Mineração de Dados

Aula 2: Regressão Linear e Seleção de Modelos

Rafael Izbicki

Resumindo até agora

Em um problema de predição:

- Observamos um conjunto de treinamento $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$. $\mathbf{X}$ são chamados de preditores, ou variáveis explicativas, ou variáveis independentes, ou *features*. Y é chamado de resposta, ou 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 significa criar g tal que o risco quadrático $R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$ seja baixo (dependendo da situação, outras funções de risco podem ser mais adequadas; veremos mais sobre isso depois)

Vamos ver ao longo do curso como, em diferentes contextos, criar g tal que $R(g)$ seja baixo, e como comparar diferentes g's.

Resumindo até agora

Em um problema de predição:

- Observamos um conjunto de treinamento $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$. $\mathbf{X}$ são chamados de preditores, ou variáveis explicativas, ou variáveis independentes, ou *features*. Y é chamado de resposta, ou 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 significa criar g tal que o risco quadrático $R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$ seja baixo (dependendo da situação, outras funções de risco podem ser mais adequadas; veremos mais sobre isso depois)

Vamos ver ao longo do curso como, em diferentes contextos, criar g tal que $R(g)$ seja baixo, e como comparar diferentes g's.

Resumindo até agora

Em um problema de predição:

- Observamos um conjunto de treinamento $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$. $\mathbf{X}$ são chamados de preditores, ou variáveis explicativas, ou variáveis independentes, ou *features*. Y é chamado de resposta, ou 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 significa criar g tal que o risco quadrático $R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$ seja baixo (dependendo da situação, outras funções de risco podem ser mais adequadas; veremos mais sobre isso depois)

Vamos ver ao longo do curso como, em diferentes contextos, criar g tal que $R(g)$ seja baixo, e como comparar diferentes g's.

Resumindo até agora

Em um problema de predição:

- Observamos um conjunto de treinamento $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$. $\mathbf{X}$ são chamados de preditores, ou variáveis explicativas, ou variáveis independentes, ou *features*. Y é chamado de resposta, ou 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 significa criar g tal que o risco quadrático $R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$ seja baixo (dependendo da situação, outras funções de risco podem ser mais adequadas; veremos mais sobre isso depois)

Vamos ver ao longo do curso como, em diferentes contextos, criar g tal que $R(g)$ seja baixo, e como comparar diferentes g's.

Qual a melhor função $g(\mathbf{x})$?

O problema de encontrar g que minimize $R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$ possui solução analítica!!

Seja $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ a função de regressão (note que não estamos assumindo que ela é linear).

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

O problema então está resolvido? Não!! $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ depende de quantidades desconhecidas!

Contudo, este resultado indica que uma estimativa boa da função de regressão r nos fornece um bom preditor para y com base em $\mathbf{x}$.

Qual a melhor função $g(\mathbf{x})$?

O problema de encontrar g que minimize $R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$ possui solução analítica!!

Seja $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ a função de regressão (note que não estamos assumindo que ela é linear).

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

O problema então está resolvido? Não!! $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ depende de quantidades desconhecidas!

Contudo, este resultado indica que uma estimativa boa da função de regressão r nos fornece um bom preditor para y com base em $\mathbf{x}$.

Qual a melhor função $g(\mathbf{x})$?

O problema de encontrar g que minimize $R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$ possui solução analítica!!

Seja $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ a função de regressão (note que não estamos assumindo que ela é linear).

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

O problema então está resolvido? Não!! $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ depende de quantidades desconhecidas!

Contudo, este resultado indica que uma estimativa boa da função de regressão r nos fornece um bom preditor para y com base em $\mathbf{x}$.

Qual a melhor função $g(\mathbf{x})$?

O problema de encontrar g que minimize $R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$ possui solução analítica!!

Seja $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ a função de regressão (note que não estamos assumindo que ela é linear).

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

O problema então está resolvido? Não!! $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ depende de quantidades desconhecidas!

Contudo, este resultado indica que uma estimativa boa da função de regressão r nos fornece um bom preditor para y com base em $\mathbf{x}$.

Qual a melhor função $g(\mathbf{x})$?

O problema de encontrar g que minimize $R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$ possui solução analítica!!

Seja $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ a função de regressão (note que não estamos assumindo que ela é linear).

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

O problema então está resolvido? Não!! $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ depende de quantidades desconhecidas!

Contudo, este resultado indica que uma estimativa boa da função de regressão r nos fornece um bom preditor para y com base em $\mathbf{x}$.

Uma forma de se criar/estimar g é através de uma regressão linear. Ao longo das próximas aulas veremos algumas metodologias mais sofisticadas para isso.

Antes disso, precisamos entender mais a fundo os elementos de um problema de predição.

Uma forma de se criar/estimar g é através de uma regressão linear. Ao longo das próximas aulas veremos algumas metodologias mais sofisticadas para isso.

Antes disso, precisamos entender mais a fundo os elementos de um problema de predição.

Notação

Resposta	Covariáveis			
Y_1	X_{11}	$\dots$	X_{1d}	$(= \mathbf{X}_1)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	
Y_n	X_{n1}	$\dots$	X_{nd}	$(= \mathbf{X}_n)$

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_{11}	$\dots$	X_{1d}	$(= \mathbf{X}_1)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	
Y_n	X_{n1}	$\dots$	X_{nd}	$(= \mathbf{X}_n)$

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

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

Resultado 1

Seja $(\mathbf{X}, Y)$ um par de variáveis aleatórias (note que $\mathbf{X}$ pode ser um vetor).

A função g que minimiza $R(g) := \mathbb{E}[(g(\mathbf{X}) - Y)^2]$ é dada por $g(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}]$.

Em outras palavras, para $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}]$, temos $R(r) \leq R(g)$ para toda função $g(\mathbf{x})$.

Resultado 1

Seja $(\mathbf{X}, Y)$ um par de variáveis aleatórias (note que $\mathbf{X}$ pode ser um vetor).

A função g que minimiza $R(g) := \mathbb{E}[(g(\mathbf{X}) - Y)^2]$ é dada por $g(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}]$.

Em outras palavras, para $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}]$, temos $R(r) \leq R(g)$ para toda função $g(\mathbf{x})$.

Resultado 1

Seja $(\mathbf{X}, Y)$ um par de variáveis aleatórias (note que $\mathbf{X}$ pode ser um vetor).

A função g que minimiza $R(g) := \mathbb{E}[(g(\mathbf{X}) - Y)^2]$ é dada por $g(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}]$.

Em outras palavras, para $r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}]$, temos $R(r) \leq R(g)$ para toda função $g(\mathbf{x})$.

Resultado 2

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

- $\mathbb{V}[Y|\mathbf{X} = \mathbf{x}]$ é a variância intrínseca da variável resposta, que não depende da função g, assim não pode ser reduzida.
- $(r(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})])^2$ é o viés ao quadrado do estimador $\widehat{g}$
- $\mathbb{V}[g(\mathbf{x})]$ é sua variância

Resultado 2

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

- $\mathbb{V}[Y|\mathbf{X} = \mathbf{x}]$ é a variância intrínseca da variável resposta, que não depende da função g, assim não pode ser reduzida.
- $(r(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})])^2$ é o viés ao quadrado do estimador $\widehat{g}$
- $\mathbb{V}[g(\mathbf{x})]$ é sua variância

Resultado 2

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

- $\mathbb{V}[Y|\mathbf{X} = \mathbf{x}]$ é a variância intrínseca da variável resposta, que não depende da função g, assim não pode ser reduzida.
- $(r(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})])^2$ é o viés ao quadrado do estimador $\widehat{g}$
- $\mathbb{V}[g(\mathbf{x})]$ é sua variância

