

# Aprendizado de máquinas

Tiago M. Magalhães

Departamento de Estatística - ICE-UFJF

Juiz de Fora, 19 de junho de 2024



# Roteiro

- 1 Introdução
- 2 Função de risco
- 3 Seleção de Modelos
- 4 Balanço entre viés e variância
- 5 Agradecimentos
- 6 Referências bibliográficas



# Roteiro

- 1 Introdução
- 2 Função de risco
- 3 Seleção de Modelos
- 4 Balanço entre viés e variância
- 5 Agradecimentos
- 6 Referências bibliográficas



# Introdução

De modo geral, o objetivo de um modelo de regressão é determinar a relação entre uma variável aleatória  $Y \in \mathbb{R}$  e um vetor  $\mathbf{x} = (x_1, \dots, x_d)^\top \in \mathbb{R}^p$ .

Mais especificamente, busca-se estimar a função de regressão

$$r(\mathbf{x}) := \mathbb{E}(Y | \mathbf{X} = \mathbf{x}) \quad (1)$$

como forma de descrever tal relação.



# Introdução

De modo geral, o objetivo de um modelo de regressão é determinar a relação entre uma variável aleatória  $Y \in \mathbb{R}$  e um vetor  $\mathbf{x} = (x_1, \dots, x_d)^\top \in \mathbb{R}^p$ .

Mais especificamente, busca-se estimar a função de regressão

$$r(\mathbf{x}) := \mathbb{E}(Y | \mathbf{X} = \mathbf{x}) \quad (1)$$

como forma de descrever tal relação.



# Introdução

Quando  $Y$  é uma variável quantitativa, nós temos um problema de regressão. Para as situações em que  $Y$  é qualitativa nós temos um problema de classificação.



# Introdução

Quando  $Y$  é uma variável quantitativa, nós temos um problema de regressão. Para as situações em que  $Y$  é qualitativa nós temos um problema de classificação.



# Introdução

No entanto, a distinção nem sempre é tão nítida (James et al., 2023, p. 28). Por exemplo, os modelo de regressão normal e logístico fazem parte da classe dos modelos lineares generalizados Nelder e Wedderburn (1972).



# Introdução

O objetivo de (1) pode ser dividido em duas classes:

- **Objetivo inferencial:** Quais preditores são importantes? Qual é a relação entre cada preditor e a variável resposta? Qual é o efeito da mudança de valor de um dos preditores na variável resposta?



# Introdução

O objetivo de (1) pode ser dividido em duas classes:

- **Objetivo inferencial:** Quais preditores são importantes? Qual é a relação entre cada preditor e a variável resposta? Qual é o efeito da mudança de valor de um dos preditores na variável resposta?



# Introdução

- **Objetivo preditivo:** Como podemos criar uma função

$$g : \mathbb{R}^d \longrightarrow \mathbb{R}$$

que tenha bom poder preditivo? Isto é, como criar  $g$  tal que, dadas novas observações independentes e identicamente distribuídas (IID)  $(\mathbf{X}_{n+1}, Y_{n+1}), \dots, (\mathbf{X}_{n+m}, Y_{n+m})$ ,



# Introdução

- **Objetivo preditivo:** Como podemos criar uma função

$$g : \mathbb{R}^d \longrightarrow \mathbb{R}$$

que tenha bom poder preditivo? Isto é, como criar  $g$  tal que, dadas novas observações independentes e identicamente distribuídas (IID)  $(\mathbf{X}_{n+1}, Y_{n+1}), \dots, (\mathbf{X}_{n+m}, Y_{n+m})$ , nós tenhamos

$$g(\mathbf{x}_{n+1}) \approx y_{n+1}, \dots, g(\mathbf{x}_{n+m}) \approx y_{n+m}?$$



# Introdução

- **Objetivo preditivo:** Como podemos criar uma função

$$g : \mathbb{R}^d \longrightarrow \mathbb{R}$$

que tenha bom poder preditivo? Isto é, como criar  $g$  tal que, dadas novas observações independentes e identicamente distribuídas (IID)  $(\mathbf{X}_{n+1}, Y_{n+1}), \dots, (\mathbf{X}_{n+m}, Y_{n+m})$ , nós tenhamos

$$g(\mathbf{x}_{n+1}) \approx y_{n+1}, \dots, g(\mathbf{x}_{n+m}) \approx y_{n+m}?$$



# Introdução

Breiman (2001) argumenta que existem duas culturas no uso de modelos estatísticos (em especial modelos de regressão).



