

# Tópicos Especiais em Computação I

## Aprendizado Estatístico

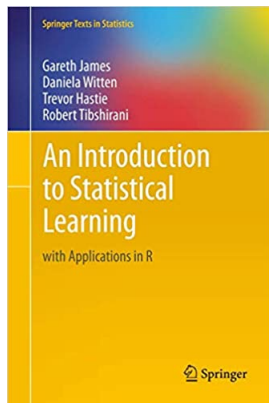
Patrícia Lucas

Bacharelado em Sistemas de Informação  
IFNMG - Campus Salinas

Salinas  
Março 2021

# Referência

## Aprendizado Estatístico



## Capítulo 2: Statistical Learning.

An Introduction to Statistical Learning: with Applications in R. G. James, D. Witten, T. Hastie, and R. Tibshirani. Springer, 2013.

# Visão geral

## Aprendizado Estatístico

A aprendizagem estatística refere-se a um vasto conjunto de ferramentas para a compreensão de dados.

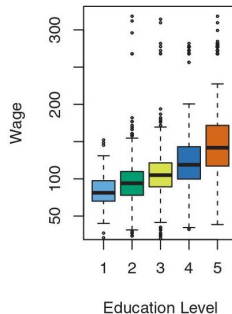
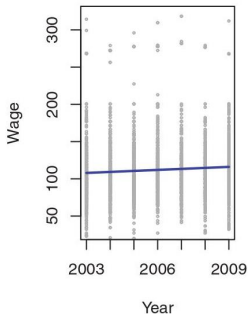
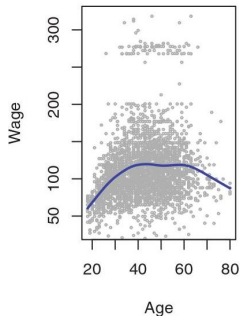
Essas ferramentas podem ser classificadas como supervisionadas ou não supervisionadas.

- Na aprendizagem supervisionada, para cada observação  $x_i$  (atributos preditores), há uma medição de resposta associada  $y_i$ . O objetivo é ajustar um modelo que relacione a resposta aos preditores, com o objetivo de prever com precisão a resposta para observações futuras (previsão) ou melhor compreender a relação entre a resposta e os preditores (inferência).
- Na aprendizagem não supervisionada, apenas dados de entrada são fornecidos nos exemplos. No entanto, mesmo sem rótulos (atributos de saída), podemos aprender relacionamentos e estruturas a partir de tais dados.

# Visão geral

## Aprendizado Estatístico

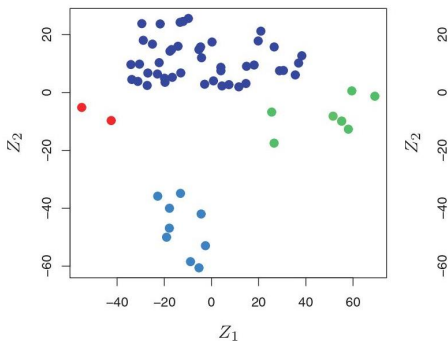
Exemplo de problema onde o **aprendizado supervisionado** é usado:



# Visão geral

## Aprendizado Estatístico

Exemplo de problema onde o **aprendizado não supervisionado** é usado:



Isso é conhecido como um problema de cluster e, ao contrário do exemplo anterior, aqui não estamos tentando prever uma variável de saída.

# Visão geral

## Aprendizado Estatístico

Suponha que sejamos consultores estatísticos contratados por um cliente para aconselhar sobre como melhorar as vendas de um determinado produto.

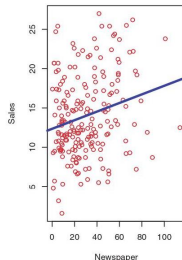
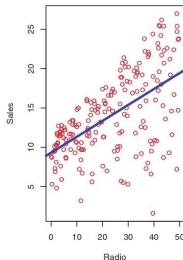
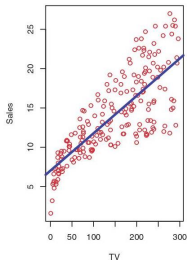
O conjunto de dados de publicidade consiste nas vendas desse produto em 200 mercados diferentes, juntamente com orçamentos de publicidade para o produto em cada um desses mercados para três meios de comunicação diferentes: TV, rádio e jornal.

Se determinarmos que existe uma associação entre publicidade e vendas, podemos instruir nosso cliente a ajustar os orçamentos de publicidade, aumentando indiretamente as vendas.

# Visão geral

## Aprendizado Estatístico

Os orçamentos de publicidade são variáveis de entrada, enquanto a entrada de vendas é uma variável de saída.



Variáveis de entrada ou variáveis independentes: são os orçamentos de publicidade (com TV, Rádio e Jornal). Vamos chamá-las de  $X_1$ ,  $X_2$  e  $X_3$ .

