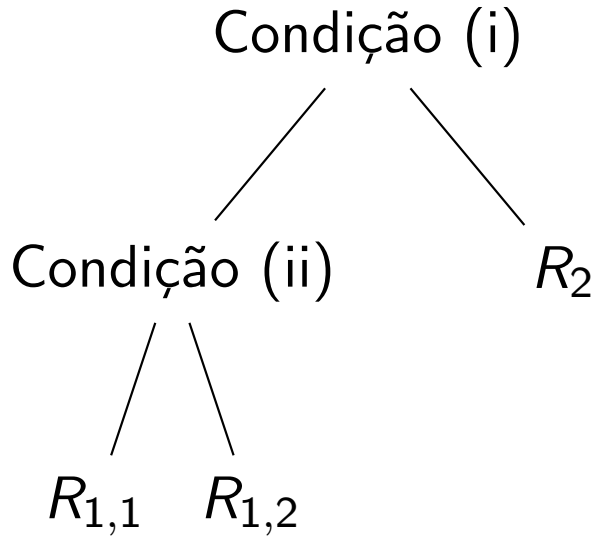
$$\mathcal{P}(T) = \sum_{R} \sum_{\mathbf{x}_k \in R} (y_k - \widehat{y}_R)^2,$$

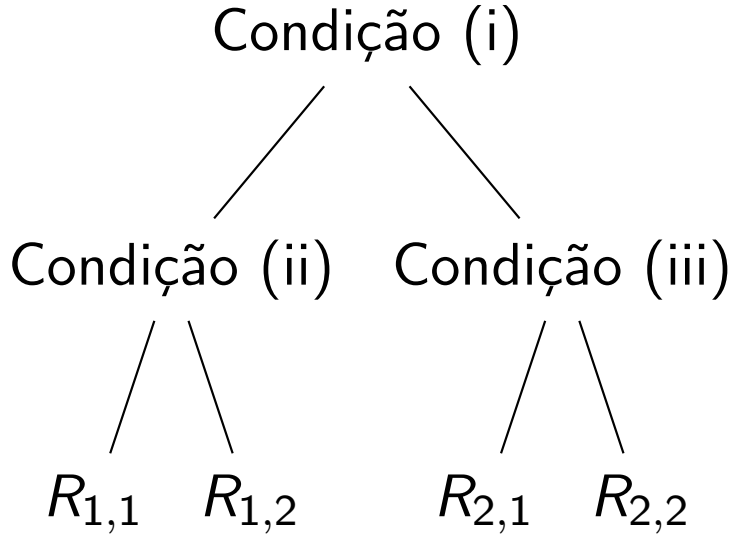
$\widehat{y}_R$: valor predito para uma observação pertencente à região R

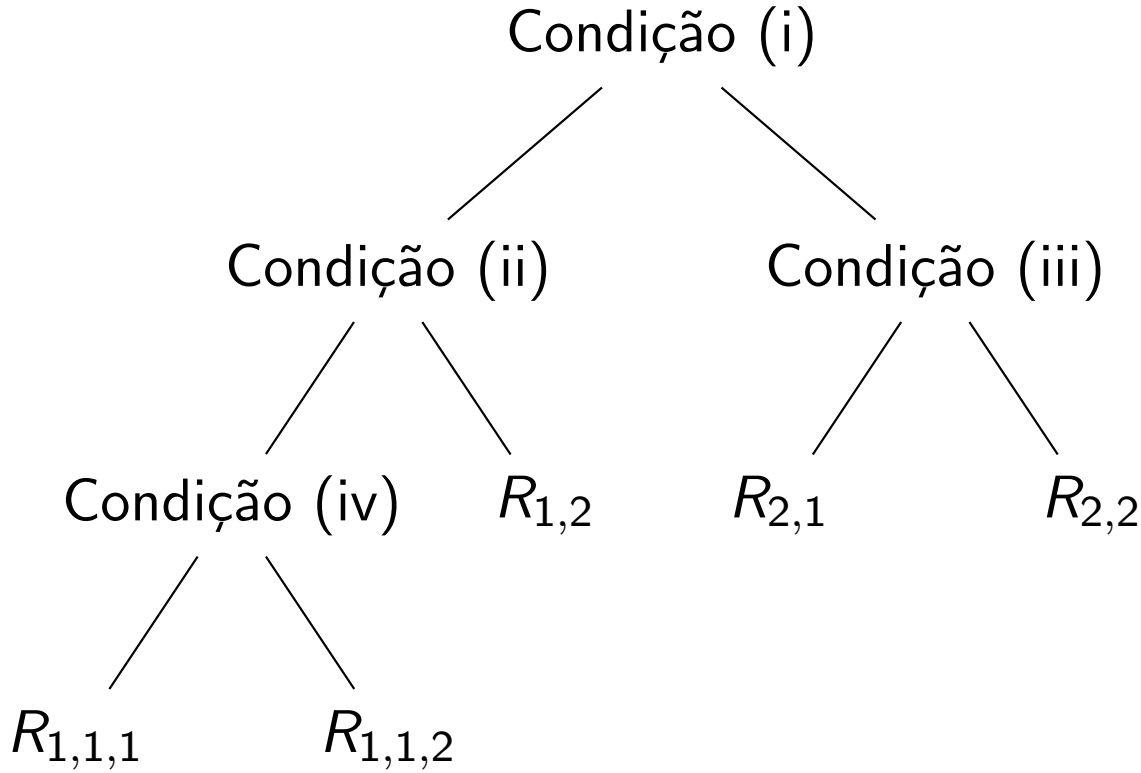
Etapa 1: Divisões binárias recursivas

Como encontrar T com $P(T)$ pequeno?


Condição (i)
R_1
R_2


Condição (i)
Condição (ii)
R_2
$R_{1,1}$
$R_{1,2}$


Condição (i)
Condição (ii)
Condição (iii)
$R_{1,1}$
$R_{1,2}$
$R_{2,1}$
$R_{2,2}$


Condição (i)
Condição (ii)
Condição (iii)
Condição (iv)
$R_{1,2}$
$R_{2,1}$
$R_{2,2}$
$R_{1,1,1}$
$R_{1,1,2}$

Prosseguimos até criar uma árvore grande.

Problema: overfitting.

