Machine Learning
Aula 4

Prof. Thiago A. N. De Andrade

Universidade Federal de Santa Maria
Departamento de Estatística

2025-09-15

Aviso aos estudantes

> - Este é um material novo e atualizado, elaborado especialmente para nosso curso **Machine Learning - UFSM 2025.2**. Entretanto, **não se configura em conteúdo original**. É apenas uma compilação resumida de conteúdos presentes nas referências citadas. Em resumo: é indispensável consultar as referências indicadas.
>
> - As imagens não são autorais e os respectivos créditos são reservados aos autores.
>
> - Este material foi integralmente produzido em `R Markdown`, utilizando o pacote `xaringan`, que possibilita a criação de apresentações **ninja**.

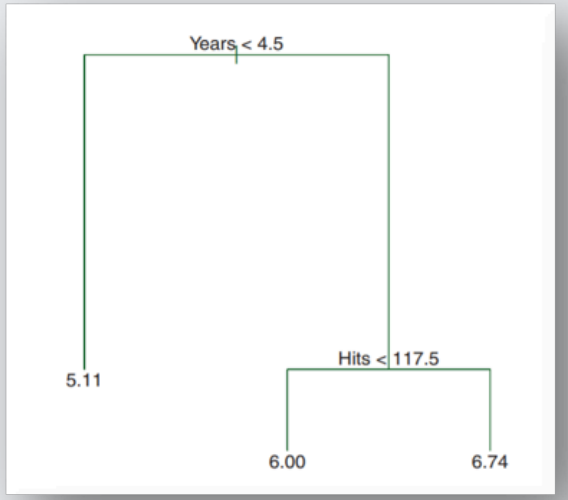
Supervised Learning

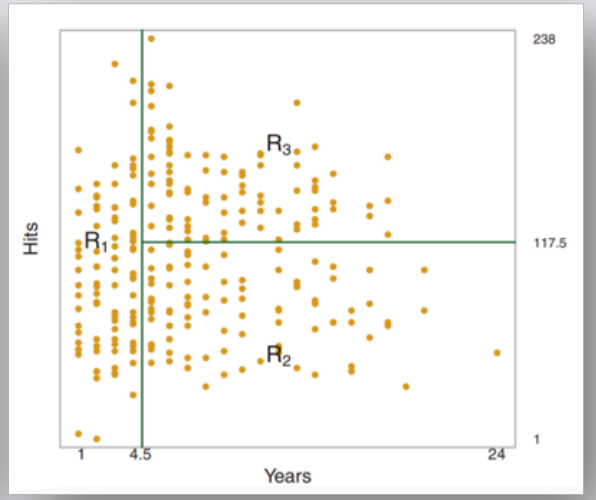
Árvores de Regressão e Classificação

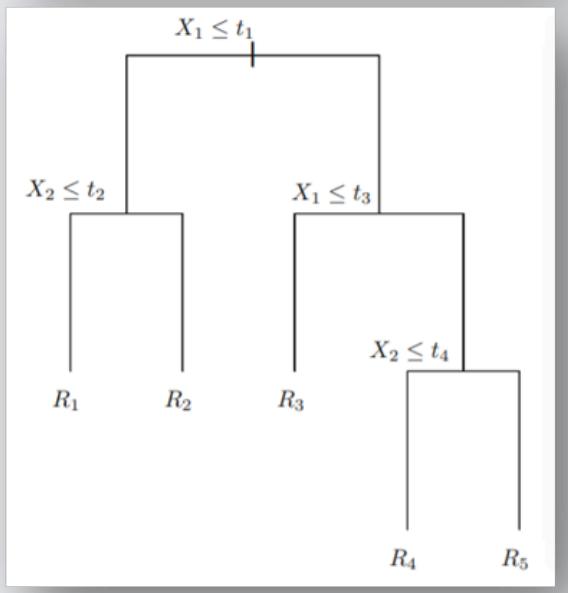
Métodos baseados em árvores

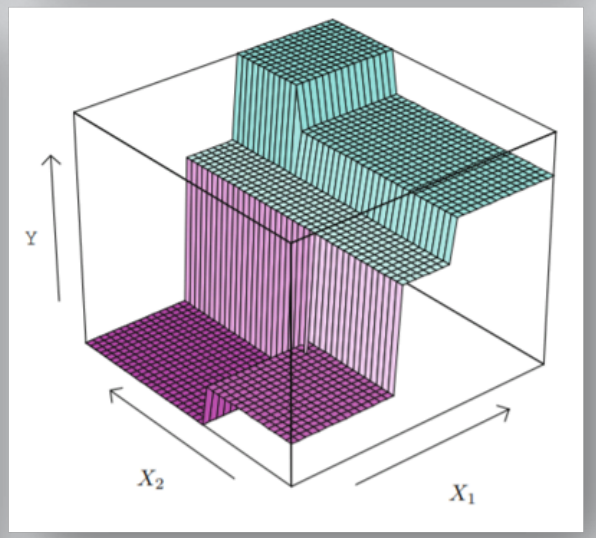
- Uma árvore de decisão é um algoritmo de aprendizado supervisionado **não paramétrico**, utilizado para tarefas de **classificação** e **regressão**.
- Possui uma estrutura hierárquica em forma de árvore, composta por **nó raiz**, **ramificações**, **nós internos** e **folhas**.
- Essa abordagem segmenta o espaço do preditor em regiões cada vez mais homogêneas.
- Como esse processo pode ser resumido em uma árvore, essas abordagens são conhecidas como **métodos baseados em árvores**.


ÁRVORE DE REGRESSÃO
Condição
Sim
Não
Condição
Condição
Sim
Não
Sim
Não
Resultado
Resultado
Resultado
Resultado


Years < 4.5
5.11
Hits < 117.5
6.00
6.74


Hits
238
117.5
1
1
4.5
24
Years
R1
R2
R3


$X_1 \leq t_1$
$X_2 \leq t_2$
$X_1 \leq t_3$
$X_2 \leq t_4$
R_1
R_2
R_3
R_4
R_5


Y
X_2
X_1

Vantagens e Desvantagens

- Métodos baseados em árvores são **simples e interpretáveis**.
- Entretanto, isoladamente, **não costumam competir em precisão** com outros métodos de aprendizado supervisionado.
