Modelos Avançados de Aprendizagem Estatística

Aprendizagem Estatística Supervisionada

Prof. Washington Santos da Silva

07/06/2021

Mestrado Profissional em Administração

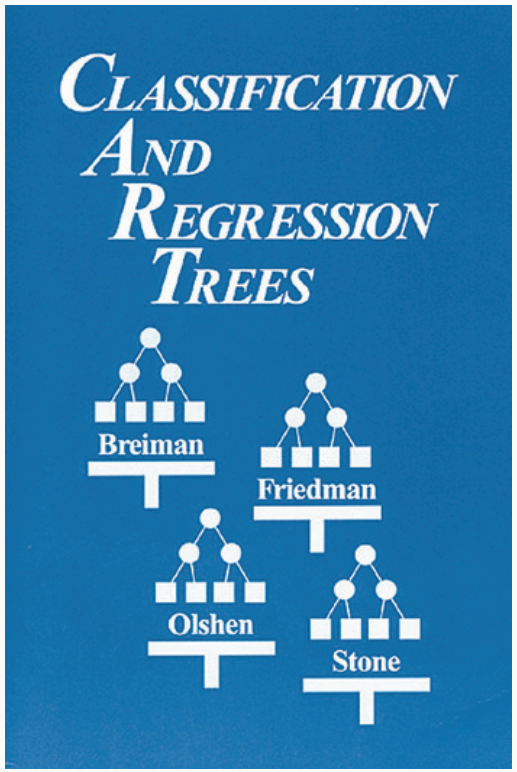
Modelos Avançados de Aprendizagem Estatística

Referências

Modelos Avançados de Aprendizagem Estatística

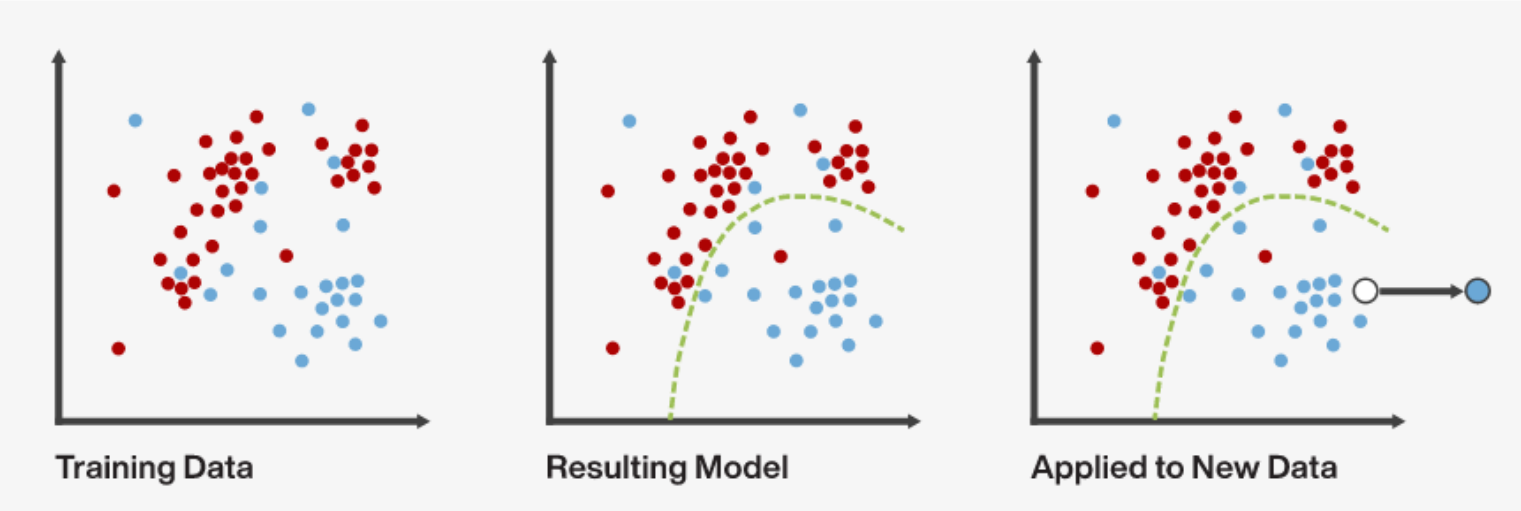
Aprendizagem Estatística

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**.