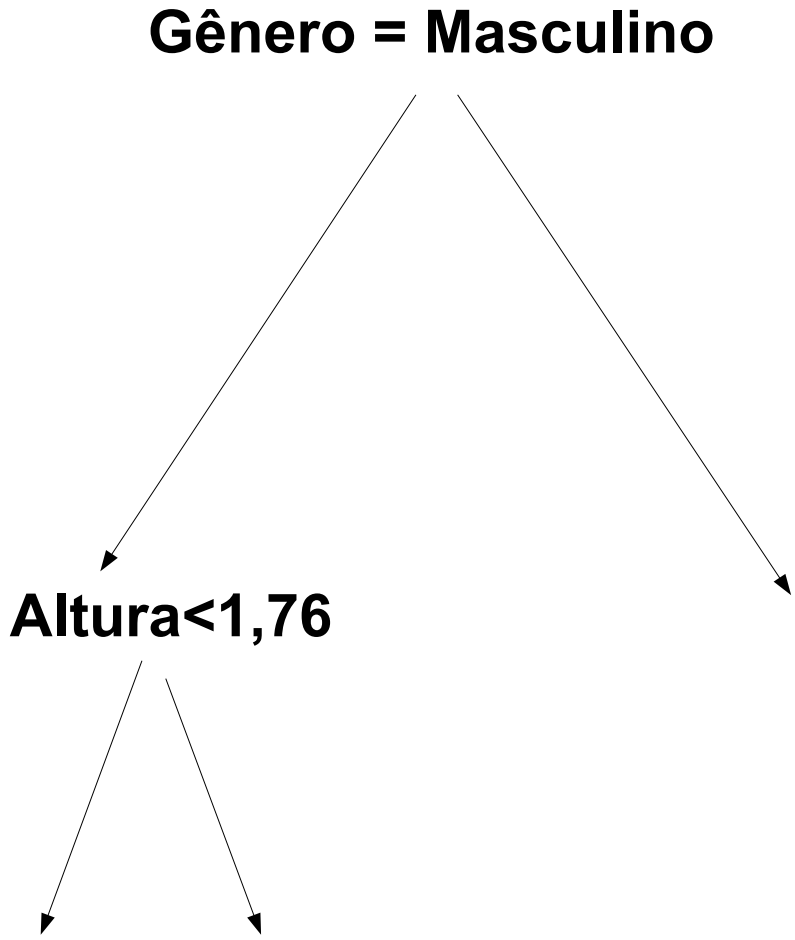
Prosseguimos até criar uma árvore grande.

Problema: overfitting.

Etapa 2: Poda

Retiramos cada nó da árvore, um por vez

É fácil adicionar um variável discreta X_i!



Bagging e Florestas Aleatórias

Combinando Predições

Imagine que temos duas funções de predição para Y, $g_1(\mathbf{x})$ e $g_2(\mathbf{x})$.

Se g_1 e g_2 são:

(i) não correlacionados

(ii) não viesados

(iii) têm mesma variância, então

$$R(g) \leq R(g_i),$$

$g(\mathbf{x}) = (g_1(\mathbf{x}) + g_2(\mathbf{x}))/2$

Assim, é melhor combinar as predições.

Se g_1 e g_2 são:

(i) não correlacionados

(ii) não viesados

(iii) têm mesma variância, então

$$R(g) \leq R(g_i),$$

$g(\mathbf{x}) = (g_1(\mathbf{x}) + g_2(\mathbf{x}))/2$

Assim, é melhor combinar as predições.

Se g_1 e g_2 são:

(i) não correlacionados

(ii) não viesados

(iii) têm mesma variância, então

$$R(g) \leq R(g_i),$$

$g(\mathbf{x}) = (g_1(\mathbf{x}) + g_2(\mathbf{x}))/2$

Assim, é melhor combinar as predições.

Random Forests/Bagging: usar isso para melhorar predições de árvores

Criamos B árvores e combinamos seus resultados

Para criar árvores próximas de não-viesadas, não as podamos.

Random Forests/Bagging: usar isso para melhorar predições de árvores

Criamos B árvores e combinamos seus resultados

Para criar árvores próximas de não-viesadas, não as podamos.

Random Forests/Bagging: usar isso para melhorar predições de árvores

Criamos B árvores e combinamos seus resultados

Para criar árvores próximas de não-viesadas, não as podamos.

Bagging

Ideia: Criamos B amostras bootstrap da amostra original

Para cada um delas, criamos uma árvore não podada.

Função de predição:

$$g(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} g^b(\mathbf{x})$$

Bagging

Ideia: Criamos B amostras bootstrap da amostra original

Para cada um delas, criamos uma árvore não podada.

Função de predição:

$$g(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} g^b(\mathbf{x})$$

Bagging

Ideia: Criamos B amostras bootstrap da amostra original

Para cada um delas, criamos uma árvore não podada.

Função de predição:

$$g(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} g^b(\mathbf{x})$$

Perdemos interpretação de árvores

Medida de importância para cada covariável: a média de quanto ela foi importante em cada árvore.

Perdemos interpretação de árvores

Medida de importância para cada covariável: a média de quanto ela foi importante em cada árvore.

Random Forests - Florestas Aleatórias

Objetivo: diminuir a correlação entre os diferentes g^b's

Mesma ideia de bagging, mas cada nó só pode escolher uma dentre $m < d$ covariáveis.

O subconjunto de covariáveis é escolhido aleatoriamente para cada nó.

$\Rightarrow$ R.

Random Forests - Florestas Aleatórias

Objetivo: diminuir a correlação entre os diferentes g^b's

Mesma ideia de bagging, mas cada nó só pode escolher uma dentre $m < d$ covariáveis.

O subconjunto de covariáveis é escolhido aleatoriamente para cada nó.

$\Rightarrow$ R.

Random Forests - Florestas Aleatórias

Objetivo: diminuir a correlação entre os diferentes g^b's

Mesma ideia de bagging, mas cada nó só pode escolher uma dentre $m < d$ covariáveis.

O subconjunto de covariáveis é escolhido aleatoriamente para cada nó.

$\Rightarrow$ R.

Random Forests - Florestas Aleatórias

Objetivo: diminuir a correlação entre os diferentes g^b's

Mesma ideia de bagging, mas cada nó só pode escolher uma dentre $m < d$ covariáveis.

O subconjunto de covariáveis é escolhido aleatoriamente para cada nó.

$\Rightarrow$ R.

Resumindo

Métodos não paramétricos:

- KNN: intuitivo; funciona bem se há muitas covariáveis redundantes
- Árvores: fácil de interpretar; predições não tão boas
- Bagging e Florestas: combinar árvores; funcionam bem se há muitas covariáveis irrelevantes

Resumindo

Métodos não paramétricos:

- KNN: intuitivo; funciona bem se há muitas covariáveis redundantes
- Árvores: fácil de interpretar; predições não tão boas
- Bagging e Florestas: combinar árvores; funcionam bem se há muitas covariáveis irrelevantes