# Introdução

A primeira cultura, *data modeling culture*, é a que domina a comunidade estatística. Em geral, nela se assume que o modelo utilizado para  $r(\mathbf{x})$  é o correto. Pois o principal objetivo está na interpretação dos parâmetros envolvidos no modelo.



# Introdução

A primeira cultura, *data modeling culture*, é a que domina a comunidade estatística. Em geral, nela se assume que o modelo utilizado para  $r(\mathbf{x})$  é o correto. Pois o principal objetivo está na interpretação dos parâmetros envolvidos no modelo.



# Introdução

Em particular há interesse em testes de hipóteses e intervalos de confiança para esses parâmetros. Sob essa abordagem, testar se as suposições do modelo são válidas é de fundamental importância.



# Introdução

Em particular há interesse em testes de hipóteses e intervalos de confiança para esses parâmetros. Sob essa abordagem, testar se as suposições do modelo são válidas é de fundamental importância. Ainda que predição muitas vezes faça parte dos objetivos, o foco em geral está na inferência.



# Introdução

Em particular há interesse em testes de hipóteses e intervalos de confiança para esses parâmetros. Sob essa abordagem, testar se as suposições do modelo são válidas é de fundamental importância. Ainda que predição muitas vezes faça parte dos objetivos, o foco em geral está na inferência.



# Introdução

A segunda cultura, *algorithmic modeling culture*, é a que domina a comunidade de aprendizado de máquina (*machine learning*).

Neste meio, o principal objetivo é a predição de novas observações. Não se assume que o modelo utilizado para os dados é correto.



# Introdução

A segunda cultura, *algorithmic modeling culture*, é a que domina a comunidade de aprendizado de máquina (*machine learning*).

Neste meio, o principal objetivo é a predição de novas observações. Não se assume que o modelo utilizado para os dados é correto.



# Introdução

O modelo é utilizado apenas para criar bons algoritmos para prever bem novas observações.

Muitas vezes não há nenhum modelo probabilístico explícito por trás dos algoritmos utilizados.



# Introdução

O modelo é utilizado apenas para criar bons algoritmos para prever bem novas observações.

Muitas vezes não há nenhum modelo probabilístico explícito por trás dos algoritmos utilizados.



# Roteiro

- 1 Introdução
- 2 Função de risco**
- 3 Seleção de Modelos
- 4 Balanço entre viés e variância
- 5 Agradecimentos
- 6 Referências bibliográficas



# Função de risco

O primeiro passo para construir boas funções de predição é criar um critério para medir o desempenho de uma dada função de predição  $g : \mathbb{R}^d \rightarrow \mathbb{R}$ .



# Função de risco

Em um contexto de regressão, faremos isso através de seu risco quadrático, embora essa não seja a única opção:

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

em que  $(\mathbf{X}, Y)$  é uma nova observação que não foi usada para estimar  $g$ . Quanto menor o risco, melhor é a função de predição  $g$ .



# Função de risco

Em um contexto de regressão, faremos isso através de seu risco quadrático, embora essa não seja a única opção:

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

em que  $(\mathbf{X}, Y)$  é uma nova observação que não foi usada para estimar  $g$ . Quanto menor o risco, melhor é a função de predição  $g$ .



# Função de risco

Quando nós medimos o desempenho de um estimador com base em seu risco quadrático, criar uma boa função de predição  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  equivale a encontrar um bom estimador para a função de regressão  $r(\mathbf{x})$ .



# Função de risco

Assim, responder à pergunta do porquê estimar a função de regressão é, nesse sentido, o melhor caminho para se criar uma função para prever novas observações  $Y$  com base em covariáveis observadas  $x$ . Isso é mostrado no seguinte teorema:



# Função de risco

## Teorema

Suponha que definimos o risco de uma função de predição  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  via perda quadrática:  $R_{\text{pred}}(g) = \mathbb{E} \{ [Y - g(\mathbf{X})]^2 \}$ , em que  $(\mathbf{X}, Y)$  é uma nova observação que não foi usada para estimar  $g$ .