Resultado 2

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

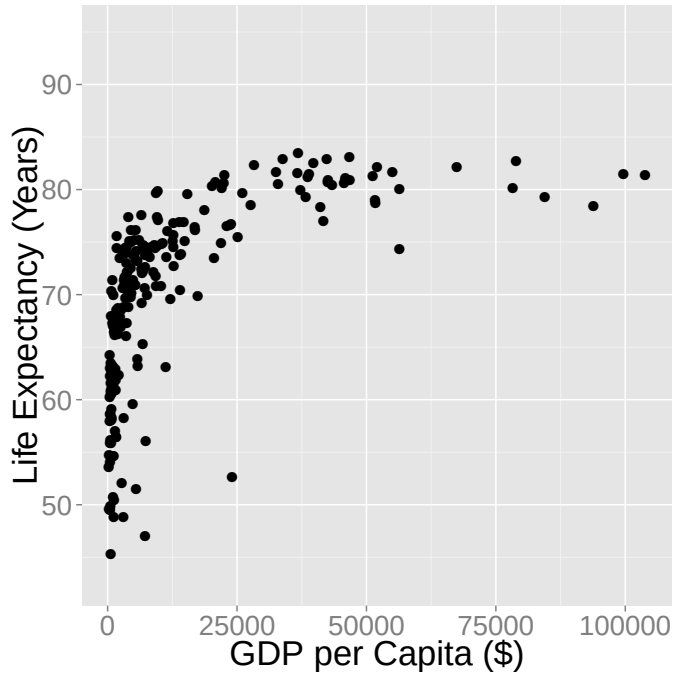
- $\mathbb{V}[Y|\mathbf{X} = \mathbf{x}]$ é a variância intrínseca da variável resposta, que não depende da função g, assim não pode ser reduzida.
- $(r(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})])^2$ é o viés ao quadrado do estimador $\widehat{g}$
- $\mathbb{V}[g(\mathbf{x})]$ é sua variância

Resultado 2

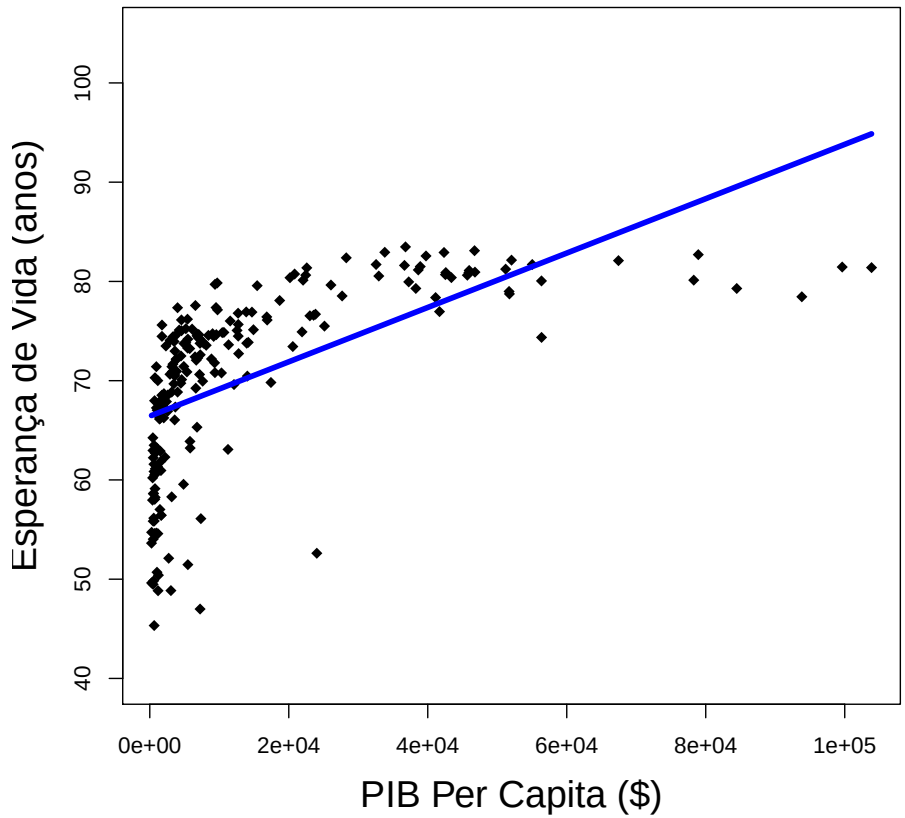
$$\mathbb{E}\left[(Y - g(\mathbf{X}))^2 \middle| \mathbf{X} = \mathbf{x}\right] = \mathbb{V}[Y|\mathbf{X} = \mathbf{x}] + (r(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})])^2 + \mathbb{V}[g(\mathbf{x})]$$

- $\mathbb{V}[Y|\mathbf{X} = \mathbf{x}]$ é a variância intrínseca da variável resposta, que não depende da função g, assim não pode ser reduzida.
- $(r(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})])^2$ é o viés ao quadrado do estimador $\widehat{g}$
- $\mathbb{V}[g(\mathbf{x})]$ é sua variância

Exemplo de Regressão Linear



Exemplo de Regressão Linear



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

Neste caso, péssimo poder preditivo!!

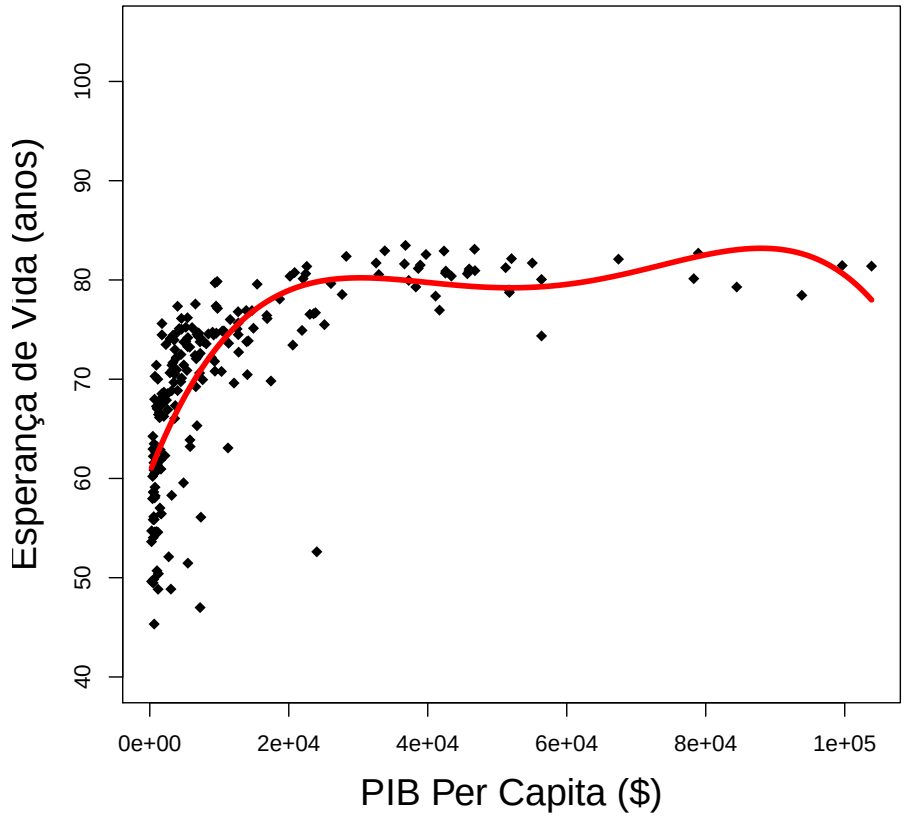
Podemos usar a metodologia de regressão linear para ajustar polinômios:

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

Neste caso, péssimo poder preditivo!!