Resumindo

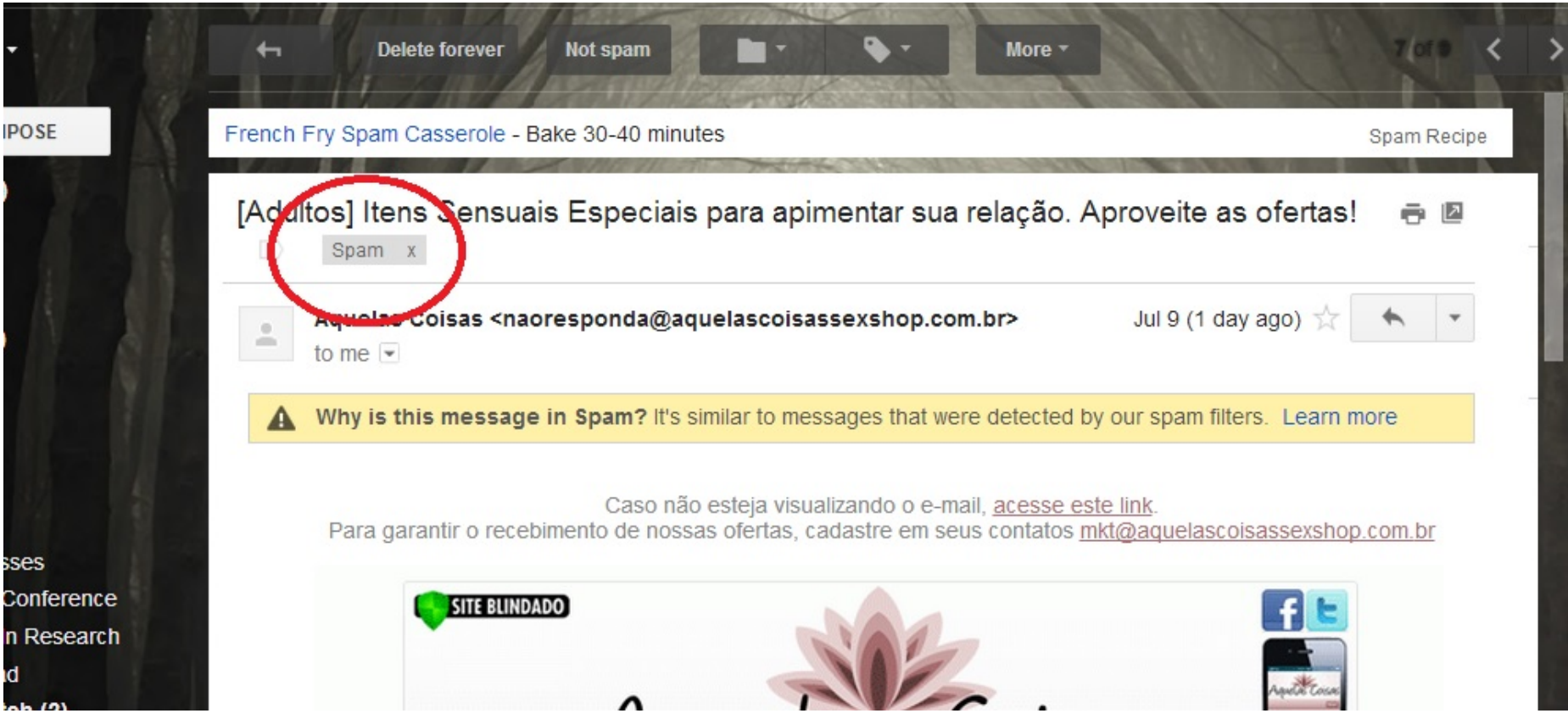
Métodos não paramétricos:

- KNN: intuitivo; funciona bem se há muitas covariáveis redundantes
- Árvores: fácil de interpretar; predições não tão boas
- Bagging e Florestas: combinar árvores; funcionam bem se há muitas covariáveis irrelevantes

Classificação

Em muitos problemas, a variável Y é uma variável qualitativa.