# Função de risco

Suponhamos também que medimos o risco de um estimador da função de regressão via perda quadrática:  $R_{\text{reg}}(g) = \mathbb{E} \{ [r(\mathbf{X}) - g(\mathbf{X})]^2 \}$ . Então

$$R_{\text{pred}}(g) = R_{\text{reg}}(g) + \mathbb{E} [\text{Var}(Y|\mathbf{X})].$$



# Função de risco

Suponhamos também que medimos o risco de um estimador da função de regressão via perda quadrática:  $R_{\text{reg}}(g) = \mathbb{E} \{ [r(\mathbf{X}) - g(\mathbf{X})]^2 \}$ . Então

$$R_{\text{pred}}(g) = R_{\text{reg}}(g) + \mathbb{E} [\text{Var}(Y|\mathbf{X})].$$



# Função de risco

Assim, visto que o termo  $\mathbb{E}[\text{Var}(Y|\mathbf{X})]$  não depende de  $g(\mathbf{x})$ , estimar bem a função de regressão  $r(\mathbf{x})$  é fundamental para se criar uma boa função de predição sob a ótica do risco quadrático, já que a melhor função de predição para  $Y$  é a função de regressão  $r(\mathbf{x})$ :

$$\operatorname{argmin}_g R_{\text{pred}}(g) = \operatorname{argmin}_g R_{\text{reg}}(g) = r(\mathbf{x}).$$

Daqui em diante, nós denotaremos por  $R$  o risco preditivo  $R_{\text{pred}}$ .



# Função de risco

Assim, visto que o termo  $\mathbb{E}[\text{Var}(Y|\mathbf{X})]$  não depende de  $g(\mathbf{x})$ , estimar bem a função de regressão  $r(\mathbf{x})$  é fundamental para se criar uma boa função de predição sob a ótica do risco quadrático, já que a melhor função de predição para  $Y$  é a função de regressão  $r(\mathbf{x})$ :

$$\operatorname{argmin}_g R_{\text{pred}}(g) = \operatorname{argmin}_g R_{\text{reg}}(g) = r(\mathbf{x}).$$

Daqui em diante, nós denotaremos por  $R$  o risco preditivo  $R_{\text{pred}}$ .



# Função de risco

Em um contexto de predição, a definição de risco condicional (isto é, considerando  $g$  fixo) possui grande apelo frequentista. Digamos que nós observamos um novo conjunto,

$$(\mathbf{X}_{n+1}, Y_{n+1}), \dots, (\mathbf{X}_{n+m}, Y_{n+m}),$$

IID à amostra observada.



# Função de risco

Em um contexto de predição, a definição de risco condicional (isto é, considerando  $g$  fixo) possui grande apelo frequentista. Digamos que nós observamos um novo conjunto,

$$(\mathbf{X}_{n+1}, Y_{n+1}), \dots, (\mathbf{X}_{n+m}, Y_{n+m}),$$

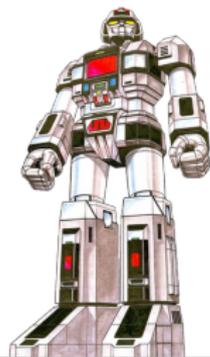
IID à amostra observada.



# Função de risco

Então, pela Lei dos Grandes Números, nós sabemos que, se  $m$  é grande,

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



# Função de risco

Em outras palavras, se  $R(g)$  possui valor baixo, então

$$g(\mathbf{x}_{n+1}) \approx y_{n+1}, \dots, g(\mathbf{x}_{n+m}) \approx y_{n+m},$$

e, portanto, nós teremos boas predições em novas observações.



# Função de risco

Em outras palavras, se  $R(g)$  possui valor baixo, então

$$g(\mathbf{x}_{n+1}) \approx y_{n+1}, \dots, g(\mathbf{x}_{n+m}) \approx y_{n+m},$$

e, portanto, nós teremos boas previsões em novas observações.



# Função de risco

Essa conclusão vale para qualquer função de perda  $L$  utilizada para definir o risco, uma vez que

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



# Função de risco

O objetivo dos métodos de regressão sob a ótica preditivista é, portanto, fornecer, em diversos contextos, métodos que apresentem bons estimadores de  $r(\mathbf{x})$ , isto é, estimadores com risco baixo.