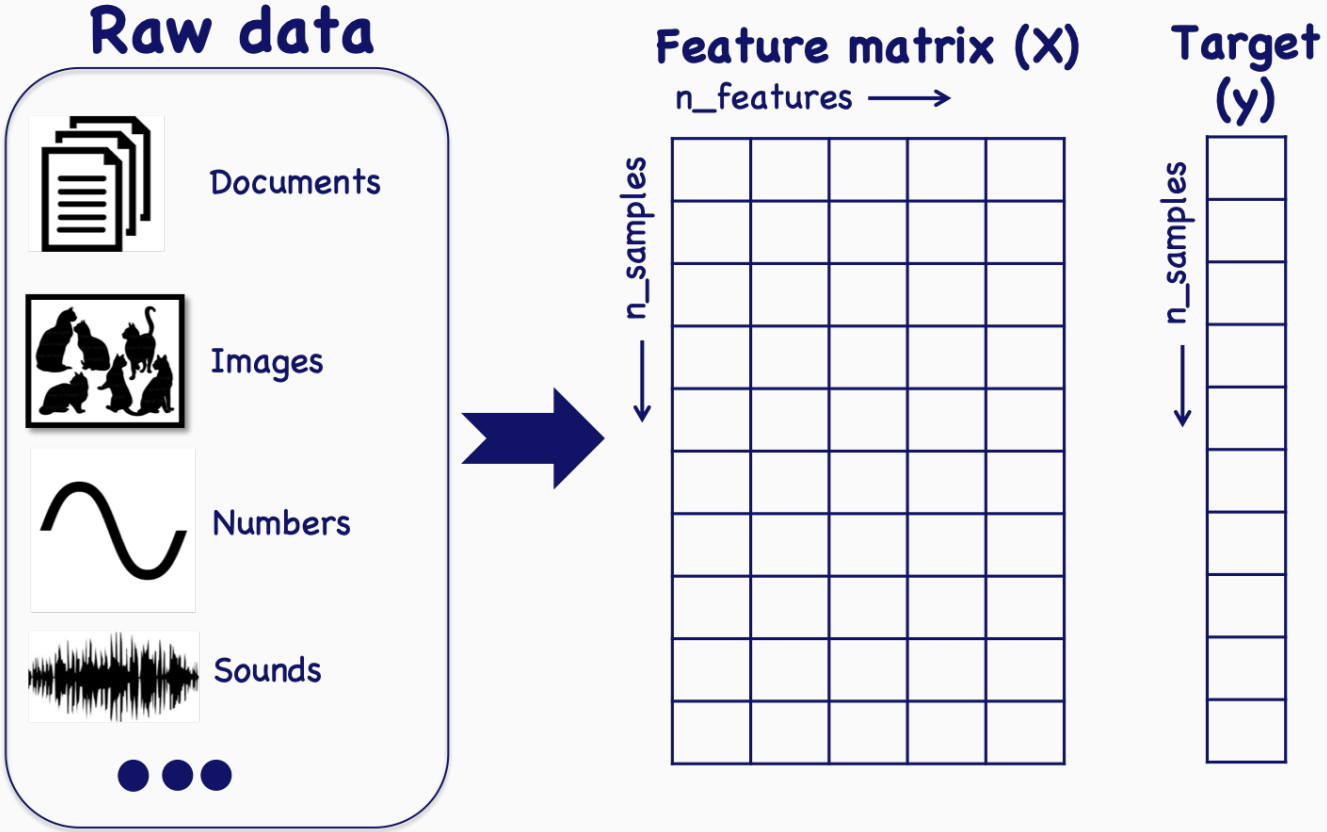


Aprendizagem Supervisionada

Em termos gerais, a aprendizagem supervisionada envolve a construção de um modelo estatístico para prever ou estimar uma **variável resposta** (*output*) com base em uma ou mais variáveis preditoras (*inputs*). O objetivo é treinar um modelo da forma $y = f(x)$, para prever y com base em x.