Variável de saída ou variável dependente: são as vendas. Vamos chamá-la de  $Y$ .

# Visão geral

## Aprendizado Estatístico

**Objetivo:** desenvolver um modelo preciso que possa ser usado para prever vendas com base nos três orçamentos de mídia.

Para isso, podemos assumir que existe uma relação entre  $Y$  e  $X$ :

$$Y = f(X) + \epsilon \quad (1)$$

Onde:

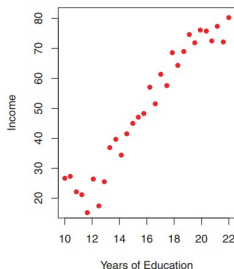
- $f(X)$  é uma função desconhecida de  $X_1$ ,  $X_2$  e  $X_3$ .
- $\epsilon$  é um erro aleatório, independente de  $X$  e com média zero.



# Visão geral

## Aprendizado Estatístico

Exemplo: o gráfico sugere que alguém pode ser capaz de prever a renda usando anos de educação.

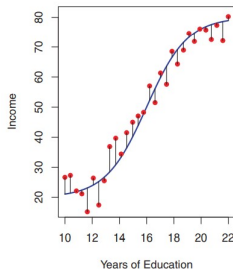


No entanto, a função  $f$  que conecta a variável de entrada à variável de saída é geralmente desconhecida e devemos estimá-la com base nos pontos observados.

# Visão geral

## Aprendizado Estatístico

Aqui, a curva azul mostra a função  $f$  e as linhas verticais representam os termos de erro ( $\epsilon$ ).



Notamos que algumas das 30 observações estão acima da curva azul e algumas abaixo dela, ou seja, o erro médio é aproximadamente 0.

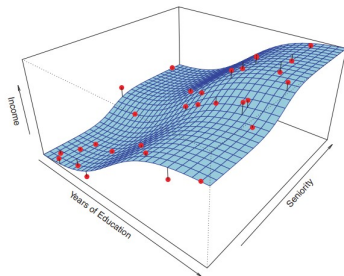
# Visão geral

## Aprendizado Estatístico

Em geral, a função  $f$  pode envolver mais de uma variável de entrada.

Nesse exemplo, representamos a renda em função dos anos de educação e da idade.

Agora,  $f$  é uma superfície bidimensional que deve ser estimada com base nos dados observados.



# Visão geral

## Aprendizado Estatístico

Em essência, a aprendizagem estatística se refere a um conjunto de abordagens para estimar  $f$ .

### **Por que desejamos estimar $f$ ?**

- Para fazer previsões para  $Y$ .
- Para fazer inferências, ou seja, para entender como  $Y$  é afetado quando  $X_1, \dots, X_p$  mudam.

## Como podemos estimar $f$ ?

- Métodos paramétricos: envolvem uma abordagem baseada em modelo.
- Métodos não-paramétricos: não fazem suposições explícitas sobre a forma de  $f$ .

# Métodos paramétricos

## Aprendizado Estatístico

Primeiro, fazemos uma suposição sobre a forma de  $f$ . Por exemplo, uma suposição muito simples é que  $f$  é linear em  $X$ :

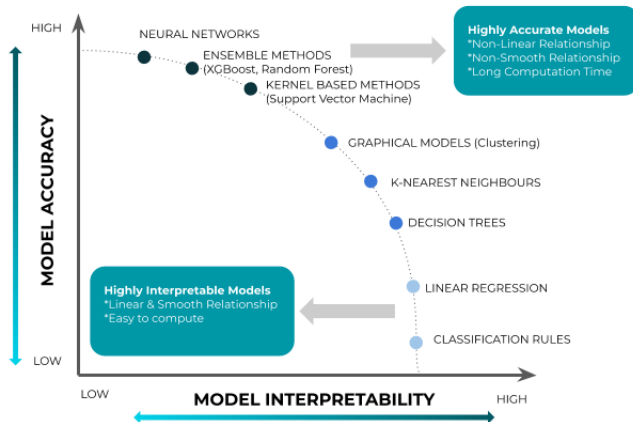
$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad (2)$$

Uma vez que assumimos que  $f$  é linear, o problema de estimar  $f$  é bastante simplificado, pois precisamos estimar apenas os coeficientes  $\beta_0, \beta_1, \dots, \beta_p$ .

Após a seleção de um modelo, precisamos de um procedimento que use os dados de treinamento para ajustar ou treinar o modelo.