# Roteiro

- 1 Introdução
- 2 Função de risco
- 3 Seleção de Modelos**
- 4 Balanço entre viés e variância
- 5 Agradecimentos
- 6 Referências bibliográficas



# Seleção de Modelos

Em problemas práticos, é comum ajustar vários modelos para a função de regressão  $r(\mathbf{x})$  e buscar qual deles possui maior poder preditivo, isto é, qual possui menor risco.



# Seleção de Modelos

O objetivo de um método de seleção de modelos é selecionar uma boa função  $g$ . Nesse caso, nós utilizaremos o critério do risco quadrático para avaliar a qualidade da função.



# Seleção de Modelos

O objetivo de um método de seleção de modelos é selecionar uma boa função  $g$ . Nesse caso, nós utilizaremos o critério do risco quadrático para avaliar a qualidade da função.

Logo, nós queremos escolher uma função  $g$  dentro de uma classe de candidatos  $\mathcal{G}$  que possua bom poder preditivo (baixo risco quadrático).



# Seleção de Modelos

O objetivo de um método de seleção de modelos é selecionar uma boa função  $g$ . Nesse caso, nós utilizaremos o critério do risco quadrático para avaliar a qualidade da função.

Logo, nós queremos escolher uma função  $g$  dentro de uma classe de candidatos  $\mathbb{G}$  que possua bom poder preditivo (baixo risco quadrático).



# Seleção de Modelos

Dessa forma, nós queremos evitar modelos que sofram de sub ou superajuste. Uma vez que  $R(g)$  é desconhecido, é necessário estimá-lo para avaliar a função  $g \in \mathbb{G}$ .



# Seleção de Modelos

Dessa forma, nós queremos evitar modelos que sofram de sub ou superajuste. Uma vez que  $R(g)$  é desconhecido, é necessário estimá-lo para avaliar a função  $g \in \mathbb{G}$ .



# Seleção de Modelos

O risco observado (também chamado de erro quadrático médio no conjunto de treinamento), definido por

$$\text{EQM}(g) := \frac{1}{n} \sum_{i=1}^n [Y_i - g(\mathbf{x}_i)]^2$$

é um estimador muito otimista do real risco.



# Seleção de Modelos

Se usado para fazer seleção de modelos, ele leva ao super-ajuste, um ajuste perfeito dos dados. Isto ocorre pois  $g$  foi escolhida de modo a ajustar bem  $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ .



# Data Splitting

Uma maneira de solucionar este problema é dividir o conjunto de dados em duas partes, treinamento e validação:

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



## Data Splitting

Nós usamos o conjunto de treinamento exclusivamente para estimar  $g$  e o conjunto de validação apenas para estimar  $R(g)$  via

$$\hat{R}(g) = \frac{1}{n-s} \sum_{i=s+1}^n [Y_i - g(\mathbf{X}_i)]^2 := \hat{R}(g),$$

isto é, nós avaliamos o erro quadrático médio no conjunto de validação.



## *Data Splitting*

Como o conjunto de validação não foi usado para estimar os parâmetros de  $g$ , o estimador da equação acima é consistente pela Lei dos Grandes Números.



# Data Splitting

O procedimento de dividir os dados em dois e utilizar uma parte para estimar o risco é chamado de *data splitting*.



# Validação Cruzada

Uma variação deste método é a validação cruzada, que faz uso de toda a amostra. Por exemplo, no *leave-one-out cross validation* (LOOCV; Stone, 1974), o estimador usado é dado por

$$\hat{R}(g) = \frac{1}{n} \sum_{i=1}^n [Y_i - g_{-i}(\mathbf{x}_i)]^2,$$



# Validação Cruzada

Uma variação deste método é a validação cruzada, que faz uso de toda a amostra. Por exemplo, no *leave-one-out cross validation* (LOOCV; Stone, 1974), o estimador usado é dado por

$$\hat{R}(g) = \frac{1}{n} \sum_{i=1}^n [Y_i - g_{-i}(\mathbf{x}_i)]^2,$$



# Validação Cruzada

em que  $g_{-i}$  é ajustado utilizando-se todas as observações exceto a  $i$ -ésima delas, ou seja, utilizando-se

$$(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_{i-1}, Y_{i-1}), (\mathbf{X}_{i+1}, Y_{i+1}), \dots, (\mathbf{X}_n, Y_n).$$



# Validação Cruzada



Alternativamente, pode-se usar o *k-fold cross validation*. Nessa abordagem divide-se os dados aleatoriamente em *k-folds* (lotes) disjuntos com aproximadamente o mesmo tamanho.