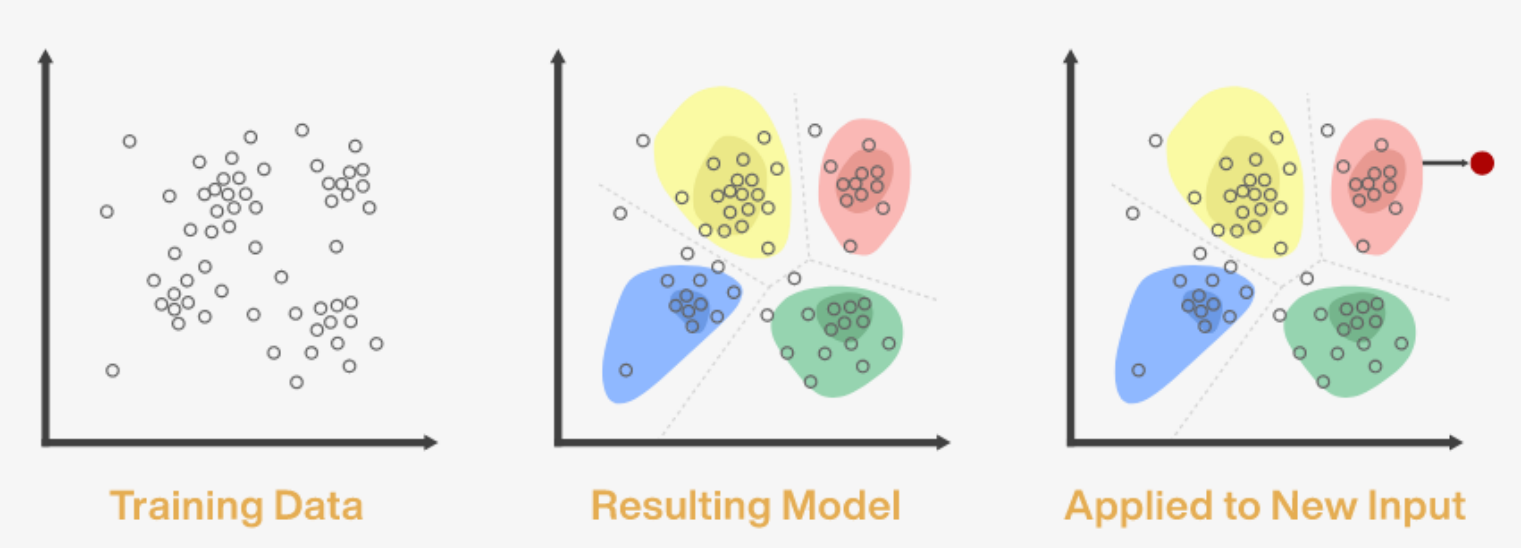


Statistical Learning: Aprendizagem Supervisionada

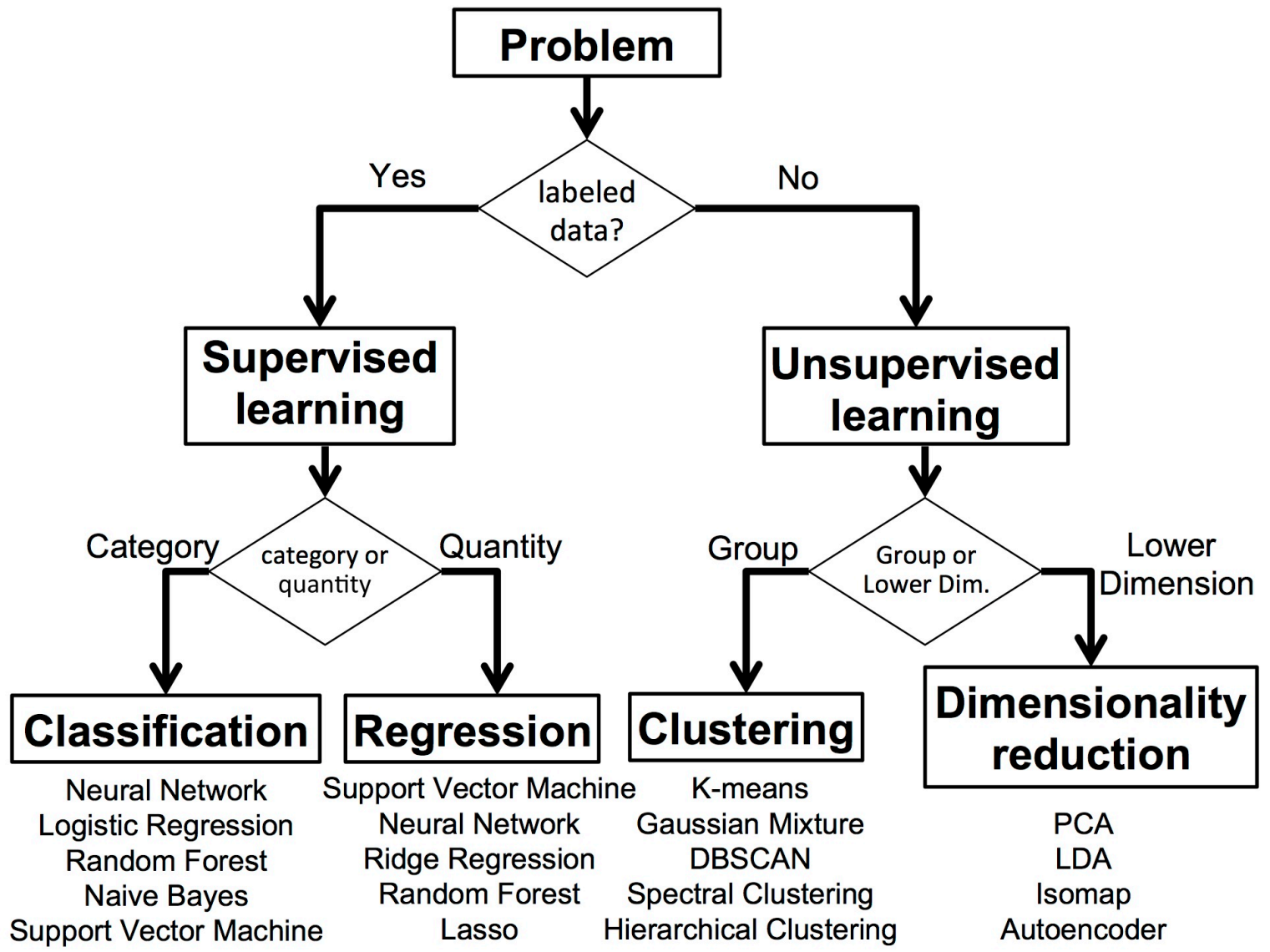


Aprendizagem Não Supervisionada

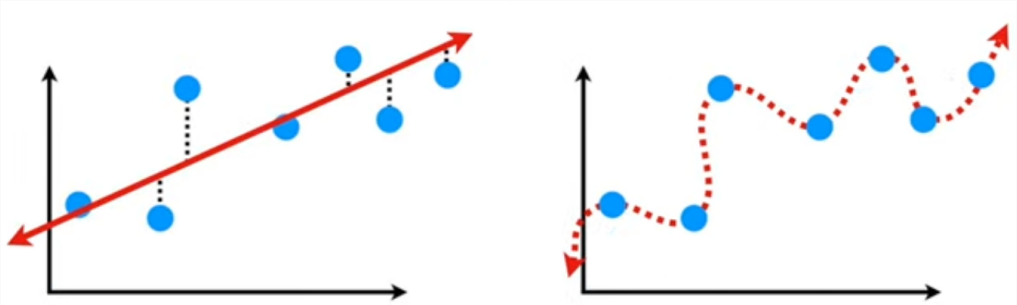
Na aprendizagem não supervisionada, há entradas (variáveis preditoras ou inputs), mas nenhuma saída (variável resposta ou outputs). No entanto, podemos descobrir ou aprender padrões a partir de tais dados.