Exemplo: Detecção de Spams

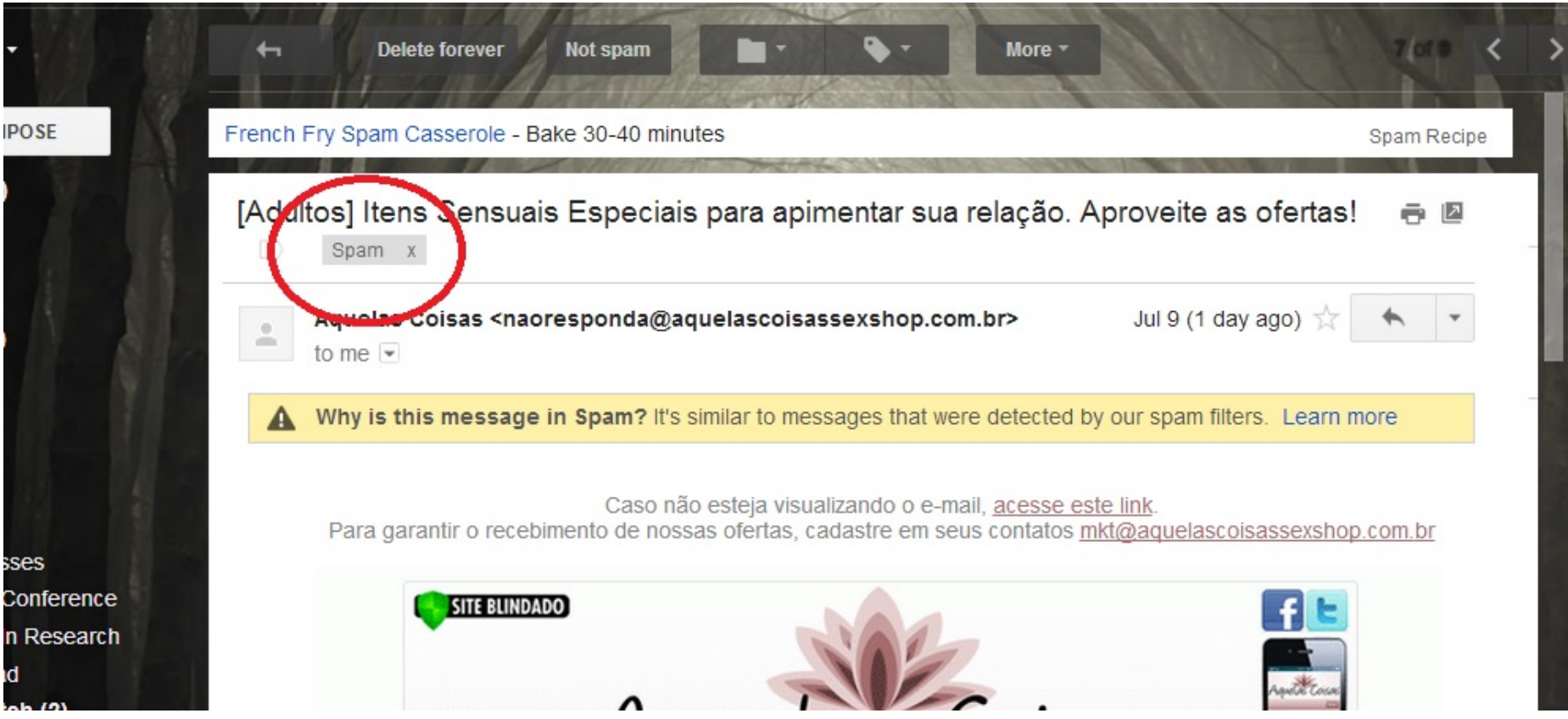


$\mathbf{X}_i \longrightarrow$ email

$Y_i \in \{\text{spam, não spam}\}$

Objetivo: prever Y_i com base em $\mathbf{X}_i$

Exemplo: Detecção de Spams

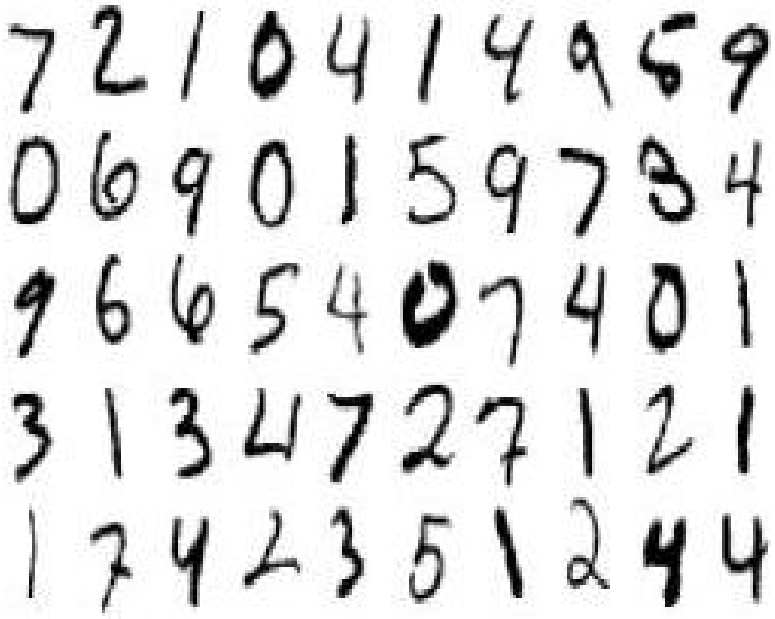


$\mathbf{X}_i \longrightarrow$ email

$Y_i \in \{\text{spam, não spam}\}$

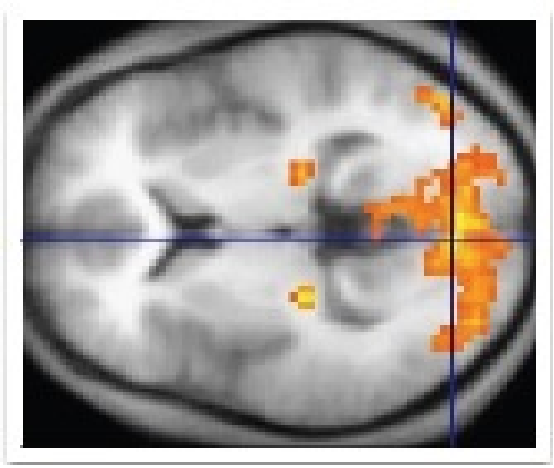
Objetivo: prever Y_i com base em $\mathbf{X}_i$

Exemplo: Reconhecimento de Dígitos



$\mathbf{X}_i \longrightarrow$ imagem de um dígito
$Y_i \in \{0, 1, \ldots, 9\}$

Exemplo: Leitura de Pensamentos



⇓⇓⇓

$\mathbf{X}_i \longrightarrow$ imagem da ressonância magnética

$Y_i \in \{\text{Minicurso chato, Não to entendendo nada, WTF???, } \ldots\}$

Y é uma variável qualitativa: problema de classificação.

$\Rightarrow$ $R(g) = \mathbb{E}[(Y - g(\mathbf{X}))^2]$ **não faz mais sentido**

É comum usar

$$R(g) := \mathbb{E}[\mathbb{I}(Y \neq g(\mathbf{X}))] = \mathbb{P}(Y \neq g(\mathbf{X}))$$

Y é uma variável qualitativa: problema de classificação.

$\Rightarrow$ $R(g) = \mathbb{E}[(Y - g(\mathbf{X}))^2]$ **não faz mais sentido**

É comum usar

$$R(g) := \mathbb{E}[\mathbb{I}(Y \neq g(\mathbf{X}))] = \mathbb{P}(Y \neq g(\mathbf{X}))$$

Y é uma variável qualitativa: problema de classificação.

$\Rightarrow R(g) = \mathbb{E}[(Y - g(\mathbf{X}))^2]$ **não faz mais sentido**

É comum usar

$$R(g) := \mathbb{E}[\mathbb{I}(Y \neq g(\mathbf{X}))] = \mathbb{P}(Y \neq g(\mathbf{X}))$$

$$R(g) := \mathbb{P}\left(Y \neq g(\mathbf{X})\right)$$

Quem minimiza $R(g)$?

$$R(g) := \mathbb{P}\left(Y \neq g(\mathbf{X})\right)$$

Quem minimiza $R(g)$?

Melhor g:

$$g(\mathbf{x}) = \arg\max_{d} \mathbb{P}(Y = d|\mathbf{x})$$

Classificador de Bayes

Para o caso binário,

$$g(\mathbf{x}) = 1 \iff \mathbb{P}(Y = 1|\mathbf{x}) \geq \frac{1}{2}.$$

Melhor g:

$$g(\mathbf{x}) = \arg\max_d \mathbb{P}(Y = d|\mathbf{x})$$

Classificador de Bayes

Para o caso binário,

$$g(\mathbf{x}) = 1 \iff \mathbb{P}(Y = 1|\mathbf{x}) \geq \frac{1}{2}.$$

Melhor g:

$$g(\mathbf{x}) = \arg\max_{d} \mathbb{P}(Y = d|\mathbf{x})$$

Classificador de Bayes

Para o caso binário,

$$g(\mathbf{x}) = 1 \iff \mathbb{P}(Y = 1|\mathbf{x}) \geq \frac{1}{2}.$$

Sugere uma abordagem simples:

(1) Estimar $\mathbb{P}(Y = c|\mathbf{x})$, para cada c.

(2) Tomar

$$g(\mathbf{x}) = \arg\max_{c \in \mathcal{C}} \widehat{P}(Y = c|\mathbf{x})$$

Plug-in classifier.

Criar um classificador resume-se a estimar $\mathbb{P}(Y = c|\mathbf{x})$.

Sugere uma abordagem simples:

(1) Estimar $\mathbb{P}(Y = c|\mathbf{x})$, para cada c.

(2) Tomar

$$g(\mathbf{x}) = \arg\max_{c \in \mathcal{C}} \widehat{P}(Y = c|\mathbf{x})$$

Plug-in classifier.

Criar um classificador resume-se a estimar $\mathbb{P}(Y = c|\mathbf{x})$.

Sugere uma abordagem simples:

(1) Estimar $\mathbb{P}(Y = c|\mathbf{x})$, para cada c.

(2) Tomar

$$g(\mathbf{x}) = \arg\max_{c \in \mathcal{C}} \widehat{P}(Y = c|\mathbf{x})$$

Plug-in classifier.

Criar um classificador resume-se a estimar $\mathbb{P}(Y = c|\mathbf{x})$.

Sugere uma abordagem simples:

(1) Estimar $\mathbb{P}(Y = c|\mathbf{x})$, para cada c.

(2) Tomar

$$g(\mathbf{x}) = \arg\max_{c \in \mathcal{C}} \widehat{P}(Y = c|\mathbf{x})$$

Plug-in classifier.

Criar um classificador resume-se a estimar $\mathbb{P}(Y = c|\mathbf{x})$.

Sugere uma abordagem simples:

(1) Estimar $\mathbb{P}(Y = c|\mathbf{x})$, para cada c.

(2) Tomar

$$g(\mathbf{x}) = \arg\max_{c \in \mathcal{C}} \widehat{P}(Y = c|\mathbf{x})$$

Plug-in classifier.