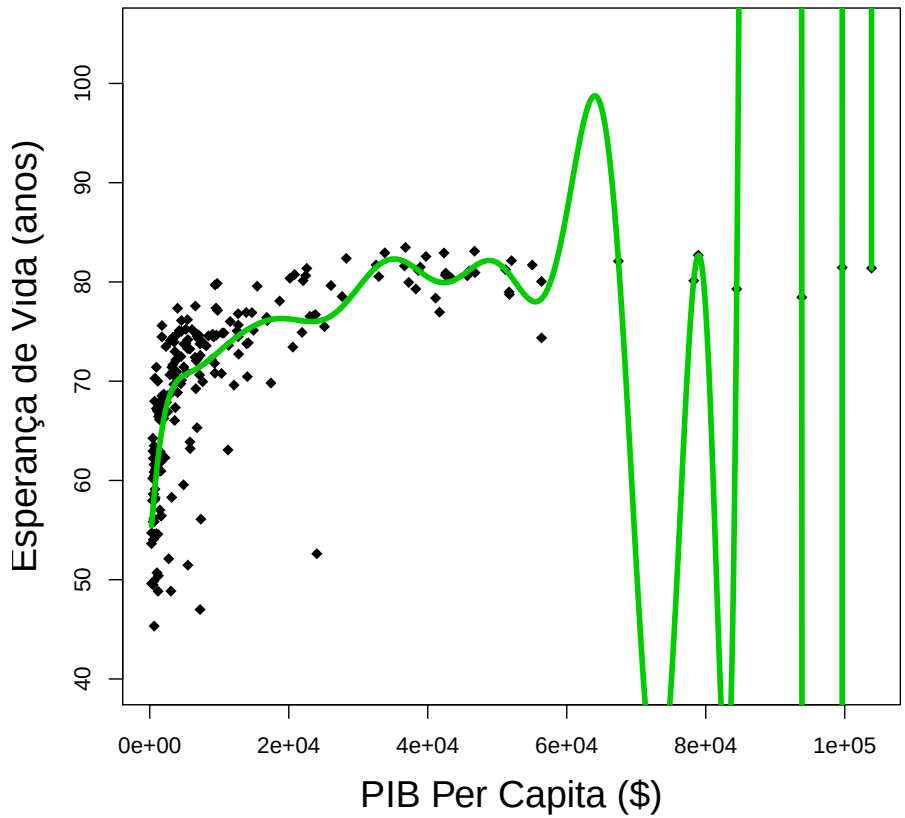
Podemos usar a metodologia de regressão linear para ajustar polinômios:

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



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

Muito melhor! E se colocamos ainda mais parâmetros?



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

Seleção de Modelos: Super e Sub-Ajuste

Seleção de Modelos: Super e Sub-Ajuste

Considere o problema de encontrar a melhor função de predição dentro da classe

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

Qual o melhor p?

$p = 50$: super-ajuste $\Rightarrow$ modelo se ajusta demais ao conjunto de treinamento, e assim possui baixo poder preditivo.

$p = 1$: sub-ajuste $\Rightarrow$ modelo é simplista demais, e assim possui baixo poder preditivo.

Como formalizar isso?

Seleção de Modelos: Super e Sub-Ajuste

Considere o problema de encontrar a melhor função de predição dentro da classe

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

Qual o melhor p?

$p = 50$: super-ajuste $\Rightarrow$ modelo se ajusta demais ao conjunto de treinamento, e assim possui baixo poder preditivo.

$p = 1$: sub-ajuste $\Rightarrow$ modelo é simplista demais, e assim possui baixo poder preditivo.

Como formalizar isso?

Seleção de Modelos: Super e Sub-Ajuste

Considere o problema de encontrar a melhor função de predição dentro da classe

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

Qual o melhor p?

$p = 50$: super-ajuste $\Rightarrow$ modelo se ajusta demais ao conjunto de treinamento, e assim possui baixo poder preditivo.

$p = 1$: sub-ajuste $\Rightarrow$ modelo é simplista demais, e assim possui baixo poder preditivo.

Como formalizar isso?

Seleção de Modelos: Super e Sub-Ajuste

Considere o problema de encontrar a melhor função de predição dentro da classe

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

Qual o melhor p?

$p = 50$: super-ajuste $\Rightarrow$ modelo se ajusta demais ao conjunto de treinamento, e assim possui baixo poder preditivo.

$p = 1$: sub-ajuste $\Rightarrow$ modelo é simplista demais, e assim possui baixo poder preditivo.

Como formalizar isso?

Seleção de Modelos: Super e Sub-Ajuste

Considere o problema de encontrar a melhor função de predição dentro da classe

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

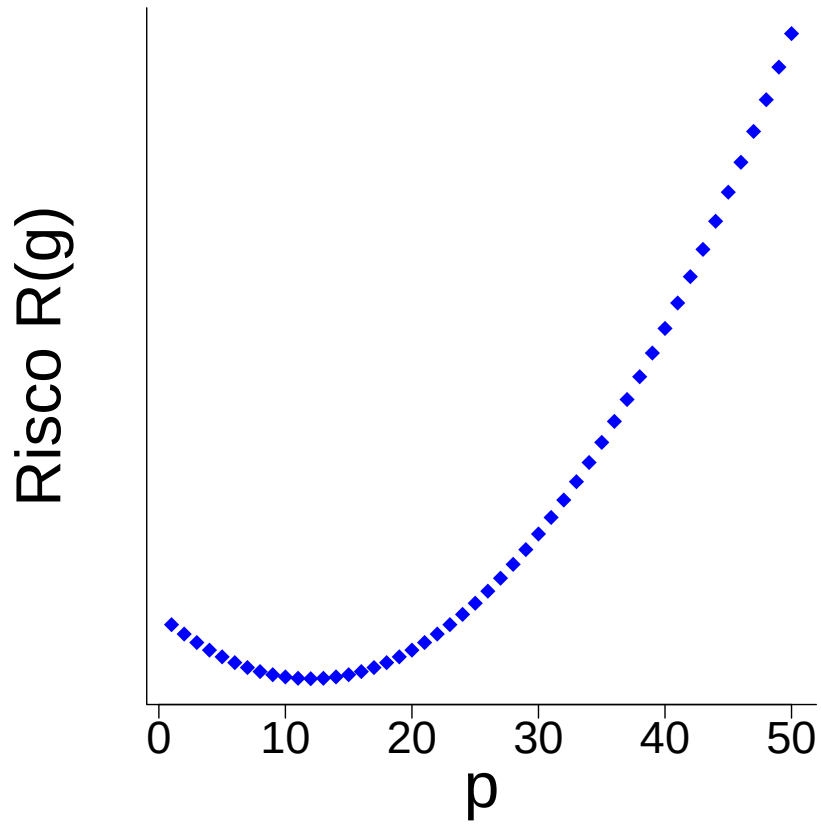
Qual o melhor p?

$p = 50$: super-ajuste $\Rightarrow$ modelo se ajusta demais ao conjunto de treinamento, e assim possui baixo poder preditivo.

$p = 1$: sub-ajuste $\Rightarrow$ modelo é simplista demais, e assim possui baixo poder preditivo.

Como formalizar isso?

Se conseguissemos calcular $R(g) = \mathbb{E}[(Y - g(X))^2]$ para cada p, obteriamos



A ideia para escolher p é estimar $R(g)$ para tentar reproduzir esta figura.

Como estimar $R(g)$?

Como estimar $R(g)$?

Estimar o risco $R(g)$ é importante para poder comparar diferentes candidatos a função de predição $g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots$.

Como podemos estimar o risco $R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$ de uma função de predição $g(\mathbf{x})$? Será que é razoável usar

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

EQM: Erro quadrático médio na amostra de treinamento

Como estimar $R(g)$?

Estimar o risco $R(g)$ é importante para poder comparar diferentes candidatos a função de predição $g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots$.

Como podemos estimar o risco $R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$ de uma função de predição $g(\mathbf{x})$? Será que é razoável usar

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

EQM: Erro quadrático médio na amostra de treinamento

Como estimar $R(g)$?

Estimar o risco $R(g)$ é importante para poder comparar diferentes candidatos a função de predição $g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots$.

Como podemos estimar o risco $R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$ de uma função de predição $g(\mathbf{x})$? Será que é razoável usar

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

EQM: Erro quadrático médio na amostra de treinamento

Como estimar $R(g)$?