Statistical Learning: Métodos



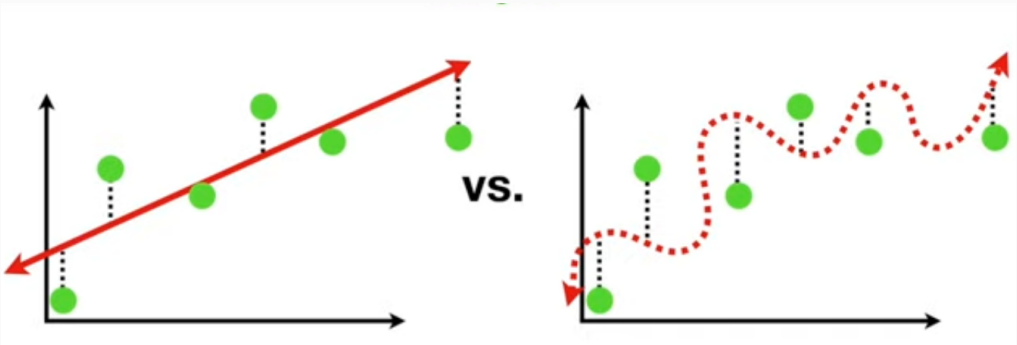
Trade-off entre Viés-Variância (Bias-Variance)



Desempenho nos dados de treino

- Modelo linear: **Viés Alto/*underfit***
- Modelo Não-Linear: **Viés Baixo**

Bias-Variance Trade-off



Desempenho nos dados de teste

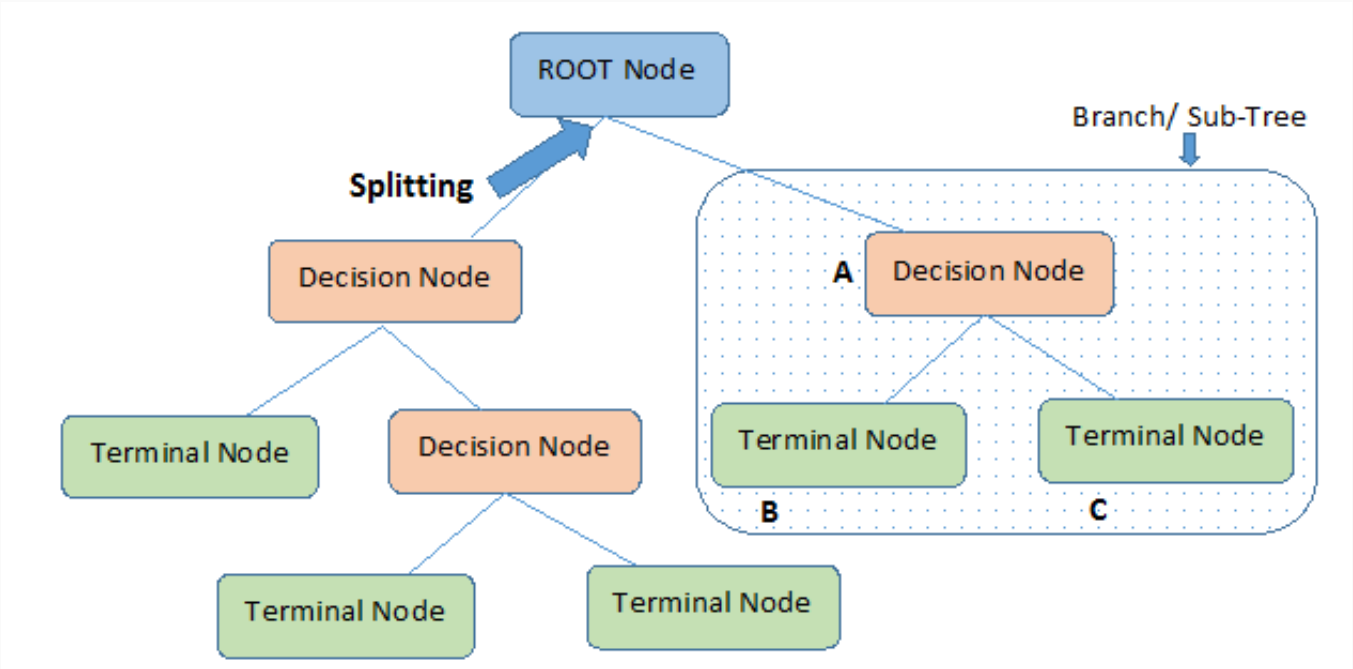
- Modelo linear: **Variância baixa**
- Modelo Não-Linear: **Variância alta**/***overfit***