Criar um classificador resume-se a estimar $\mathbb{P}(Y = c|\mathbf{x})$.

Regressão Logística

Assumindo Y binário,

$$\mathbb{P}(Y = 1|\mathbf{x}) = \frac{e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_i}}$$

Pode-se adicionar penalização

$\Rightarrow$ R.

Regressão Logística

Assumindo Y binário,

$$\mathbb{P}(Y = 1|\mathbf{x}) = \frac{e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_i}}$$

Pode-se adicionar penalização

$\Rightarrow$ R.

Regressão Logística

Assumindo Y binário,

$$\mathbb{P}(Y = 1|\mathbf{x}) = \frac{e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_i}}$$

Pode-se adicionar penalização

$\Rightarrow$ R.

Regressão Logística

Assumindo Y binário,

$$\mathbb{P}(Y = 1|\mathbf{x}) = \frac{e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{d} \beta_i x_i}}$$

Pode-se adicionar penalização

$\Rightarrow$ R.

Uma alternativa: Regressão Linear

$$\mathbb{P}(Y = 1|\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

$\widehat{\mathbb{P}}(Y = 1|\mathbf{x})$ pode ser menor que 0 ou maior que 1.

Ainda assim podemos usar o classificador

$$g(\mathbf{x}) = \mathbb{I}(\widehat{\mathbb{P}}(Y = 1|\mathbf{x}) \geq 1/2).$$

$\Rightarrow$ R.

Uma alternativa: Regressão Linear

$$\mathbb{P}(Y = 1|\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

$\widehat{\mathbb{P}}(Y = 1|\mathbf{x})$ pode ser menor que 0 ou maior que 1.

Ainda assim podemos usar o classificador

$$g(\mathbf{x}) = \mathbb{I}(\widehat{\mathbb{P}}(Y = 1|\mathbf{x}) \geq 1/2).$$

$\Rightarrow$ R.

Uma alternativa: Regressão Linear

$$\mathbb{P}(Y = 1|\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

$\widehat{\mathbb{P}}(Y = 1|\mathbf{x})$ pode ser menor que 0 ou maior que 1.

Ainda assim podemos usar o classificador

$$g(\mathbf{x}) = \mathbb{I}(\widehat{\mathbb{P}}(Y = 1|\mathbf{x}) \geq 1/2).$$

$\Rightarrow$ R.

Uma alternativa: Regressão Linear

$$\mathbb{P}(Y = 1|\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d$$

$\widehat{\mathbb{P}}(Y = 1|\mathbf{x})$ pode ser menor que 0 ou maior que 1.

Ainda assim podemos usar o classificador

$$g(\mathbf{x}) = \mathbb{I}(\widehat{\mathbb{P}}(Y = 1|\mathbf{x}) \geq 1/2).$$

$\Rightarrow$ R.

Naive Bayes

Naive Bayes

$$\mathbb{P}(Y = c|\mathbf{x}) = \frac{f(\mathbf{x}|Y = c)\mathbb{P}(Y = c)}{\sum_{s \in \mathcal{X}} f(\mathbf{x}|Y = s)\mathbb{P}(Y = s)}$$

$\mathbb{P}(Y = s)$ facilmente estimada

Para estimar $f(\mathbf{x}|Y = s)$, precisamos assumir algum modelo para as covariáveis.

Naive Bayes

$$\mathbb{P}(Y = c|\mathbf{x}) = \frac{f(\mathbf{x}|Y = c)\mathbb{P}(Y = c)}{\sum_{s \in \mathcal{X}} f(\mathbf{x}|Y = s)\mathbb{P}(Y = s)}$$

$\mathbb{P}(Y = s)$ facilmente estimada

Para estimar $f(\mathbf{x}|Y = s)$, precisamos assumir algum modelo para as covariáveis.

Naive Bayes

$$\mathbb{P}(Y = c|\mathbf{x}) = \frac{f(\mathbf{x}|Y = c)\mathbb{P}(Y = c)}{\sum_{s \in \mathcal{X}} f(\mathbf{x}|Y = s)\mathbb{P}(Y = s)}$$

$\mathbb{P}(Y = s)$ facilmente estimada

Para estimar $f(\mathbf{x}|Y = s)$, precisamos assumir algum modelo para as covariáveis.

Naive Bayes

Suposição:

$$f(\mathbf{x}|Y = s) = f(x_1, \ldots, x_d|Y = s) = \prod_{j=1}^{d} f(x_j|Y = s),$$

Não é razoável em muitos problemas, mas pode levar a bons classificadores.

Naive Bayes

Suposição:

$$f(\mathbf{x}|Y = s) = f(x_1, \ldots, x_d|Y = s) = \prod_{j=1}^{d} f(x_j|Y = s),$$

Não é razoável em muitos problemas, mas pode levar a bons classificadores.

Naive Bayes

Suposição:

$$f(\mathbf{x}|Y = s) = f(x_1, \ldots, x_d|Y = s) = \prod_{j=1}^{d} f(x_j|Y = s),$$

Não é razoável em muitos problemas, mas pode levar a bons classificadores.

Naive Bayes

Podemos estimar $f(x_j|Y = s)$ assumindo, e.g.,

$$X_j|Y = s \sim N(\mu_{j,s}, \sigma^2_{j,s}),\ j = 1, \ldots, p$$

Parâmetros podem ser estimados via EMV

Naive Bayes

Podemos estimar $f(x_j|Y = s)$ assumindo, e.g.,

$$X_j|Y = s \sim N(\mu_{j,s}, \sigma^2_{j,s}),\ j = 1, \ldots, p$$

Parâmetros podem ser estimados via EMV

Assim,

$$\widehat{f}(\mathbf{x}|Y=c) = \prod_{k=1}^{d} \widehat{f}(x_k|Y=c) = \prod_{k=1}^{d} \frac{1}{\sqrt{2\pi\widehat{\sigma^2}_{k,s}}} e^{-\left(\frac{(x_k-\widehat{\mu_{k,s}})^2}{2\widehat{\sigma^2_{k,s}}}\right)}$$

Sofisticação: redes Bayesianas

$\Rightarrow$ R.

Assim,

$$\widehat{f}(\mathbf{x}|Y=c) = \prod_{k=1}^{d} \widehat{f}(x_k|Y=c) = \prod_{k=1}^{d} \frac{1}{\sqrt{2\pi\widehat{\sigma^2}_{k,s}}} e^{-\left(\frac{(x_k-\widehat{\mu_{k,s}})^2}{2\widehat{\sigma^2_{k,s}}}\right)}$$

Sofisticação: redes Bayesianas

$\Rightarrow$ R.

Assim,

$$\widehat{f}(\mathbf{x}|Y=c) = \prod_{k=1}^{d} \widehat{f}(x_k|Y=c) = \prod_{k=1}^{d} \frac{1}{\sqrt{2\pi\widehat{\sigma^2}_{k,s}}} e^{-\left(\frac{(x_k-\widehat{\mu_{k,s}})^2}{2\widehat{\sigma^2_{k,s}}}\right)}$$

Sofisticação: redes Bayesianas

$\Rightarrow$ R.