# Validação Cruzada

Sejam  $L_1, \dots, L_k \subset \{1, \dots, n\}$  os índices associados a cada um dos lotes. A ideia do *k-fold cross validation* é criar  $k$  estimadores da função de regressão,  $\hat{g}_{-1}, \dots, \hat{g}_{-k}$ , em que  $\hat{g}_{-j}$  é criado usando todas as observações do banco menos aquelas do lote  $L_j$ .



# Validação Cruzada



O estimador do risco é dado por

$$\hat{R}(g) = \frac{1}{n} \sum_{j=1}^k \sum_{i \in L_j} [Y_i - g_{-j}(\mathbf{x}_i)]^2.$$

Note que quando  $k = n$ , nós recuperamos o LOOCV.



# Validação Cruzada



O estimador do risco é dado por

$$\hat{R}(g) = \frac{1}{n} \sum_{j=1}^k \sum_{i \in L_j} [Y_i - g_{-j}(\mathbf{x}_i)]^2.$$

Note que quando  $k = n$ , nós recuperamos o LOOCV.



# Validação Cruzada

## Observação

Segue, da Lei dos Grandes Números, que a estimação do risco baseada na divisão treinamento versus validação fornece um estimador consistente para o erro condicional.



# Treino versus Validação versus Teste

Do mesmo modo que o EQM no conjunto de treinamento é muito otimista pois cada  $g \in \mathbb{G}$  é escolhida de modo a (aproximadamente) minimizá-lo, o EQM no conjunto de validação avaliado na  $g$  com menor EQM no conjunto de validação também é otimista, principalmente se muitos métodos de predição forem avaliados no conjunto de validação.



# Treino versus Validação versus Teste

Do mesmo modo que o EQM no conjunto de treinamento é muito otimista pois cada  $g \in \mathbb{G}$  é escolhida de modo a (aproximadamente) minimizá-lo, o EQM no conjunto de validação avaliado na  $g$  com menor EQM no conjunto de validação também é otimista, principalmente se muitos métodos de predição forem avaliados no conjunto de validação.



# Treino versus Validação versus Teste

Isso pode levar a um super-ajuste ao conjunto de validação, isto é, pode levar à escolha de uma função  $g$  que tem bom desempenho no conjunto de validação, mas não em novas observações.

Uma forma de contornar esse problema é dividir o conjunto original em três partes: treinamento, validação e teste.



# Treino versus Validação versus Teste

Isso pode levar a um super-ajuste ao conjunto de validação, isto é, pode levar à escolha de uma função  $g$  que tem bom desempenho no conjunto de validação, mas não em novas observações.

Uma forma de contornar esse problema é dividir o conjunto original em três partes: treinamento, validação e teste.



# Treino versus Validação versus Teste

Os conjuntos de treinamento e validação são usados como descrito anteriormente.

Já o conjunto de teste é usado para estimar o erro do melhor estimador da regressão determinado com a utilização do conjunto de validação.



# Treino versus Validação versus Teste

Os conjuntos de treinamento e validação são usados como descrito anteriormente.

Já o conjunto de teste é usado para estimar o erro do melhor estimador da regressão determinado com a utilização do conjunto de validação. Assim, poderemos saber se o risco da melhor  $g$  encontrada de fato é pequeno.



# Treino versus Validação versus Teste

Os conjuntos de treinamento e validação são usados como descrito anteriormente.

Já o conjunto de teste é usado para estimar o erro do melhor estimador da regressão determinado com a utilização do conjunto de validação. Assim, poderemos saber se o risco da melhor  $g$  encontrada de fato é pequeno.



# Roteiro

- 1 Introdução
- 2 Função de risco
- 3 Seleção de Modelos
- 4 Balanço entre viés e variância**
- 5 Agradecimentos
- 6 Referências bibliográficas



# Balço entre viés e variância

Um grande apelo para o uso do risco quadrático é sua grande interpretabilidade: o risco quadrático (condicional no novo  $\mathbf{x}$  observado) pode ser decomposto como

$$\begin{aligned}\mathbb{E} \left\{ [Y - \hat{g}(\mathbf{X})]^2 \mid \mathbf{X} = \mathbf{x} \right\} &= \text{Var}(Y \mid \mathbf{X} = \mathbf{x}) \\ &+ \{r(\mathbf{x}) - \mathbb{E}[\hat{g}(\mathbf{x})]\}^2 \\ &+ \text{Var}[\hat{g}(\mathbf{x})].\end{aligned}$$