# O *trade-off* entre precisão e interpretabilidade

## Aprendizado Estatístico

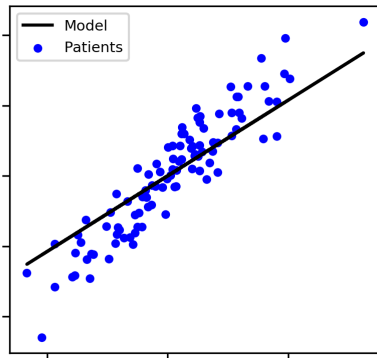
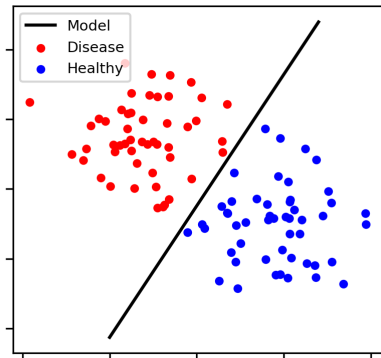


Por que escolheríamos usar um método mais restritivo em vez de uma abordagem mais flexível?

# Aprendizado Supervisionado

## Aprendizado Estatístico

### Regressão X Classificação





# Avaliação da precisão do modelo

## Aprendizado Estatístico

Por que é necessário introduzir tantas abordagens diferentes de aprendizagem estatística, em vez de apenas um único método melhor?

Não há almoço grátis em estatística: nenhum método domina todos os outros sobre todos os conjuntos de dados possíveis.

Portanto, é uma tarefa importante decidir, para qualquer conjunto de dados, qual método produz os melhores resultados.

Selecionar a melhor abordagem pode ser uma das partes mais desafiadoras do desempenho do aprendizado estatístico na prática.

# Medindo a qualidade do ajuste

## Aprendizado Estatístico

A fim de avaliar o desempenho de um método de aprendizado estatístico em um determinado conjunto de dados, precisamos de alguma forma para medir o quão bem suas previsões realmente correspondem aos dados observados.

Ou seja, precisamos quantificar até que ponto o valor de resposta previsto para uma determinada observação está próximo do valor de resposta verdadeiro para essa observação.

Para minimizar o erro de teste esperado, precisamos selecionar um método de aprendizado estatístico que alcance simultaneamente baixa variância e baixo viés.

# O *trade-off* entre viés e variância

## Aprendizado Estatístico

### Variância:

- A variância refere-se à mudança que  $\hat{f}$  sofreria se a estimássemos usando um conjunto de dados de treinamento diferente.
- Como os dados de treinamento são usados para se ajustar ao método estatístico de aprendizado, diferentes conjuntos de dados de treinamento resultam em um  $\hat{f}$  diferente.
- Idealmente, a estimativa para  $f$  não deve variar muito entre os conjuntos de treinamento.
- Quando um método tem alta variância, pequenas alterações nos dados de treinamento podem resultar em grandes alterações em  $\hat{f}$ .

# O *trade-off* entre viés e variância

Aprendizado Estatístico

## Viés:

- O viés refere-se ao erro.
- Por exemplo: a regressão linear assume que há uma relação linear entre  $Y$  e  $X_1, X_2, \dots, X_p$ . É improvável que qualquer problema da vida real realmente tenha uma relação linear tão simples e, portanto, a realização da regressão linear resultará, sem dúvida, em algum viés na estimativa de  $f$ .

# O *trade-off* entre viés e variância

## Aprendizado Estatístico

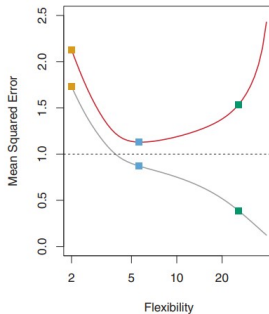
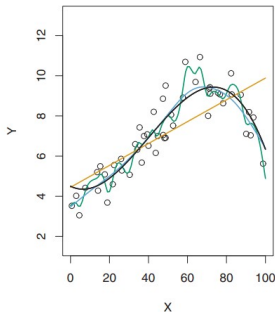


Gráfico à direita:

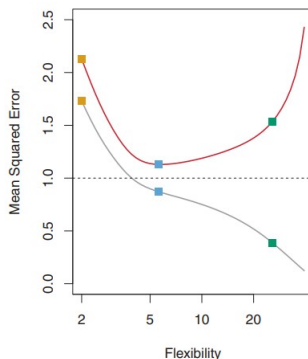
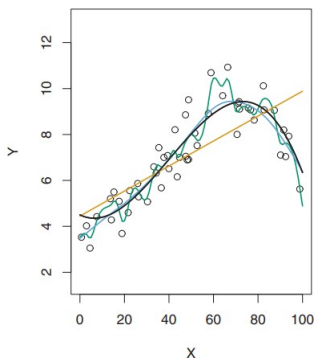
- Curva vermelha: erro de teste.
- Curva cinza: erro de treino.

# O *trade-off* entre viés e variância

## Aprendizado Estatístico

**Underfitting:** ocorre em modelos com alto viés e baixa variância.  
Exemplo: linha amarela.

**Overfitting:** ocorre em modelos com baixo viés e alta variância.  
Exemplo: curva verde.



# O *trade-off* entre viés e variância

Aprendizado Estatístico