Estimar o risco $R(g)$ é importante para poder comparar diferentes candidatos a função de predição $g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots$.

Como podemos estimar o risco $R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$ de uma função de predição $g(\mathbf{x})$? Será que é razoável usar

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

EQM: Erro quadrático médio na amostra de treinamento

Em geral, não!!! Isto ocorre porque em geral g é escolhida de modo que $EQM(g)$ seja baixo.

Como estimar $R(g)$?

Este é exatamente o caso de regressão linear, em que os parâmetros são rescolhidos de modo a minimizar o erro quadrático médio

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overbrace{\beta_0 - \beta_1 x_{i,1} - \ldots - \beta_p x_{i,p}}^{g_\beta(X_i)})^2$$

(i.e., buscamos $g_\beta(x)$ que minimize $EQM(g_\beta)$)

Assim, o erro quadrático médio avaliado no conjunto de treinamento em geral é um estimador muito otimista do erro; em geral o erro é muito maior que o predito por ele.

Leva ao super-ajuste

Como estimar $R(g)$?

Este é exatamente o caso de regressão linear, em que os parâmetros são rescolhidos de modo a minimizar o erro quadrático médio

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overbrace{\beta_0 - \beta_1 x_{i,1} - \ldots - \beta_p x_{i,p}}^{g_\beta(X_i)})^2$$

(i.e., buscamos $g_\beta(x)$ que minimize $EQM(g_\beta)$)

Assim, o erro quadrático médio avaliado no conjunto de treinamento em geral é um estimador muito otimista do erro; em geral o erro é muito maior que o predito por ele.

Leva ao super-ajuste

Como estimar $R(g)$?

Este é exatamente o caso de regressão linear, em que os parâmetros são rescolhidos de modo a minimizar o erro quadrático médio

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overbrace{\beta_0 - \beta_1 x_{i,1} - \ldots - \beta_p x_{i,p}}^{g_\beta(X_i)})^2$$

(i.e., buscamos $g_\beta(x)$ que minimize $EQM(g_\beta)$)

Assim, o erro quadrático médio avaliado no conjunto de treinamento em geral é um estimador muito otimista do erro; em geral o erro é muito maior que o predito por ele.

Leva ao super-ajuste

Como estimar $R(g)$?

Este é exatamente o caso de regressão linear, em que os parâmetros são rescolhidos de modo a minimizar o erro quadrático médio

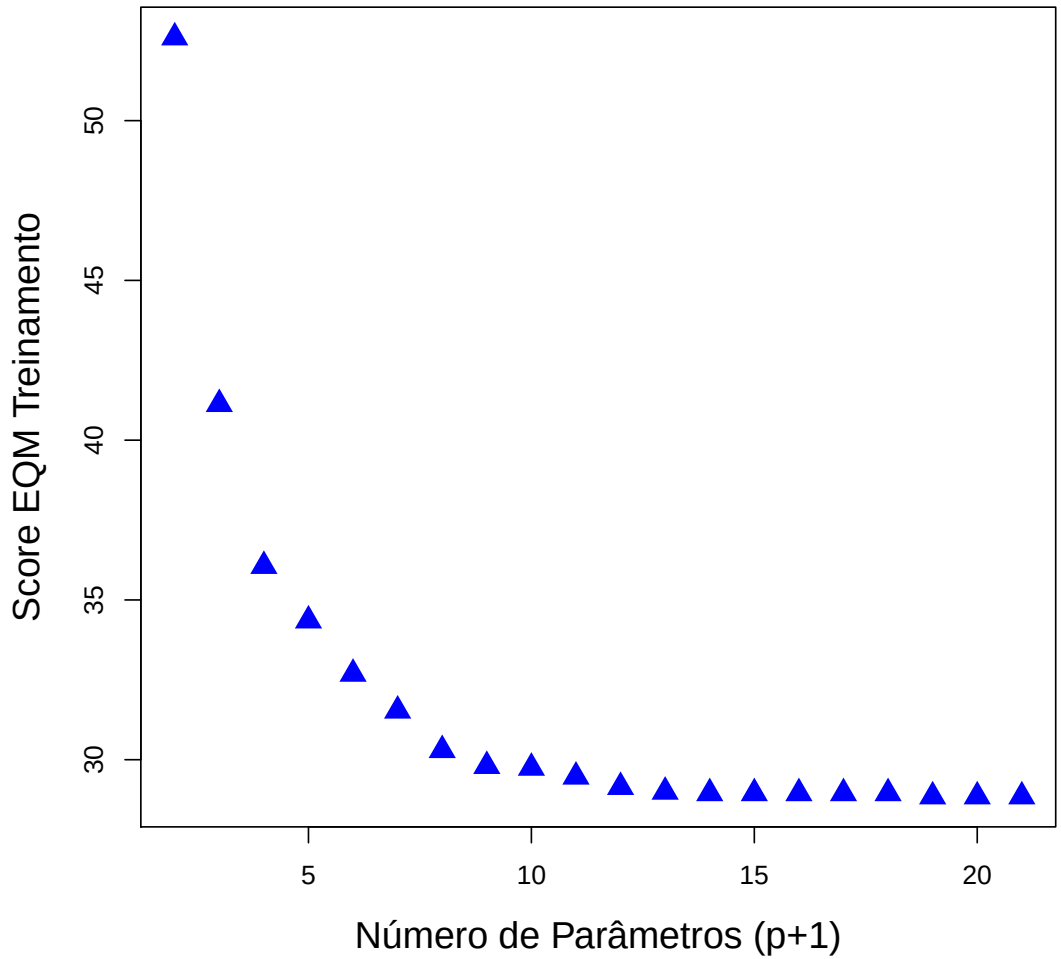
$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overbrace{\beta_0 - \beta_1 x_{i,1} - \ldots - \beta_p x_{i,p}}^{g_\beta(X_i)})^2$$

(i.e., buscamos $g_\beta(x)$ que minimize $EQM(g_\beta)$)

Assim, o erro quadrático médio avaliado no conjunto de treinamento em geral é um estimador muito otimista do erro; em geral o erro é muito maior que o predito por ele.

Leva ao super-ajuste

EQM do conjunto de treinamento no exemplo.



Vamos ver duas formas de estimar o risco:

- Penalização
- Validação Cruzada

Primeira linha: Penalização

Quanto mais parâmetros no modelo, mais $EQM(g)$ subestima $R(g)$. Isto é, maior a diferença entre $EQM(g)$ e $R(g)$.

A ideia é então criar uma medida de complexidade g, $\mathcal{P}(g)$. Por exemplo, $\mathcal{P}(g)$ pode ser quantos parâmetros há no modelo.

Podemos então compensar o quão subestimado $R(g)$ é adicionando estas duas quantidades:

$$R(g) \approx EQM(g) + \mathcal{P}(g)$$

Diversas funções de penalização existem, com diferentes motivações teóricas. Exemplos:

- $\mathcal{P}(g) = 2/n * p * \widehat{\sigma}^2$ (AIC)
- $\mathcal{P}(g) = log(n)/n * p * \widehat{\sigma}^2$ (BIC)

onde p é o número de parâmetros de g e $\widehat{\sigma}^2$ é uma estimativa de $\mathbb{V}[Y|x]$.

Primeira linha: Penalização

Quanto mais parâmetros no modelo, mais $EQM(g)$ subestima $R(g)$. Isto é, maior a diferença entre $EQM(g)$ e $R(g)$.

A ideia é então criar uma medida de complexidade g, $\mathcal{P}(g)$. Por exemplo, $\mathcal{P}(g)$ pode ser quantos parâmetros há no modelo.

Podemos então compensar o quão subestimado $R(g)$ é adicionando estas duas quantidades:

$$R(g) \approx EQM(g) + \mathcal{P}(g)$$

Diversas funções de penalização existem, com diferentes motivações teóricas. Exemplos:

- $\mathcal{P}(g) = 2/n * p * \widehat{\sigma}^2$ (AIC)
- $\mathcal{P}(g) = log(n)/n * p * \widehat{\sigma}^2$ (BIC)

onde p é o número de parâmetros de g e $\widehat{\sigma}^2$ é uma estimativa de $\mathbb{V}[Y|x]$.