# Balço entre viés e variância

Assim, o risco pode ser decomposto em três termos:

- $\text{Var}(Y|X = \mathbf{x})$  é a variância intrínseca da variável resposta, que não depende da função  $\hat{g}$  escolhida e, assim, não pode ser reduzida;



# Balço entre viés e variância

Assim, o risco pode ser decomposto em três termos:

- $\text{Var}(Y|\mathbf{X} = \mathbf{x})$  é a variância intrínseca da variável resposta, que não depende da função  $\hat{g}$  escolhida e, assim, não pode ser reduzida;
- $\{r(\mathbf{x}) - \mathbb{E}[\hat{g}(\mathbf{x})]\}^2$  é o quadrado do viés do estimador  $\hat{g}$  e



# Balço entre viés e variância

Assim, o risco pode ser decomposto em três termos:

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



# Balço entre viés e variância

Assim, o risco pode ser decomposto em três termos:

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

É importante observar que os valores dos dois últimos itens podem ser reduzidos se escolhermos  $\hat{g}$  adequado.



# Balço entre viés e variância

Assim, o risco pode ser decomposto em três termos:

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

É importante observar que os valores dos dois últimos itens podem ser reduzidos se escolhermos  $\hat{g}$  adequado.



# Balço entre viés e variância

Grosso modo, modelos com muitos parâmetros possuem viés relativamente baixo, mas variância alta, já que é necessário estimar todos eles.



# Balço entre viés e variância

Já modelos com poucos parâmetros possuem variância baixa, mas viés muito alto, já que são demasiado simplistas para descrever o modelo gerador dos dados.



# Balço entre viés e variância

Assim, com a finalidade de obter um bom poder preditivo, deve-se escolher um número de parâmetros nem tão alto, nem tão baixo.



# Balço entre viés e variância

Note que, enquanto em inferência paramétrica tradicionalmente buscam-se estimadores não viesados para os parâmetros de interesse, em inferência preditiva é comum abrir mão de se ter um estimador viesado para, em troca, conseguir-se uma variância menor e, assim, um risco menor.



# Roteiro

- 1 Introdução
- 2 Função de risco
- 3 Seleção de Modelos
- 4 Balanço entre viés e variância
- 5 Agradecimentos**
- 6 Referências bibliográficas



# Agradecimentos

Para mais detalhes sobre o assunto, ver Izbicki e dos Santos (2020), de onde o conteúdo desta apresentação foi majoritariamente tirado. Qualquer imprecisão poderá ter sido ocasionada na montagem dos slides.



# Roteiro

- 1 Introdução
- 2 Função de risco
- 3 Seleção de Modelos
- 4 Balanço entre viés e variância
- 5 Agradecimentos
- 6 Referências bibliográficas



# Referências bibliográficas I

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

Izbicki, R. e dos Santos, T. M. (2020), *Aprendizado de máquina: uma abordagem estatística*, UICLAP, São Carlos.

James, G., Witten, D., Hastie, T. e Tibshirani, R. (2023), *An introduction to statistical learning: with applications in R*, 2nd edn, Springer, New York, NY.



## Referências bibliográficas II

Nelder, J. A. e Wedderburn, R. W. M. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society. Series A (General)* **135**(3), 370–384.

Stone, M. (1974), 'Cross-validatory choice and assessment of statistical predictions', *Journal of the Royal Statistical Society. Series B (Methodological)* **36**(2), 111–133.



# Obrigado!

✉ [tiago.magalhaes@ufjf.br](mailto:tiago.magalhaes@ufjf.br)

🌐 [ufjf.br/tiago\\_magalhaes](http://ufjf.br/tiago_magalhaes)

🌐 Departamento de Estatística, Sala 319

↑ ↑ ↓ ↓ ← → ← → B A



# Obrigado!

✉ [tiago.magalhaes@ufjf.br](mailto:tiago.magalhaes@ufjf.br)

🌐 [ufjf.br/tiago\\_magalhaes](http://ufjf.br/tiago_magalhaes)

🌐 Departamento de Estatística, Sala 319

↑ ↑ ↓ ↓ ← → ← → B A

