

# Árvores de Regressão: Processo de Treinamento

Introdução ao Aprendizado de Máquina

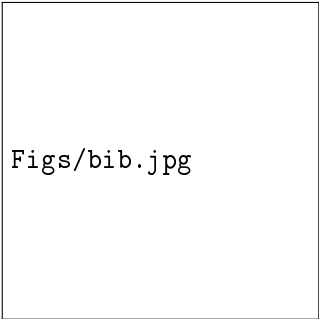
Jean Marcelo Mira Junior

Ramiro Luiz Nunes

14 de maio de 2024

# Sumário

# Base de Científica



Figs/bib.jpg

## Classification and Regression Trees

Leo Breiman, Jerome H. Friedman, Richard A.  
Olshen, Charles J. Stone

# Diferença entre Árvores de Decisão e Regressão

## ■ Árvores de Decisão:

- *Objetivo*: Prever categorias (ex: aprovado/reprovado).
- *Funcionamento*: Modelo em forma de árvore para decisões recursivas.
- *Vantagens*: Fácil interpretação, robustez a outliers.
- *Desvantagens*: Risco de sobreajuste, menos eficaz em grandes dados.

## ■ Árvores de Regressão:

- *Objetivo*: Prever valores contínuos (ex: preço de casa).
- *Funcionamento*: Usa funções matemáticas para ajustar dados.
- *Vantagens*: Alta precisão, útil em relações complexas.
- *Desvantagens*: Modelo complexo, sensível a outliers.

# Diferença entre AID e CART

- **AID (Detecção Automática de Interação):**
  - Um programa precursor na estrutura de árvore para regressão.
  - Foco na construção de uma "árvore honesta" através de um processo específico de poda e estimativa.
- **CART (Árvores de Classificação e Regressão):**
  - Uma abordagem mais avançada que se baseia no AID.
  - Compartilha algumas funcionalidades com o AID, como combinações de variáveis e tratamento de dados ausentes, mas também oferece recursos adicionais:
    - Sem restrições no número de valores de variáveis (diferente do AID).
    - Subamostragem para melhor generalização.
    - Critérios de divisão diferentes dos usados no AID.

# Construção dos Ramos

## ■ Processo de construção envolve:

- *Definição e Coleta de Dados*: Identificar o problema, coletar e preparar dados.
- *Seleção do Algoritmo de Divisão*: Escolher critérios (Gini, entropia) e estratégia de divisão (binária, multivariada).
- *Crescimento da Árvore*: Aplicar critérios de divisão, criar ramos e repetir até atingir critérios de parada.
- *Poda da Árvore*: Podar para evitar sobreajuste, utilizando técnicas pré e pós-pruna.
- *Avaliação e Seleção do Modelo*: Usar validação cruzada para testar desempenho e selecionar o melhor modelo.

# Critério de Divisão

## ■ Critérios e Considerações

### ■ *Critérios:*

- **MSE (Erro Quadrático Médio):** Precisão de previsão em regressão.
- **Ganho de Informação:** Discriminação em classificação baseada em entropia ou Gini.
- **Entropia e Índice de Gini:** Medem a pureza; baixos valores indicam alta pureza.

### ■ *Fatores a Considerar:*

- **Tipo de Problema:** Gini e entropia para classificação; MSE para regressão.
- **Características dos Dados:** Distribuição de classes e presença de outliers.
- **Complexidade do Modelo:** Ganho de informação aumenta complexidade; MSE simplifica.

# Métricas de Avaliação

## ■ MSE (Erro Quadrático Médio)

- **Objetivo:** Avaliar a qualidade de modelos de regressão penalizando grandes erros de previsão.
- **Funcionamento:** Calcula a média dos quadrados das diferenças entre os valores reais e previstos.
- **Vantagens:**
  - Penaliza severamente grandes erros.
  - Utilizável em técnicas de otimização como gradiente descendente.
- **Desvantagens:**
  - Alta sensibilidade a outliers.
  - Difícil interpretação direta.



# Fórmula do MSE

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (1)$$

onde:

- $n$  é o número total de observações.
- $y_i$  é o valor real da  $i$ -ésima observação.
- $\hat{y}_i$  é o valor previsto para a  $i$ -ésima observação.

## Exemplo Prático do MSE

- Valores Reais: [6.575, 6.421, 7.185, 6.998, 7.147]
- Valores Previstos: [6.1, 6.0, 7.2, 6.5, 6.3]
- Diferenças Quadráticas:
  - $(6.575 - 6.1)^2 = 0.225625$
  - $(6.421 - 6.0)^2 = 0.177241$
  - $(7.185 - 7.2)^2 = 0.000225$
  - $(6.998 - 6.5)^2 = 0.248004$
  - $(7.147 - 6.3)^2 = 0.717409$
- Soma das Diferenças Quadráticas: 1.368504
- $MSE = 1.368504 / 5 = 0.273701$

# Métricas de Avaliação

## ■ MAE (Erro Absoluto Médio)

- **Objetivo:** Avaliar a qualidade de modelos de regressão, medindo o erro médio em termos absolutos.
- **Funcionamento:** Calcula a média dos valores absolutos das diferenças entre os valores reais e previstos.
- **Vantagens:**
  - Menos sensível a outliers, oferecendo uma visão mais robusta do desempenho.
  - Fácil interpretação em termos dos valores originais dos dados.
- **Desvantagens:**
  - Não penaliza severamente grandes erros, o que pode levar a ajustes menos precisos em alguns casos.

# Fórmula do MAE

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2)$$

onde:

- $n$  é o número total de observações.
- $y_i$  é o valor real da  $i$ -ésima observação.
- $\hat{y}_i$  é o valor previsto para a  $i$ -ésima observação.

# Exemplo Prático do MAE

- Valores Reais: [6.575, 6.421, 7.185, 6.998, 7.147]
- Valores Previstos: [6.1, 6.0, 7.2, 6.5, 6.3]
- Diferenças Absolutas:
  - $|6.575 - 6.1| = 0.475$
  - $|6.421 - 6.0| = 0.421$
  - $|7.185 - 7.2| = 0.015$
  - $|6.998 - 6.5| = 0.498$
  - $|7.147 - 6.3| = 0.847$
- Soma das Diferenças Absolutas: 2.256
- $MAE = 2.256 / 5 = 0.4512$

# Métodos de Podagem

## ■ O que é Poda?

- Processo de remover ramos desnecessários para simplificar a árvore e melhorar a generalização do modelo.

## ■ Tipos de Poda:

- **Pré-Poda:** Evita ramos excessivos durante a construção, definindo critérios como profundidade máxima, número mínimo de amostras e valor mínimo de impureza.
- **Pós-Poda:** Remove ramos após a construção da árvore, utilizando validação cruzada e critérios de custo-complexidade.

## ■ Métodos Populares de Pós-Poda:

- Redução de Erro e Poda Pessimista: Removem ramos que aumentam o erro de validação.
- Poda Baseada em Custo-Complexidade: Elimina ramos que elevam a complexidade sem aumentar a precisão.

Texto Aqui

Texto Aqui



Texto Aqui

Texto Aqui

Texto Aqui

Texto Aqui

# Conclusões

# Conclusões

# Conclusões

## Referências

- BREIMAN, L.; FRIEDMAN, J.; STONE, C.J.; OLSHEN, R.A. Classification and Regression Trees. Taylor & Francis, 1984. ISBN 9780412048418.



## Árvores de Regressão: Processo de Treinamento

---

Jean Marcelo Mira Junior  
Ramiro Luiz Nunes