Primeira linha: Penalização

Quanto mais parâmetros no modelo, mais $EQM(g)$ subestima $R(g)$. Isto é, maior a diferença entre $EQM(g)$ e $R(g)$.

A ideia é então criar uma medida de complexidade g, $\mathcal{P}(g)$. Por exemplo, $\mathcal{P}(g)$ pode ser quantos parâmetros há no modelo.

Podemos então compensar o quão subestimado $R(g)$ é adicionando estas duas quantidades:

$$R(g) \approx EQM(g) + \mathcal{P}(g)$$

Diversas funções de penalização existem, com diferentes motivações teóricas. Exemplos:

- $\mathcal{P}(g) = 2/n * p * \widehat{\sigma}^2$ (AIC)
- $\mathcal{P}(g) = log(n)/n * p * \widehat{\sigma}^2$ (BIC)

onde p é o número de parâmetros de g e $\widehat{\sigma}^2$ é uma estimativa de $\mathbb{V}[Y|x]$.

Primeira linha: Penalização

Quanto mais parâmetros no modelo, mais $EQM(g)$ subestima $R(g)$. Isto é, maior a diferença entre $EQM(g)$ e $R(g)$.

A ideia é então criar uma medida de complexidade g, $\mathcal{P}(g)$. Por exemplo, $\mathcal{P}(g)$ pode ser quantos parâmetros há no modelo.

Podemos então compensar o quão subestimado $R(g)$ é adicionando estas duas quantidades:

$$R(g) \approx EQM(g) + \mathcal{P}(g)$$

Diversas funções de penalização existem, com diferentes motivações teóricas. Exemplos:

- $\mathcal{P}(g) = 2/n * p * \widehat{\sigma}^2$ (AIC)
- $\mathcal{P}(g) = log(n)/n * p * \widehat{\sigma}^2$ (BIC)

onde p é o número de parâmetros de g e $\widehat{\sigma}^2$ é uma estimativa de $\mathbb{V}[Y|x]$.

Primeira linha: Penalização

Quanto mais parâmetros no modelo, mais $EQM(g)$ subestima $R(g)$. Isto é, maior a diferença entre $EQM(g)$ e $R(g)$.

A ideia é então criar uma medida de complexidade g, $\mathcal{P}(g)$. Por exemplo, $\mathcal{P}(g)$ pode ser quantos parâmetros há no modelo.

Podemos então compensar o quão subestimado $R(g)$ é adicionando estas duas quantidades:

$$R(g) \approx EQM(g) + \mathcal{P}(g)$$

Diversas funções de penalização existem, com diferentes motivações teóricas. Exemplos:

- $\mathcal{P}(g) = 2/n * p * \widehat{\sigma}^2$ (AIC)
- $\mathcal{P}(g) = log(n)/n * p * \widehat{\sigma}^2$ (BIC)

onde p é o número de parâmetros de g e $\widehat{\sigma}^2$ é uma estimativa de $\mathbb{V}[Y|x]$.

Primeira linha: Penalização

A ideia é escolher a função g da classe $\mathbb{G}$ que minimize

$$EQM(g) + \mathcal{P}(g)$$

- Se g tem muitos parâmetros, $EQM(g)$ é em geral muito baixo (overfitting), mas em compensação $\mathcal{P}(g)$ é alto.
- Se g tem poucos parâmetros, $EQM(g)$ é em geral muito alto (underfitting), mas em compensação $\mathcal{P}(g)$ é baixo.

Assim, o modelo que minimiza $EQM(g) + \mathcal{P}(g)$ deve ter bom ajuste ($EQM(g)$ baixo) e poucos parâmetros ($\mathcal{P}(g)$ baixo), de forma que não haja nem overfitting nem underfitting.

Primeira linha: Penalização

A ideia é escolher a função g da classe $\mathbb{G}$ que minimize

$$EQM(g) + \mathcal{P}(g)$$

- Se g tem muitos parâmetros, $EQM(g)$ é em geral muito baixo (overfitting), mas em compensação $\mathcal{P}(g)$ é alto.
- Se g tem poucos parâmetros, $EQM(g)$ é em geral muito alto (underfitting), mas em compensação $\mathcal{P}(g)$ é baixo.

Assim, o modelo que minimiza $EQM(g) + \mathcal{P}(g)$ deve ter bom ajuste ($EQM(g)$ baixo) e poucos parâmetros ($\mathcal{P}(g)$ baixo), de forma que não haja nem overfitting nem underfitting.

Primeira linha: Penalização

A ideia é escolher a função g da classe $\mathbb{G}$ que minimize

$$EQM(g) + \mathcal{P}(g)$$

- Se g tem muitos parâmetros, $EQM(g)$ é em geral muito baixo (overfitting), mas em compensação $\mathcal{P}(g)$ é alto.
- Se g tem poucos parâmetros, $EQM(g)$ é em geral muito alto (underfitting), mas em compensação $\mathcal{P}(g)$ é baixo.

Assim, o modelo que minimiza $EQM(g) + \mathcal{P}(g)$ deve ter bom ajuste ($EQM(g)$ baixo) e poucos parâmetros ($\mathcal{P}(g)$ baixo), de forma que não haja nem overfitting nem underfitting.

Primeira linha: Penalização

A ideia é escolher a função g da classe $\mathbb{G}$ que minimize

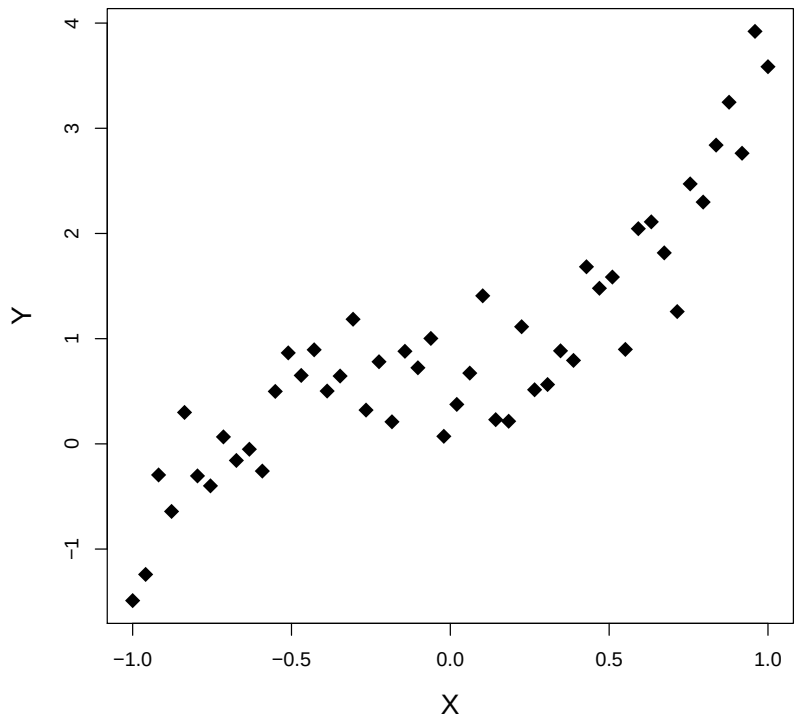
$$EQM(g) + \mathcal{P}(g)$$

- Se g tem muitos parâmetros, $EQM(g)$ é em geral muito baixo (overfitting), mas em compensação $\mathcal{P}(g)$ é alto.
- Se g tem poucos parâmetros, $EQM(g)$ é em geral muito alto (underfitting), mas em compensação $\mathcal{P}(g)$ é baixo.

Assim, o modelo que minimiza $EQM(g) + \mathcal{P}(g)$ deve ter bom ajuste ($EQM(g)$ baixo) e poucos parâmetros ($\mathcal{P}(g)$ baixo), de forma que não haja nem overfitting nem underfitting.