- Abordagens mais avançadas, como *bagging*, *random forests* e *boosting*, constroem várias árvores e combinam os resultados para obter uma previsão de consenso.
- Essa combinação geralmente proporciona **grande ganho em precisão**, ao custo de **reduzir a interpretabilidade**.

Características das Árvores de Decisão

- Podem considerar várias variáveis explicativas.
- As perguntas em cada nó sempre se baseiam nas **variáveis explicativas**, nunca na resposta.
- O objetivo é dividir os dados em grupos **cada vez mais homogêneos**.
- Decisões importantes: variáveis escolhidas em cada nó, ponto de corte, profundidade da árvore etc.

No R

```
# Regressão linear
modelo_regressao <- linear_reg(
  mode = ,
  engine = ,
  penalty = tune(),
  mixture = )
```

```
# Árvore de decisão
modelo_arvore <- decision_tree(
  mode =, engine =,
  min_n = tune(),
  tree_depth = tune(),
  cost_complexity = tune()
)
```

```
# Regressão logística
modelo_reglog <- logistic_reg(
  mode = ,
  engine = ,
  penalty = tune(),
  mixture =
)
```

```
# Random Forest
modelo_rf <- rand_forest(
  mode =, engine =,
  min_n = tune(),
  mtry = tune(),
  trees = tune()
)
```

Hiperparâmetros

min_n – Número mínimo de observações que um nó deve conter para ser considerado candidato à divisão. Controla se a árvore cria **ramos muito específicos** ou mantém apenas divisões mais gerais.

tree_depth – Profundidade máxima da árvore (também chamada de altura). Define quantas **perguntas sucessivas** podem ser feitas sobre as variáveis explicativas.

cost_complexity – Parâmetro de **poda por complexidade** (penalização). Exige uma **redução mínima do risco/impureza penalizada** para aceitar uma divisão. **Valores maiores geram árvores mais simples**.

Convenções

Dados: $\mathcal{D} = (x_i, y_i)_{i=1}^{n}$, com $x_i = (x_{i1}, \ldots, x_{ip})$ e n observações, em que

- p é o número de preditores (features).
- Classificação: $y_i \in 1, \ldots, C$, em que C é o número de classes.
- Regressão: $y_i \in \mathbb{R}$.