Seleção de Modelos

Seleção de Modelos

Como estimar risco

$$R(g) := \mathbb{E}[\mathbb{I}(Y \neq g(\mathbf{X}))] = \mathbb{P}(Y \neq g(\mathbf{X}))?$$

Treinamento e validação;

$$\widehat{R}(g) := \frac{1}{m} \sum_{k=1}^{m} \mathbb{I}(Y_k' \neq g(\mathbf{X}_k')),$$

onde $(\mathbf{X}_1', Y_1'), \ldots, (\mathbf{X}_m', Y_m')$ é o conjunto de validação.

Seleção de Modelos

Como estimar risco

$$R(g) := \mathbb{E}[\mathbb{I}(Y \neq g(\mathbf{X}))] = \mathbb{P}(Y \neq g(\mathbf{X}))?$$

Treinamento e validação;

$$\widehat{R}(g) := \frac{1}{m} \sum_{k=1}^{m} \mathbb{I}(Y_k' \neq g(\mathbf{X}_k')),$$

onde $(\mathbf{X}_1', Y_1'), \ldots, (\mathbf{X}_m', Y_m')$ é o conjunto de validação.

Seleção de Modelos

Como estimar risco

$$R(g) := \mathbb{E}[\mathbb{I}(Y \neq g(\mathbf{X}))] = \mathbb{P}\,(Y \neq g(\mathbf{X}))?$$

Treinamento e validação;

$$\widehat{R}(g) := \frac{1}{m}\sum_{k=1}^{m} \mathbb{I}(Y_k^{'} \neq g(\mathbf{X}_k^{'})),$$

onde $(\mathbf{X}_1', Y_1'), \ldots, (\mathbf{X}_m', Y_m')$ é o conjunto de validação.

Teorema

$\mathbb{G} = \{g_1, \ldots, g_N\}$: classificadores estimados via cj. treinamento.

g^*: classificador que minimiza $R(g)$ dentre $g \in \mathbb{G}$;

$\widehat{g}$: classificador que minimiza $\widehat{R}(g)$ dentre $g \in \mathbb{G}$.

Com prob ao menos $1 - \epsilon$,

$$|R(\widehat{g}) - R(g^*)| \leq \sqrt{\frac{2}{m} \log \frac{2N}{\epsilon}}$$

Teorema

$\mathbb{G} = \{g_1, \ldots, g_N\}$: classificadores estimados via cj. treinamento.

g^*: classificador que minimiza $R(g)$ dentre $g \in \mathbb{G}$;

$\widehat{g}$: classificador que minimiza $\widehat{R}(g)$ dentre $g \in \mathbb{G}$.

Com prob ao menos $1 - \epsilon$,

$$|R(\widehat{g}) - R(g^*)| \leq \sqrt{\frac{2}{m} \log \frac{2N}{\epsilon}}$$

Teorema

$\mathbb{G} = \{g_1, \ldots, g_N\}$: classificadores estimados via cj. treinamento.

g^*: classificador que minimiza $R(g)$ dentre $g \in \mathbb{G}$;

$\widehat{g}$: classificador que minimiza $\widehat{R}(g)$ dentre $g \in \mathbb{G}$.

Com prob ao menos $1 - \epsilon$,

$$|R(\widehat{g}) - R(g^*)| \leq \sqrt{\frac{2}{m} \log \frac{2N}{\epsilon}}$$

Teorema

$\mathbb{G} = \{g_1, \ldots, g_N\}$: classificadores estimados via cj. treinamento.

g^*: classificador que minimiza $R(g)$ dentre $g \in \mathbb{G}$;

$\widehat{g}$: classificador que minimiza $\widehat{R}(g)$ dentre $g \in \mathbb{G}$.

Com prob ao menos $1 - \epsilon$,

$$|R(\widehat{g}) - R(g^*)| \leq \sqrt{\frac{2}{m} \log \frac{2N}{\epsilon}}$$

Teoria VC: extensão dessa ideia para um número infinito de classificadores

Perda 0-1 é razoável?

$$R(g) := \mathbb{E}[\mathbb{I}(Y \neq g(\mathbf{X}))]$$

Exemplo: Y indica se uma pessoa tem uma **doença rara**. Nossa amostra tem poucos pacientes com $Y = 1$.

O classificador $g(\mathbf{x}) \equiv 0$ tem $R(g)$ baixo...

Perda 0-1 é razoável?

$$R(g) := \mathbb{E}[\mathbb{I}(Y \neq g(\mathbf{X}))]$$

Exemplo: Y indica se uma pessoa tem uma **doença rara**. Nossa amostra tem poucos pacientes com $Y = 1$.

O classificador $g(\mathbf{x}) \equiv 0$ tem $R(g)$ baixo...

Perda 0-1 é razoável?

$$R(g) := \mathbb{E}[\mathbb{I}(Y \neq g(\mathbf{X}))]$$

Exemplo: Y indica se uma pessoa tem uma **doença rara**. Nossa amostra tem poucos pacientes com $Y = 1$.

O classificador $g(\mathbf{x}) \equiv 0$ tem $R(g)$ baixo...

Matriz de confusão

	Valor verdadeiro	
Valor Predito	Y=0	Y=1
Y=0	VN	FN
Y=1	FP	VP

- Sensibilidade: VP/(VP+FN)
- Especificidade: VN/(VN+FP)

Matriz de confusão

	Valor verdadeiro	
Valor Predito	Y=0	Y=1
Y=0	VN	FN
Y=1	FP	VP

- Sensibilidade: VP/(VP+FN)
- Especificidade: VN/(VN+FP)

Problema relacionado: $Y = 1$ é raro $\Rightarrow \mathbb{P}(Y = 1|\mathbf{x})$ baixo.

$g(\mathbf{x}) = \mathbb{I}(\mathbb{P}(Y = 1|\mathbf{x}) \geq 1/2) = 0$ para quase todo $\mathbf{x}$

Evitar isso: buscar cortes diferentes de $1/2$:

$$g(\mathbf{x}) = \mathbb{I}(\mathbb{P}(Y = 1|\mathbf{x}) \geq K)$$

Problema relacionado: $Y = 1$ é raro $\Rightarrow \mathbb{P}(Y = 1|\mathbf{x})$ baixo.