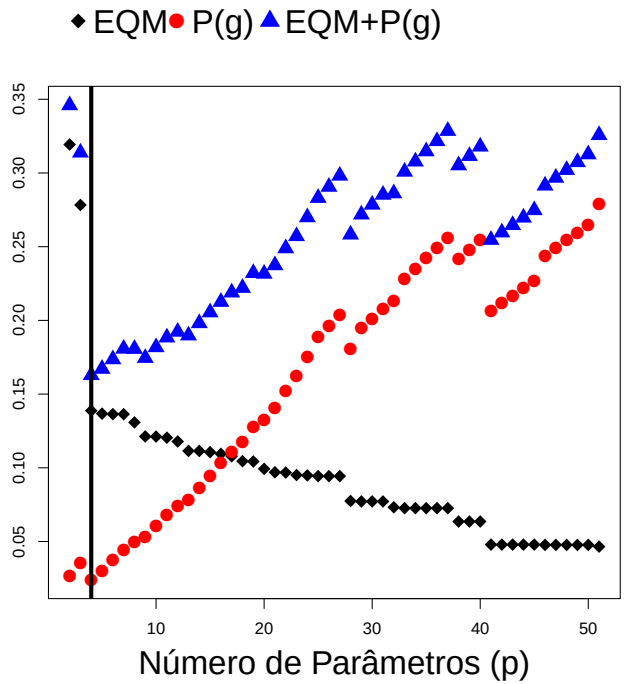
Primeira linha: Penalização

Exemplo.



Primeira linha: Penalização

Exemplo.

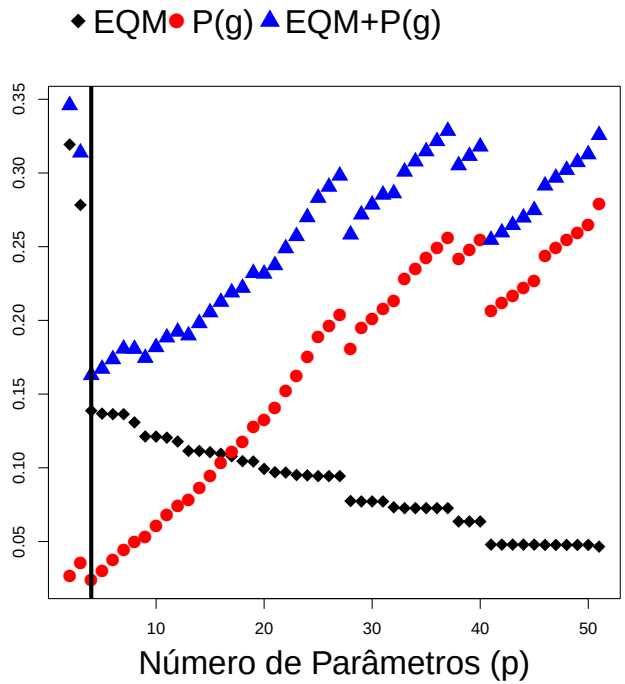


O gráfico mostra o EQM, a penalidade e o AIC calculado para $g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \ldots + \widehat{\beta}_{p-1} x^{p-1}$ com $p \in \{2, 3, \ldots, 50\}$.

O melhor modelo segundo este critério é obtido tomando-se $p = 4$ (linha vertical no gráfico).

Primeira linha: Penalização

Exemplo.

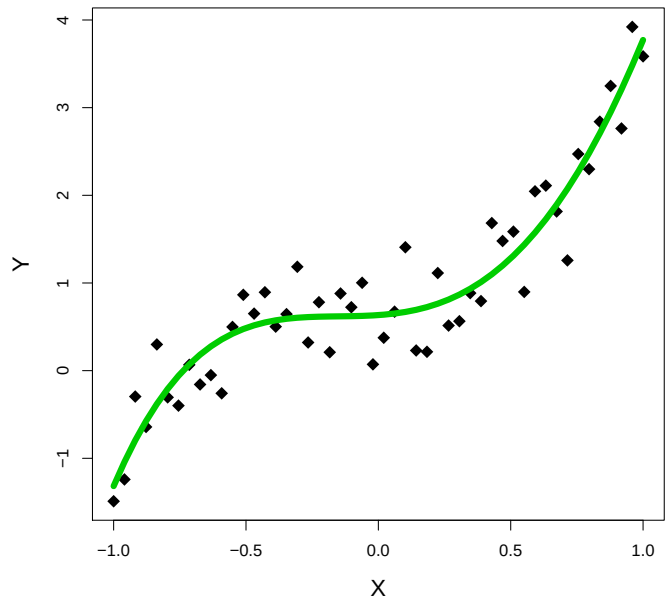


O gráfico mostra o EQM, a penalidade e o AIC calculado para $g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \ldots + \widehat{\beta}_{p-1} x^{p-1}$ com $p \in \{2, 3, \ldots, 50\}$.

O melhor modelo segundo este critério é obtido tomando-se $p = 4$ (linha vertical no gráfico).

Primeira linha: Penalização

Modelo ajustado com $p = 4$ (regressão cúbica):



Resumo: A ideia de métodos de seleção de modelos com base em penalização é a de minimizar uma medida de bondade de ajuste (nos exemplos que vimos, EQM) mais uma penalização por complexidade do modelo (no exemplo vimos o AIC).

Assim, buscamos um modelo que ao mesmo tempo modelo bem os dados, mas que não tenha mais parâmetros que o necessário.
Deste modo, espera-se que tal modelo possua bom poder preditivo.

Resumo: A ideia de métodos de seleção de modelos com base em penalização é a de minimizar uma medida de bondade de ajuste (nos exemplos que vimos, EQM) mais uma penalização por complexidade do modelo (no exemplo vimos o AIC).

Assim, buscamos um modelo que ao mesmo tempo modelo bem os dados, mas que não tenha mais parâmetros que o necessário.
Deste modo, espera-se que tal modelo possua bom poder preditivo.

Segunda linha: Validação Cruzada

Ideia: dividimos o conjunto de dados em dois, treinamento e validação:

$$\overbrace{(X_1, Y_1), (X_2, Y_2), \ldots, (X_s, Y_s)}^{\text{Treinamento (e.g., 70\%)}}, \overbrace{(X_{s+1}, Y_{s+1}), \ldots, (X_n, Y_n)}^{\text{Validação (e.g., 30\%)}}$$

Usamos o conjunto de treinamento para estimar g (e.g., estimar os coeficientes da regressão linear), mas usamos o conjunto de validação apenas para estimar $R(g)$ via

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

Como o conjunto de validação não foi usado para estimar os parâmetros de g, o estimador acima é consistente pela lei dos grandes números.

Segunda linha: Validação Cruzada

Ideia: dividimos o conjunto de dados em dois, treinamento e validação:

$$\overbrace{(X_1, Y_1), (X_2, Y_2), \ldots, (X_s, Y_s)}^{\text{Treinamento (e.g., 70\%)}}, \overbrace{(X_{s+1}, Y_{s+1}), \ldots, (X_n, Y_n)}^{\text{Validação (e.g., 30\%)}}$$

Usamos o conjunto de treinamento para estimar g (e.g., estimar os coeficientes da regressão linear), mas usamos o conjunto de validação apenas para estimar $R(g)$ via

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

Como o conjunto de validação não foi usado para estimar os parâmetros de g, o estimador acima é consistente pela lei dos grandes números.

Segunda linha: Validação Cruzada

Ideia: dividimos o conjunto de dados em dois, treinamento e validação:

$$\overbrace{(X_1, Y_1), (X_2, Y_2), \ldots, (X_s, Y_s)}^{\text{Treinamento (e.g., 70\%)}}, \overbrace{(X_{s+1}, Y_{s+1}), \ldots, (X_n, Y_n)}^{\text{Validação (e.g., 30\%)}}$$

Usamos o conjunto de treinamento para estimar g (e.g., estimar os coeficientes da regressão linear), mas usamos o conjunto de validação apenas para estimar $R(g)$ via

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

Como o conjunto de validação não foi usado para estimar os parâmetros de g, o estimador acima é consistente pela lei dos grandes números.