Nó (subconjunto de índices): $\mathcal{N} \subseteq 1, \ldots, n$; tamanho $n_{\mathcal{N}} = |\mathcal{N}|$. Filhos de um corte: $L, R \subset \mathcal{N}$ com $L \cap R = \varnothing$ e $L \cup R = \mathcal{N}$; tamanhos $n_L = |L|$, $n_R = |R|$.

- Proporções de classe no nó $\mathcal{N}$ (classificação):

$$p_{k,\mathcal{N}} = \frac{1}{n_{\mathcal{N}}} \sum_{i \in \mathcal{N}} 1 y_i = k, \quad k = 1, \ldots, C,$$

em que $1_{(\cdot)}$ é a função indicadora.

- Média da resposta no nó $\mathcal{N}$ (regressão):

$$\bar{y}_{\mathcal{N}} = \frac{1}{n_{\mathcal{N}}} \sum_{i \in \mathcal{N}} y_i.$$

- Cortes (splits):
  - Numérico: $s = (j, t)$, com $j \in 1, \ldots, p$ e limiar $t \in \mathbb{R}$, $L = i \in \mathcal{N} : x_{ij} \leq t$ e $R = i \in \mathcal{N} : x_{ij} > t$.
  - Categórico: $s = (j, S)$, com S subconjunto dos níveis de X_j, $L = i \in \mathcal{N} : x_{ij} \in S$ e $R = i \in \mathcal{N} : x_{ij} \notin S$.

Medidas de Impureza

Impureza	Tarefa	Fórmula	Descrição
Gini	Classificação	$\mathcal{I}_{\text{Gini}}(\mathcal{N}) = 1 - \sum k = 1^{C} p_{k,\mathcal{N}}^2$	p_k é a proporção da classe k, $k = 1, \ldots, C$.
Entropia	Classificação	$\mathcal{I}_H(\mathcal{N}) = -\sum_{k=1}^{C} p_{k,\mathcal{N}} \log_2 p_{k,\mathcal{N}}$	Base 2 (bits); outra base muda apenas a **escala**.
Variância	Regressão	$\mathcal{I}_{\text{MSE}}(\mathcal{N}) = \frac{1}{n_{\mathcal{N}}} \sum_{i \in \mathcal{N}} \left(y_i - \bar{y}_{\mathcal{N}}\right)^2$	Impureza de um nó $\mathcal{N}$ (MSE no nó); $\bar{y}_{\mathcal{N}}$ é a média dentro de $\mathcal{N}$.

Entropia

1. O conceito de **entropia da informação** foi introduzido por Claude Shannon (1948).

2. A entropia quantifica o **nível médio de incerteza, surpresa ou informação** associado aos resultados possíveis de uma variável aleatória.

3. Para uma variável discreta X com espaço de estados $\mathcal{X}$, a entropia (em bits) é:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x)\, \log_2 p(x) \;=\; \mathbb{E}\big[-\log_2 p(X)\big].$$

Exemplo aplicado:

Considere $X \sim \text{Bernoulli}(p)$ com $q = 1 - p$.

Balanceado ($p = q = 0.5$):

$$\begin{aligned} H(X) &= -\sum_{i=1}^{2} \tfrac{1}{2} \log_2\left(\tfrac{1}{2}\right) \\ &= -2 \cdot \tfrac{1}{2} \cdot (-1) \\ &= 1 \end{aligned}$$

Desbalanceado ($p = 0.7$, $q = 0.3$):

$$\begin{aligned} H(X) &= -p \log_2 p \; - \; q \log_2 q \\ &= -0.7 \log_2(0.7) \\ &\quad - 0.3 \log_2(0.3) \\ &\approx 0.8813 \; < \; 1 \end{aligned}$$

Ganho de Informação (quando a impureza é entropia)

- Mede a **redução de incerteza** sobre a variável-alvo Y após observar o atributo X_j:

$$IG(Y; X_j) = H(Y) - H(Y \mid X_j) = H(Y) - \sum_v \mathbb{P}(X_j{=}v)\, H\big(Y \mid X_j{=}v\big).$$