Árvores de Decisão (Decision Trees): Regressão

Vantagens:

- Capacidade nativa para lidar com variáveis numéricas e categóricas
- Tratam (poucos) dados faltantes adequadamente
- Robustas a valores extremos
- Preparação dos dados (um pouco) mais simples
- Modelam não-linearidades
- Podem ser treinadas rapidamente em grandes bancos de dados

Árvores de Decisão (Decision Trees): Regressão



Árvores de Decisão (Decision Trees): Regressão

Procedimento Geral

1. Dividimos o espaço das preditoras (conjunto de possíveis valores para $X_1, X_2, \ldots, X_p$) em J regiões distintas não sobrepostas $R_1, R_2, \ldots, R_j$.
2. Para cada observação que cai na região R_j, fazemos a mesma previsão, que é simplesmente a média do valores da resposta (y_i) para os dados de treinamento.
3. O objetivo é encontrar "retângulos"que minimizam a Soma dos Quadrados dos Resíduos (*SQR*):

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})$$

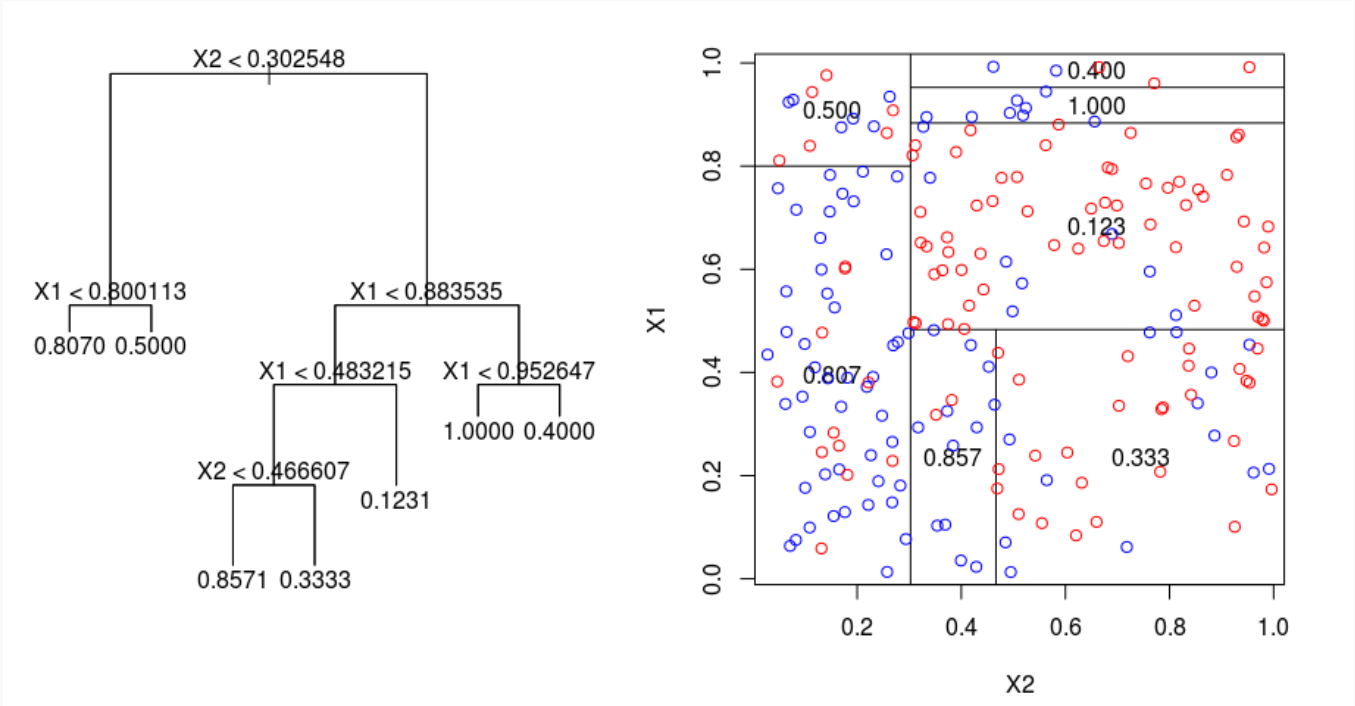
4. Selecionamos o preditor X_j e o ponto de corte s tal que dividir o espaço das preditoras nas regiões $\{X|X_j < s\}$ e $\{X|X_j \geq s\}$ nos leva a maior redução possível na SQR.