Segunda linha: Validação Cruzada

Quando o tamanho amostral é pequeno, dividir o banco em dois pode não ser adequado pois cada conjunto (treinamento e validação) terá tamanho amostra muito pequeno.

Uma alternativa é usar a validação cruzada propriamente dita. Estimamos o risco usando

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

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

$$(X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_n, Y_n)$$

Segunda linha: Validação Cruzada

Quando o tamanho amostral é pequeno, dividir o banco em dois pode não ser adequado pois cada conjunto (treinamento e validação) terá tamanho amostra muito pequeno.

Uma alternativa é usar a validação cruzada propriamente dita. Estimamos o risco usando

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

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

$$(X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_n, Y_n)$$

Segunda linha: Validação Cruzada

Quando o tamanho amostral é pequeno, dividir o banco em dois pode não ser adequado pois cada conjunto (treinamento e validação) terá tamanho amostra muito pequeno.

Uma alternativa é usar a validação cruzada propriamente dita. Estimamos o risco usando

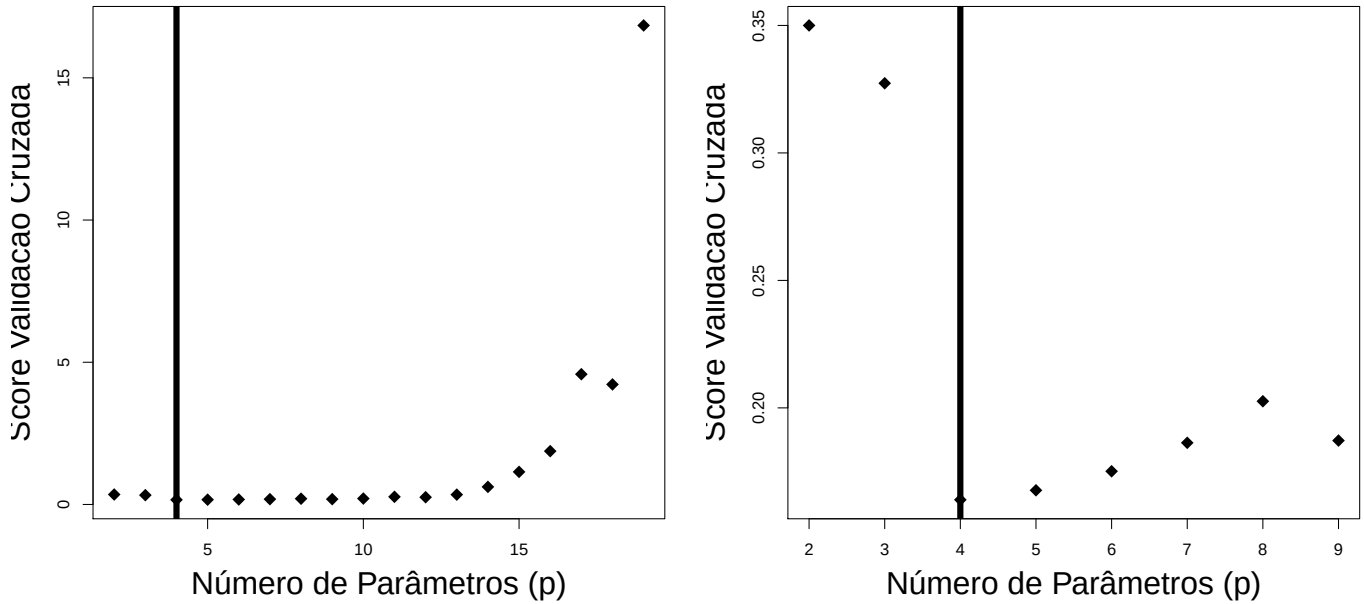
$$R(g) \approx \frac{1}{n}\sum_{i=1}^{n}(Y_i - g_{-i}(X_i))^2$$

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

$$(X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_n, Y_n)$$

Segunda linha: Validação Cruzada

Exemplo.

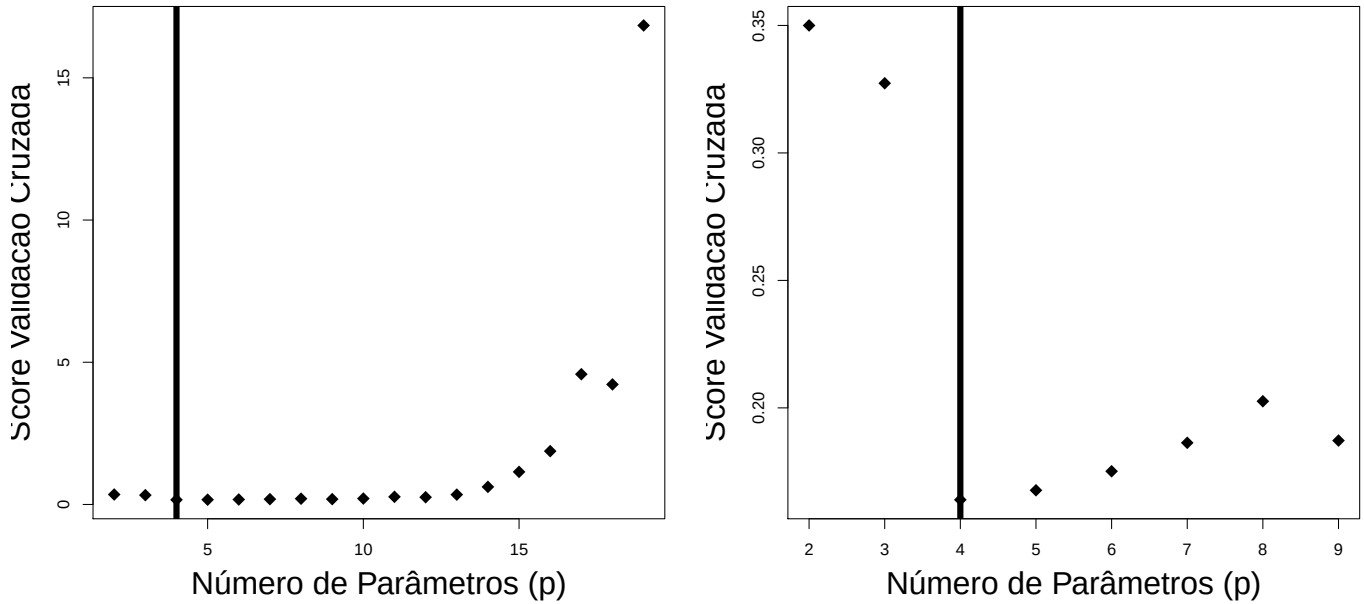


O gráfico mostra o score dado pela validação cruzada para $g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \ldots + \widehat{\beta}_{p-1} x^{p-1}$ com $p \in \{2, 3, \ldots, 50\}$.

Como no caso do AIC, o melhor modelo segundo este critério é obtido tomando-se $p = 4$ (linha vertical no gráfico).

Segunda linha: Validação Cruzada

Exemplo.



O gráfico mostra o score dado pela validação cruzada para $g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \ldots + \widehat{\beta}_{p-1} x^{p-1}$ com $p \in \{2, 3, \ldots, 50\}$.

Como no caso do AIC, o melhor modelo segundo este critério é obtido tomando-se $p = 4$ (linha vertical no gráfico).

Validação Cruzada – Parte II

K-fold cross validation: dividir o banco em k grupos de aproximadamente o mesmo tamanho. Para cada um destes grupos, estimar (treinar) g usando todas as amostras menos as contidas nesse grupo, e estimar as repostas das amostras deixadas de fora. Usar então as respostas preditas para estimar o EQM.

Quando $k = n$, temos o leave-one-out cross validation.