$g(\mathbf{x}) = \mathbb{I}(\mathbb{P}(Y = 1|\mathbf{x}) \geq 1/2) = 0$ para quase todo $\mathbf{x}$

Evitar isso: buscar cortes diferentes de $1/2$:

$$g(\mathbf{x}) = \mathbb{I}(\mathbb{P}(Y = 1|\mathbf{x}) \geq K)$$

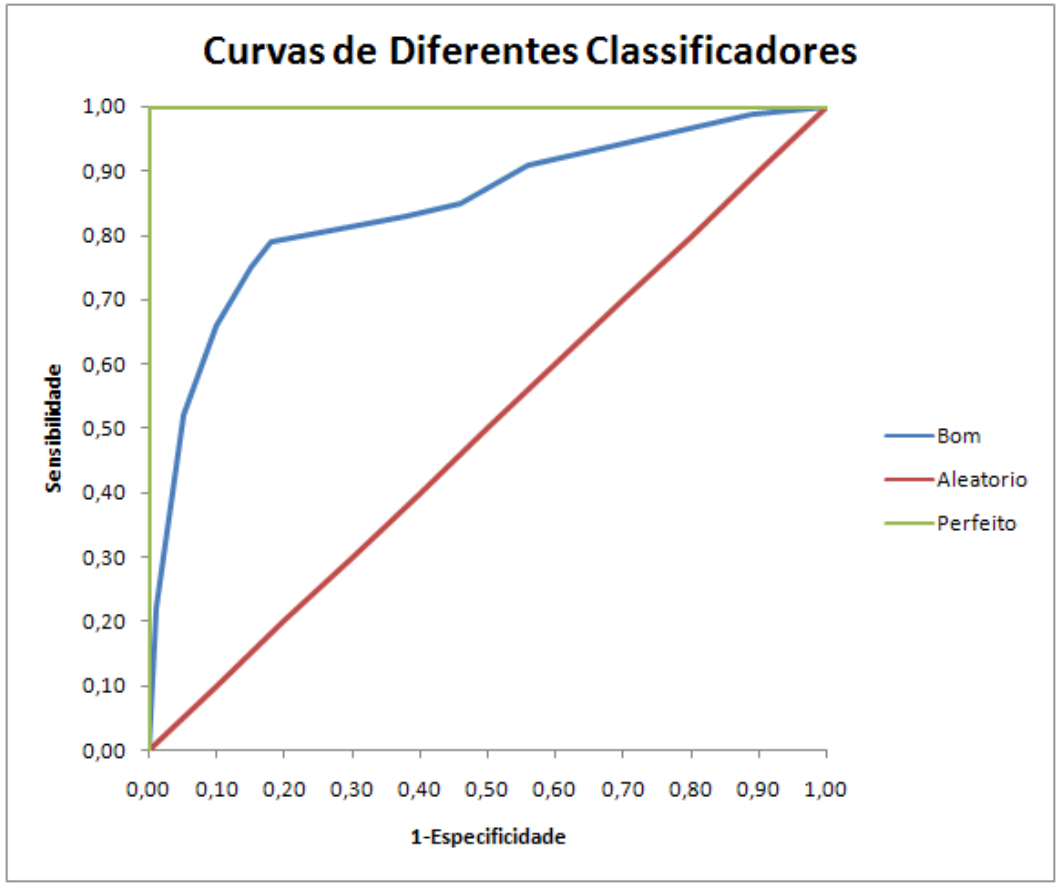
Problema relacionado: $Y = 1$ é raro $\Rightarrow \mathbb{P}(Y = 1|\mathbf{x})$ baixo.

$g(\mathbf{x}) = \mathbb{I}(\mathbb{P}(Y = 1|\mathbf{x}) \geq 1/2) = 0$ para quase todo $\mathbf{x}$

Evitar isso: buscar cortes diferentes de $1/2$:

$$g(\mathbf{x}) = \mathbb{I}(\mathbb{P}(Y = 1|\mathbf{x}) \geq K)$$

Curva ROC


Curvas de Diferentes Classificadores
Sensibilidade
1,00
0,90
0,80
0,70
0,60
0,50
0,40
0,30
0,20
0,10
0,00
0,00 0,10 0,20 0,30 0,40 0,50 0,60 0,70 0,80 0,90 1,00
1-Especificidade
Bom
Aleatorio
Perfeito

Outras funções de perda

π_0: $\mathbb{P}(Y = 0)$
π_1: $\mathbb{P}(Y = 1)$

$$R'(g) = \mathbb{E}[(\pi_1 \mathbb{I}(Y \neq g(\mathbf{X}) \text{ e } Y = 0)) + (\pi_0 \mathbb{I}(Y \neq g(\mathbf{X}) \text{ e } Y = 1))]$$

Função $g(\mathbf{x})$ que minimiza $R'(g)$:

$$g(\mathbf{x}) = \mathbb{I}(\mathbb{P}(Y = 1|\mathbf{x}) > \pi_1).$$

Outras funções de perda

π_0: $\mathbb{P}(Y = 0)$
π_1: $\mathbb{P}(Y = 1)$

$$R'(g) = \mathbb{E}[(\pi_1 \mathbb{I}(Y \neq g(\mathbf{X}) \text{ e } Y = 0)) + (\pi_0 \mathbb{I}(Y \neq g(\mathbf{X}) \text{ e } Y = 1))]$$

Função $g(\mathbf{x})$ que minimiza $R'(g)$:

$$g(\mathbf{x}) = \mathbb{I}(\mathbb{P}(Y = 1|\mathbf{x}) > \pi_1).$$

Outras funções de perda

π_0: $\mathbb{P}(Y = 0)$
π_1: $\mathbb{P}(Y = 1)$

$$R'(g) = \mathbb{E}[(\pi_1 \mathbb{I}(Y \neq g(\mathbf{X}) \text{ e } Y = 0)) + (\pi_0 \mathbb{I}(Y \neq g(\mathbf{X}) \text{ e } Y = 1))]$$

Função $g(\mathbf{x})$ que minimiza $R'(g)$:

$$g(\mathbf{x}) = \mathbb{I}(\mathbb{P}(Y = 1|\mathbf{x}) > \pi_1).$$

Revisão

- Risco para classificação: probabilidade de erro
- $g(\mathbf{x}) = \mathbb{I}(\mathbb{P}(Y = 1|\mathbf{x}) > K)$
- $\mathbb{P}(Y = 1|\mathbf{x})$ pode ser estimada de várias formas

Revisão

- Risco para classificação: probabilidade de erro
- $g(\mathbf{x}) = \mathbb{I}(\mathbb{P}(Y = 1|\mathbf{x}) > K)$
- $\mathbb{P}(Y = 1|\mathbf{x})$ pode ser estimada de várias formas

Revisão

- Risco para classificação: probabilidade de erro
- $g(\mathbf{x}) = \mathbb{I}(\mathbb{P}(Y = 1|\mathbf{x}) > K)$
- $\mathbb{P}(Y = 1|\mathbf{x})$ pode ser estimada de várias formas

Nem todo classificador é baseado em estimar $\mathbb{P}(Y = 1|\mathbf{x})$

- KNN
- Árvores
- Florestas em Regressão
- SVM, ...

Nem todo classificador é baseado em estimar $\mathbb{P}(Y = 1|\mathbf{x})$

- KNN
- Árvores
- Florestas em Regressão
- SVM, ...

Nem todo classificador é baseado em estimar $\mathbb{P}(Y = 1|\mathbf{x})$

- KNN
- Árvores
- Florestas em Regressão
- SVM, ...

Nem todo classificador é baseado em estimar $\mathbb{P}(Y = 1|\mathbf{x})$

- KNN
- Árvores
- Florestas em Regressão
- SVM, ...