Árvores de Decisão: Regressão

Procedimento Geral

4. Em seguida, repetimos o processo, procurando a melhor preditora X e o melhor ponto de corte s para dividir os dados ainda mais, a fim de minimizar a *SQR* em cada uma das regiões resultantes.
5. No entanto, desta vez, em vez de dividir todo o espaço das preditoras, dividimos uma das duas regiões previamente identificadas. Agora temos três regiões.
6. Mais uma vez, procuramos dividir ainda mais uma dessas três regiões, de modo a minimizar a *SQR*. O processo continua até que um o critério de parada é alcançado; por exemplo, podemos continuar até que nenhuma região contenha mais de cinco observações.

Árvores de Decisão (Decision Trees)



Árvores de Decisão: Regressão

Overfit

- O processo descrito pode produzir boas previsões nos dados de treinamento, mas provavelmente apresentará overfit, com desempenho insatisfatório nos dados de teste.
- Uma árvore menor com menos divisões pode levar a uma menor variância e melhor interpretação ao custo de um pequeno viés.
- Uma estratégia melhor é construir uma árvore grande T_0 e, em seguida, podá-la para obter uma subárvore.
- **Cost complexity pruning** é usado para isso.

Árvores de Decisão: Regressão

Cost complexity pruning

- consideramos uma sequência de árvores indexadas por parâmetro de ajuste não negativo α. Para cada valor de α corresponde uma subárvore $T \subset T0$ tal que:

$$\sum_{m=1}^{T} \sum_{x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

  é a menor possível.

- $|T|$ = n. de nós terminais da árvore. R_m é o retângulo correspondente ao m-ésimo nó terminal. $\hat{y}_{R_m}$ é a média dos dados de treinamento em R_m.
- α = controla o trade-off entre a complexidade da subárvore e seu ajuste aos dados de treinamento. Selecionamos o valor ótimo de $\hat{\alpha}$ usando **validação cruzada**.
- Em seguida, retornamos ao conjunto de dados completo e obtemos a subárvore correspondendo a $\hat{\alpha}$.

Validação Cruzada (Cross-Validation)

Validação Cruzada

É uma das várias **técnicas de validação de modelos** semelhantes para avaliar como os resultados de um modelo estatísticp poderão ser generalizados para um conjunto de dados independente. É usada principalmente em situações onde o objetivo da modelagem é a previsão, e desejamos estimar a qualidade de um modelo preditivo **ou para selecionar valores ótimos para parâmetros de controle**.

k-fold Cross-Validation: k = 5

Def.

O conjunto de dados total é dividido em k conjuntos. Um por um, um conjunto é selecionado como o conjunto de teste e os outros $k - 1$ conjuntos são combinados no conjunto de treinamento. Isso é repetido para cada um dos k conjuntos

Árvores de Regressão: Algoritmo

Algoritmo

1. Use *recursive binary splitting* para crescer uma grande árvore para os dados de treinamento, parando apenas quando cada nó terminal tem menos do que um número mínimo de observações.
2. Aplique *cost complexity pruning* à árvore maior de forma a uma sequência das melhores subárvores como uma função de α.
3. Use a validação cruzada k-fold para escolhaer α. Isto é, divida os dados de treinamento em k folds. Para cada $k = 1. . . . , K$:
    a) Repita os passos 1 e 2 para todo k exceto o k-ésimo fold.
    b) Calcule o RMSE para os dados na k-ésima partição deixada para teste.
    c) Tome a média dos resultados para cada valor de α para minimizar o erro médio
4. O Resultado é a subárvore do passo 2 que corresponde ao valor de α.