Validação Cruzada – Parte II

K-fold cross validation: dividir o banco em k grupos de aproximadamente o mesmo tamanho. Para cada um destes grupos, estimar (treinar) g usando todas as amostras menos as contidas nesse grupo, e estimar as repostas das amostras deixadas de fora. Usar então as respostas preditas para estimar o EQM.

Quando $k = n$, temos o leave-one-out cross validation.

Resumo da Aula: Vimos que o objetivo de um problema de predição é, com base em um conjunto de treinamento, estimar $g(x)$ tal que, dadas as covariáveis de uma nova observação, x, $g(x)$ seja próxima à resposta desta nova observação, y.

Tal objetivo pode ser traduzido como encontrar g tal que $R(g) = \mathbb{E}[(g(X) - Y)^2)]$ seja baixo.

A solução para tal problema é dada por $g(x) = \mathbb{E}[Y|x]$, ou seja, a melhor solução é dada pela função de regressão de Y em x. Como tal função é desconhecida, podemos estimá-la.

Uma maneira de se fazer isso é através de uma regressão linear. Note que não estamos assumindo que a verdadeira regressão é linear, estamos apenas usando tal método para criar um preditor de Y.

Resumo da Aula: Vimos que o objetivo de um problema de predição é, com base em um conjunto de treinamento, estimar $g(x)$ tal que, dadas as covariáveis de uma nova observação, x, $g(x)$ seja próxima à resposta desta nova observação, y.

Tal objetivo pode ser traduzido como encontrar g tal que $R(g) = \mathbb{E}[(g(X) - Y)^2)]$ seja baixo.

A solução para tal problema é dada por $g(x) = \mathbb{E}[Y|x]$, ou seja, a melhor solução é dada pela função de regressão de Y em x. Como tal função é desconhecida, podemos estimá-la.

Uma maneira de se fazer isso é através de uma regressão linear. Note que não estamos assumindo que a verdadeira regressão é linear, estamos apenas usando tal método para criar um preditor de Y.

Resumo da Aula: Vimos que o objetivo de um problema de predição é, com base em um conjunto de treinamento, estimar $g(x)$ tal que, dadas as covariáveis de uma nova observação, x, $g(x)$ seja próxima à resposta desta nova observação, y.

Tal objetivo pode ser traduzido como encontrar g tal que $R(g) = \mathbb{E}[(g(X) - Y)^2)]$ seja baixo.

A solução para tal problema é dada por $g(x) = \mathbb{E}[Y|x]$, ou seja, a melhor solução é dada pela função de regressão de Y em x. Como tal função é desconhecida, podemos estimá-la.

Uma maneira de se fazer isso é através de uma regressão linear. Note que não estamos assumindo que a verdadeira regressão é linear, estamos apenas usando tal método para criar um preditor de Y.

Resumo da Aula: Vimos que o objetivo de um problema de predição é, com base em um conjunto de treinamento, estimar $g(x)$ tal que, dadas as covariáveis de uma nova observação, x, $g(x)$ seja próxima à resposta desta nova observação, y.

Tal objetivo pode ser traduzido como encontrar g tal que $R(g) = \mathbb{E}[(g(X) - Y)^2)]$ seja baixo.

A solução para tal problema é dada por $g(x) = \mathbb{E}[Y|x]$, ou seja, a melhor solução é dada pela função de regressão de Y em x. Como tal função é desconhecida, podemos estimá-la.

Uma maneira de se fazer isso é através de uma regressão linear. Note que não estamos assumindo que a verdadeira regressão é linear, estamos apenas usando tal método para criar um preditor de Y.

Estudamos também o problema de seleção de modelos: Seja $\mathbb{G}$ um conjunto de funções $g(x)$. Ex:

$$\begin{aligned}\mathbb{G} = \{&g(x) = \widehat{\beta}_0,\\ &g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x,\\ &g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2, \ \ldots\\ &g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2 + \ldots + \widehat{\beta}_{30} x^{30}\}\end{aligned}$$

Como escolher $g \in \mathbb{G}$ com menor risco $R(g)$?

Estudamos também o problema de seleção de modelos: Seja $\mathbb{G}$ um conjunto de funções $g(x)$. Ex:

$$\begin{aligned}\mathbb{G} = \{&g(x) = \widehat{\beta}_0,\\ &g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x,\\ &g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2, \ \ldots\\ &g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2 + \ldots + \widehat{\beta}_{30} x^{30}\}\end{aligned}$$

Como escolher $g \in \mathbb{G}$ com menor risco $R(g)$?

Como escolher $g \in \mathbb{G}$ com menor risco $R(g)$?

Para resolver tal problema é necessário estimar $R(g)$. O erro quadrático médio amostral $EQM(g) := n^{-1} \sum_{i=1}^{n} (Y_i - g(X_i))^2$ é um estimador extremamente viesado $R(g)$: em geral ele subestima tal quantidade. Selecionar modelos com base no EQM leva ao overfitting.

Vimos duas alternativas para estimar $R(g)$:

- penalização: $R(g) \approx EQM(g) + \mathcal{P}(g)$
- validação cruzada: dividir a amostra em treinamento vs validação; estimar $R(g)$ usando somente o conjunto de validação

Próxima Aula: Mais sobre regressão linear e seleção de modelos

Para casa: Ler Seção 2.1 até 2.1.1 (inclusive) do livro texto, assim como a seção 2.2.1.

Como escolher $g \in \mathbb{G}$ com menor risco $R(g)$?

Para resolver tal problema é necessário estimar $R(g)$. O erro quadrático médio amostral $EQM(g) := n^{-1} \sum_{i=1}^{n} (Y_i - g(X_i))^2$ é um estimador extremamente viesado $R(g)$: em geral ele subestima tal quantidade. Selecionar modelos com base no EQM leva ao overfitting.

Vimos duas alternativas para estimar $R(g)$:

- penalização: $R(g) \approx EQM(g) + \mathcal{P}(g)$
- validação cruzada: dividir a amostra em treinamento vs validação; estimar $R(g)$ usando somente o conjunto de validação

Próxima Aula: Mais sobre regressão linear e seleção de modelos

Para casa: Ler Seção 2.1 até 2.1.1 (inclusive) do livro texto, assim como a seção 2.2.1.

Como escolher $g \in \mathbb{G}$ com menor risco $R(g)$?

Para resolver tal problema é necessário estimar $R(g)$. O erro quadrático médio amostral $EQM(g) := n^{-1} \sum_{i=1}^{n} (Y_i - g(X_i))^2$ é um estimador extremamente viesado $R(g)$: em geral ele subestima tal quantidade. Selecionar modelos com base no EQM leva ao overfitting.

Vimos duas alternativas para estimar $R(g)$:

- penalização: $R(g) \approx EQM(g) + \mathcal{P}(g)$
- validação cruzada: dividir a amostra em treinamento vs validação; estimar $R(g)$ usando somente o conjunto de validação

Próxima Aula: Mais sobre regressão linear e seleção de modelos

Para casa: Ler Seção 2.1 até 2.1.1 (inclusive) do livro texto, assim como a seção 2.2.1.

Como escolher $g \in \mathbb{G}$ com menor risco $R(g)$?

Para resolver tal problema é necessário estimar $R(g)$. O erro quadrático médio amostral $EQM(g) := n^{-1} \sum_{i=1}^{n} (Y_i - g(X_i))^2$ é um estimador extremamente viesado $R(g)$: em geral ele subestima tal quantidade. Selecionar modelos com base no EQM leva ao overfitting.

Vimos duas alternativas para estimar $R(g)$:

- penalização: $R(g) \approx EQM(g) + \mathcal{P}(g)$
- validação cruzada: dividir a amostra em treinamento vs validação; estimar $R(g)$ usando somente o conjunto de validação

Próxima Aula: Mais sobre regressão linear e seleção de modelos

Para casa: Ler Seção 2.1 até 2.1.1 (inclusive) do livro texto, assim como a seção 2.2.1.