Resumo: conceitos vistos

- Aprendizado supervisionado: classificação e regressão
- Risco: quantifica o poder preditivo de uma função
- Métodos de classificação/regressão: KNN, Árvores, Florestas, Lasso
- Ideia dos métodos: controlar balanço viés-variância; overfitting vs underfitting

Resumo: conceitos vistos

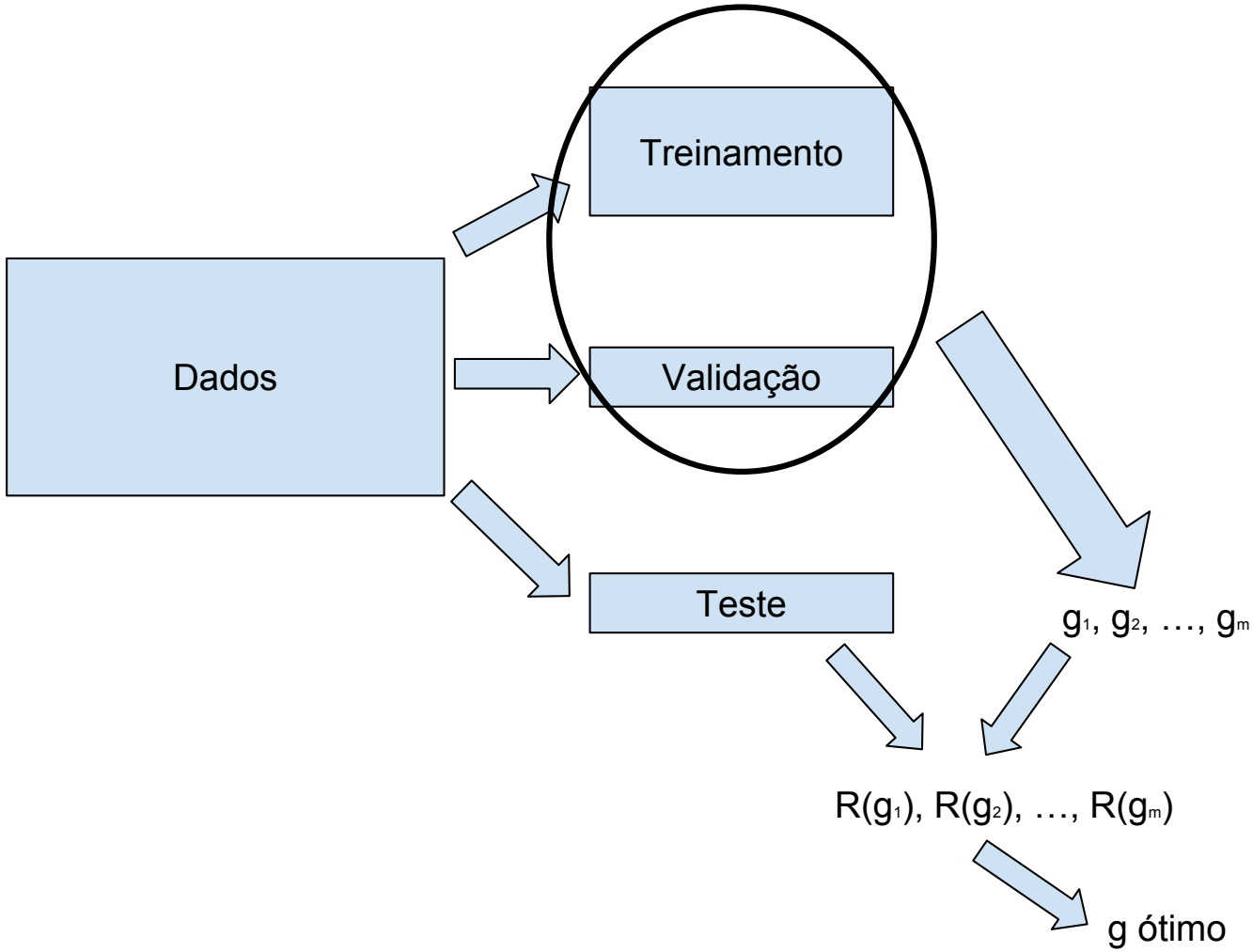
- Aprendizado supervisionado: classificação e regressão
- Risco: quantifica o poder preditivo de uma função
- Métodos de classificação/regressão: KNN, Árvores, Florestas, Lasso
- Ideia dos métodos: controlar balanço viés-variância; overfitting vs underfitting

Resumo: conceitos vistos

- Aprendizado supervisionado: classificação e regressão
- Risco: quantifica o poder preditivo de uma função
- Métodos de classificação/regressão: KNN, Árvores, Florestas, Lasso
- Ideia dos métodos: controlar balanço viés-variância; overfitting vs underfitting

Resumo: conceitos vistos

- Aprendizado supervisionado: classificação e regressão
- Risco: quantifica o poder preditivo de uma função
- Métodos de classificação/regressão: KNN, Árvores, Florestas, Lasso
- Ideia dos métodos: controlar balanço viés-variância; overfitting vs underfitting


Treinamento
Dados
Validação
Teste
$g_1, g_2, \ldots, g_m$
$R(g_1), R(g_2), \ldots, R(g_m)$
g ótimo

O que mais há por aí?

- SVM, Boosting, Redes Neurais (Profundas), RKHS e Truque do Kernel
- Mudança de domínio/conceito
- Interpretação de classificadores
- Métodos semi-supervisionados
- Active learning
- Aprendizado não supervisionado

Literatura de ML; JMLR

O que mais há por aí?

- SVM, Boosting, Redes Neurais (Profundas), RKHS e Truque do Kernel
- Mudança de domínio/conceito
- Interpretação de classificadores
- Métodos semi-supervisionados
- Active learning
- Aprendizado não supervisionado

Literatura de ML; JMLR

O que mais há por aí?

- SVM, Boosting, Redes Neurais (Profundas), RKHS e Truque do Kernel
- Mudança de domínio/conceito
- Interpretação de classificadores
- Métodos semi-supervisionados
- Active learning
- Aprendizado não supervisionado

Literatura de ML; JMLR

O que mais há por aí?

- SVM, Boosting, Redes Neurais (Profundas), RKHS e Truque do Kernel
- Mudança de domínio/conceito
- Interpretação de classificadores
- Métodos semi-supervisionados
- Active learning
- Aprendizado não supervisionado

Literatura de ML; JMLR

O que mais há por aí?

- SVM, Boosting, Redes Neurais (Profundas), RKHS e Truque do Kernel
- Mudança de domínio/conceito
- Interpretação de classificadores
- Métodos semi-supervisionados
- Active learning
- Aprendizado não supervisionado

Literatura de ML; JMLR

O que mais há por aí?

- SVM, Boosting, Redes Neurais (Profundas), RKHS e Truque do Kernel
- Mudança de domínio/conceito
- Interpretação de classificadores
- Métodos semi-supervisionados
- Active learning
- Aprendizado não supervisionado

Literatura de ML; JMLR

O que mais há por aí?

- SVM, Boosting, Redes Neurais (Profundas), RKHS e Truque do Kernel
- Mudança de domínio/conceito
- Interpretação de classificadores
- Métodos semi-supervisionados
- Active learning
- Aprendizado não supervisionado

Literatura de ML; JMLR

Obrigado!

rafaelizbicki@gmail.com

`http://rizbicki.ufscar.br/sml`