Árvores de Regressão

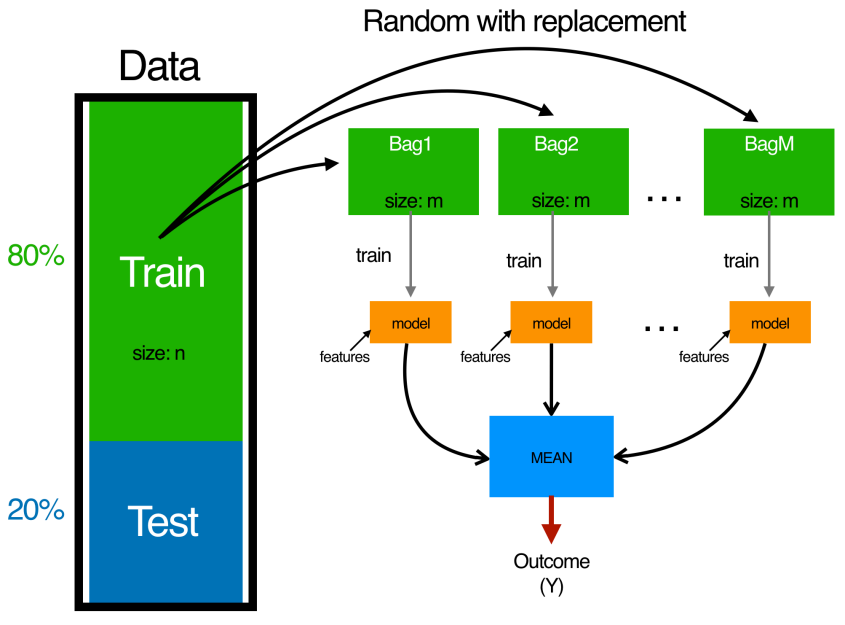
Desvantagens

- Árvores apresentam alta variância -> baixa performance do modelo
- Overfit

Qual a razão da Popularidade

- Bagging: Bootstrap Aggregation
- Boosting

Bagging: bootstrap aggregation



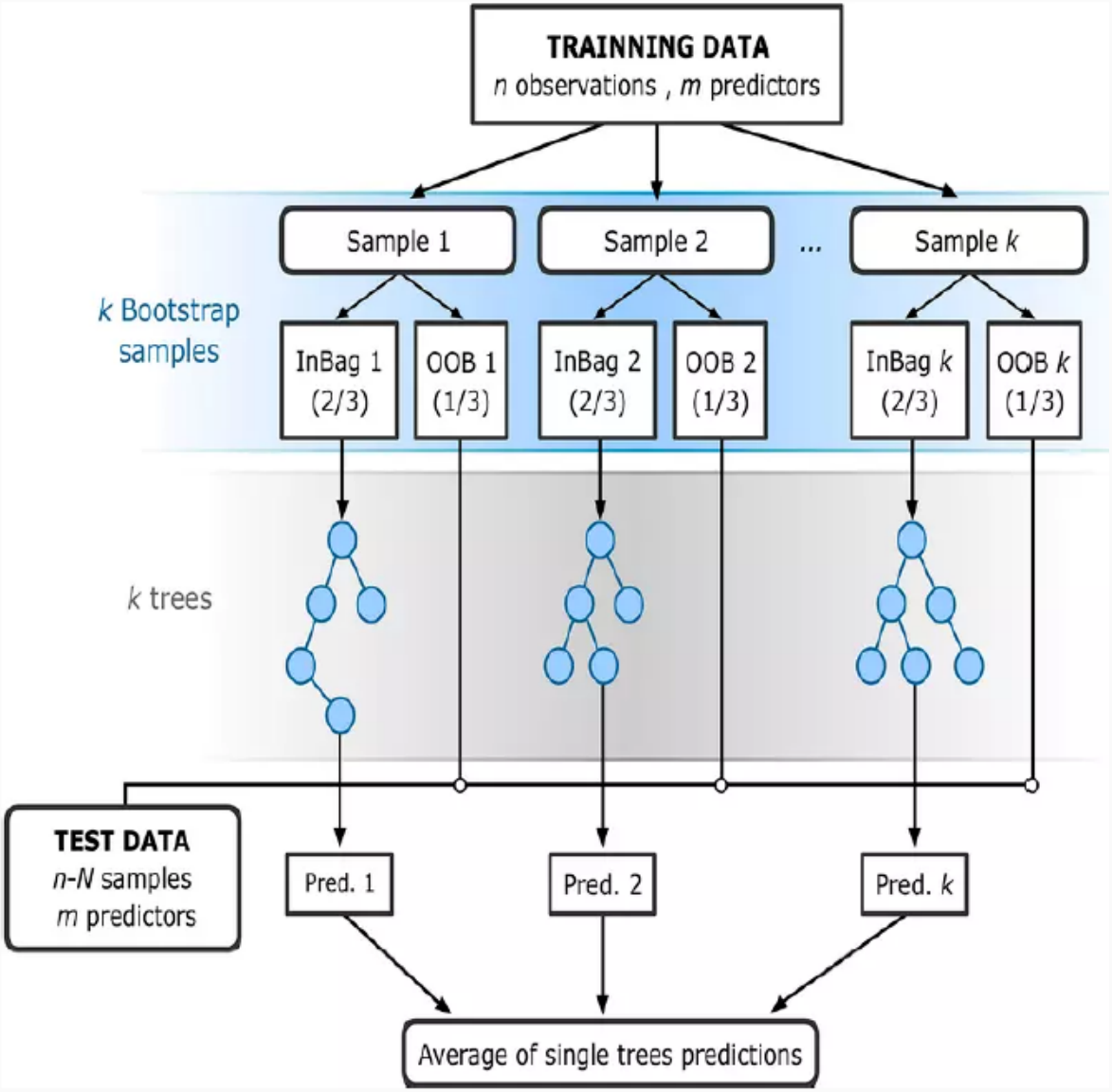
Princípio

Sejam $X_1, X_2, \ldots, X_n$ amostras independentes com variância σ^2:

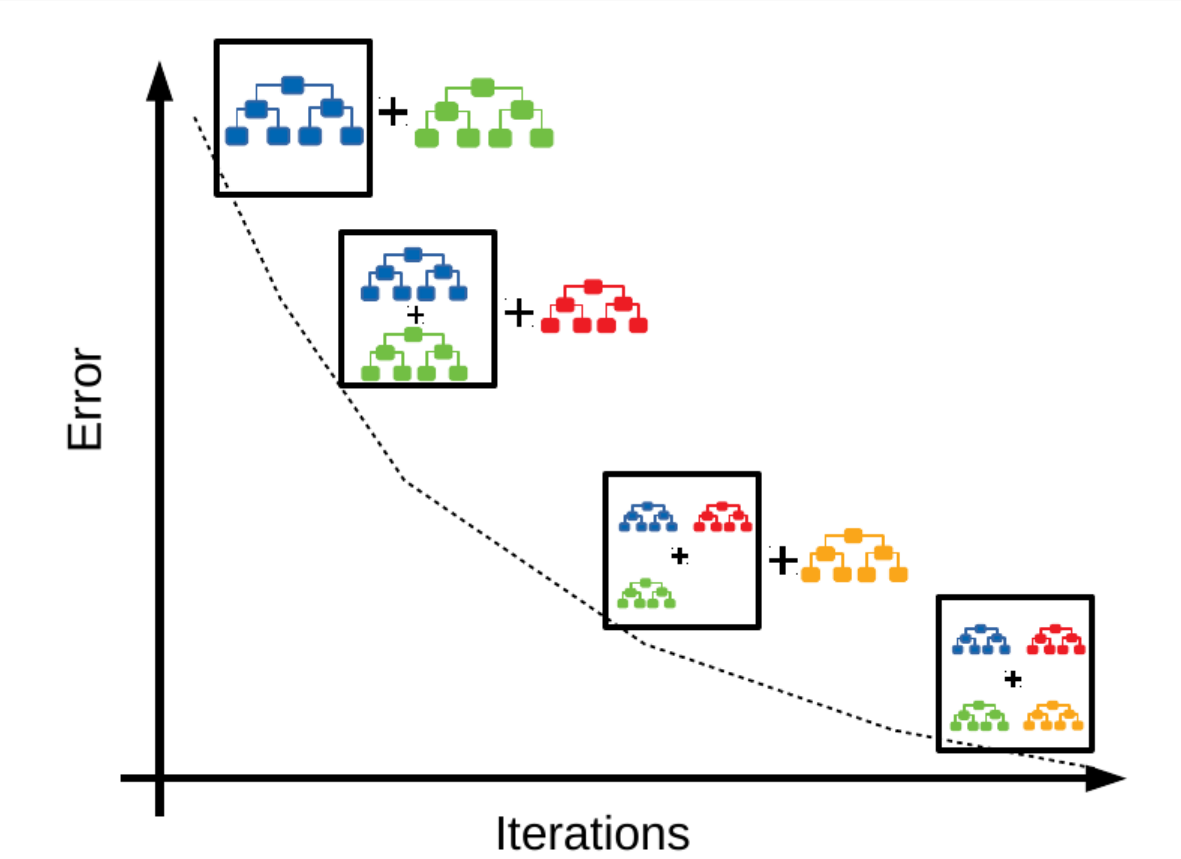
$$V(\bar{X}) = \frac{\sigma^2}{n}$$

Em palavras, **tomar a média de um conjunto de observações reduz a variância**.

Random Forests:



Gradient Boosting



Referências

Referências

JAMES, Gareth et al. An introduction to statistical learning. New York: springer, 2013. Disponível em: https://www.statlearning.com/

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerone. The Elements of Statistical Learning. 2nd. ed., Springer. 2009.