- Para um **corte binário** $s = (j, t)$ (ou $s = (j, S)$) aplicado ao nó $\mathcal{N}$ com filhos L e R:

$$IG_H(s; \mathcal{N}) = H(\mathcal{N}) - \frac{n_L}{n_\mathcal{N}} H(L) - \frac{n_R}{n_\mathcal{N}} H(R).$$

> **Generalização:** para qualquer impureza $\mathcal{I}$ (Gini, MSE, ...), usamos o **decréscimo de impureza** $\Delta_{\mathcal{I}}(s; \mathcal{N})$, obtido trocando H por $\mathcal{I}$ na expressão anterior.

Critério Geral de Divisão (objetivo local)

Dado um nó $\mathcal{N}$ e um corte candidato s (numérico $s{=}(j,t)$ ou categórico $s{=}(j,S)$), definimos:

$$\boxed{\Delta_{\mathcal{I}}(s;\mathcal{N}) = \mathcal{I}(\mathcal{N}) - \frac{n_L}{n_{\mathcal{N}}}\,\mathcal{I}(L) - \frac{n_R}{n_{\mathcal{N}}}\,\mathcal{I}(R)}$$

Seleciona-se o corte s que **maximiza** $\Delta_{\mathcal{I}}(s;\mathcal{N})$.

Partição Induzida pela Árvore e Predição nas Folhas

- A árvore define uma **partição** do espaço preditor em regiões disjuntas (folhas) $R_1, \ldots, R_M$.
- **Regressão (estimador em degraus):**

$$\hat{f}(x) = \sum_{m=1}^{M} c_m \, 1\{x \in R_m\}, \qquad c_m = \arg\min_{c} \sum_{i:x_i \in R_m} (y_i - c)^2 = \bar{y}_{R_m},$$

em que $\bar{y}_{R_m} = \dfrac{1}{n_m} \sum_{i:x_i \in R_m} y_i$ e $n_m = |\{i : x_i \in R_m\}|$.

- **Classificação (probabilidades e rótulo):**

$$\hat{p}_k(x) = \frac{1}{n_m} \sum_{i:x_i \in R_m} 1\{y_i = k\}, \qquad \hat{y}(x) = \arg\max_{k \in \{1,\ldots,C\}} \hat{p}_k(x), \;\; x \in R_m.$$

Observações

- A escolha entre **Gini** e **Entropia** raramente altera fortemente o resultado. Ambos favorecem **folhas puras**.
- Em regressão, **minimizar MSE local** equivale a **prever com a média** da folha.
- Para evitar sobreajuste: **profundidade máxima**, **min_n** por nó, **poda** via penalização de complexidade.

Resumo da notação:

- n (obs.), p (preditores), C (classes).
- Nó: $\mathcal{N}$; filhos: L, R; tamanhos: $n_{\mathcal{N}}, n_L, n_R$.
- Proporção de classe no nó: $p_{k,\mathcal{N}}$. Média no nó: $\bar{y}_{\mathcal{N}}$.
- Impureza genérica no nó: $\mathcal{I}(\mathcal{N})$ (Gini/Entropia/MSE).
- Corte: $s = (j, t)$ (numérico) ou $s = (j, S)$ (categórico).
- Partição final: regiões R_m, $m = 1, \ldots, M$.
- Critério de divisão: $\Delta_{\mathcal{I}}(s; \mathcal{N})$ (maximizar).
- Entropia: $H(\cdot)$; Ganho de informação: $IG(\cdot\,;\cdot)$.

Referências

- **A Recursive Partitioning Decision Rule for Nonparametric Classification**
- **Classification and Regression Trees**
- **The strength of weak learnability**

- **Additive Logistic Regression: A Statistical View of Boosting**
- **Random Forests**
- **XGBoost: A Scalable Tree Boosting System**
- **An Introduction to Statistical Learning**
- **Aprendizado de máquina: uma abordagem estatística**
- **Materiais Curso R**

Não deixe de entrar em contato comigo para tirar suas dúvidas: thiagoan.andrade@gmail.com

Estamos no  @thiagoan.andrade para networking e socializações

Obrigado!

Thanks!