Data Science and Big Data

Cesar Augusto Taconeli

13 de novembro, 2018

Árvores de classificação e regressão

Modelos baseados em árvores

- Modelos baseados em árvores (*tree based models*) são alternativas não paramétricas a técnicas de regressão e classificação;
- Baseiam-se em partições dos dados em subgrupos disjuntos, usando os valores das covariáveis;
- As partições configuram um conjunto de regras que são representadas, graficamente, por um diagrama denominado **árvore**;
- Em cada uma das partições produzidas, um modelo simples para a resposta (como uma média, ou uma proporção) deve ser ajustado usando os respectivos dados.

Árvores de classificação e regressão (*Classification And Regression Trees - CART*)

- Proposta por Breiman e colaboradores, em 1984.
- Permitem modelar variáveis respostas numéricas (regressão) ou categóricas (classificação) em função de covariáveis;
- Configuram alternativas a diversos métodos de regressão e classificação, baseando-se em um conjunto mínimo de pressupostos;

Árvores de classificação e regressão (*Classification And Regression Trees - CART*)

- As covariáveis podem ser categóricas, dicotômicas, discretas ou contínuas...

- Árvores de classificação e regressão são usadas em diversos algoritmos de machine learning como *boosting*, *bagging*, *random forests*...

Árvores de classificação e regressão (Iris dataset - Ronald Fisher)

- Como primeira ilustração, vamos considerar os dados sobre dimensões (comprimento e largura) de pétalas e sépalas de 150 flores, 50 de cada uma das seguintes espécies: *setosa*, *versicolor* e *virginica*.

- O objetivo é usar as quatro dimensões das flores para obter uma regra de classificação para as espécies;

- Os slides na sequência ilustram o resultado da aplicação de uma árvore de classificação.

Exemplo - Iris dataset (Ronald Fisher)

- A base de dados iris foi introduzida pelo estatístico e biólogo inglês Ronald Fisher como motivação para a técnica multivariada de análise discriminante linear;
- Os dados consistem de 50 amostras de três espécies de iris (setosa, versicolor e virgínica);
- Quatro medidas foram tomadas em cada flor: comprimento e largura da pétala e comprimento e largura da sépala;
- O objetivo da análise é produzir um modelo que seja capaz de classificar a espécie de iris segundo as dimensões da flor.

Exemplo - Iris dataset (Ronald Fisher)

Figura 1: Espécies de iris.

Exemplo - Iris dataset (Ronald Fisher)

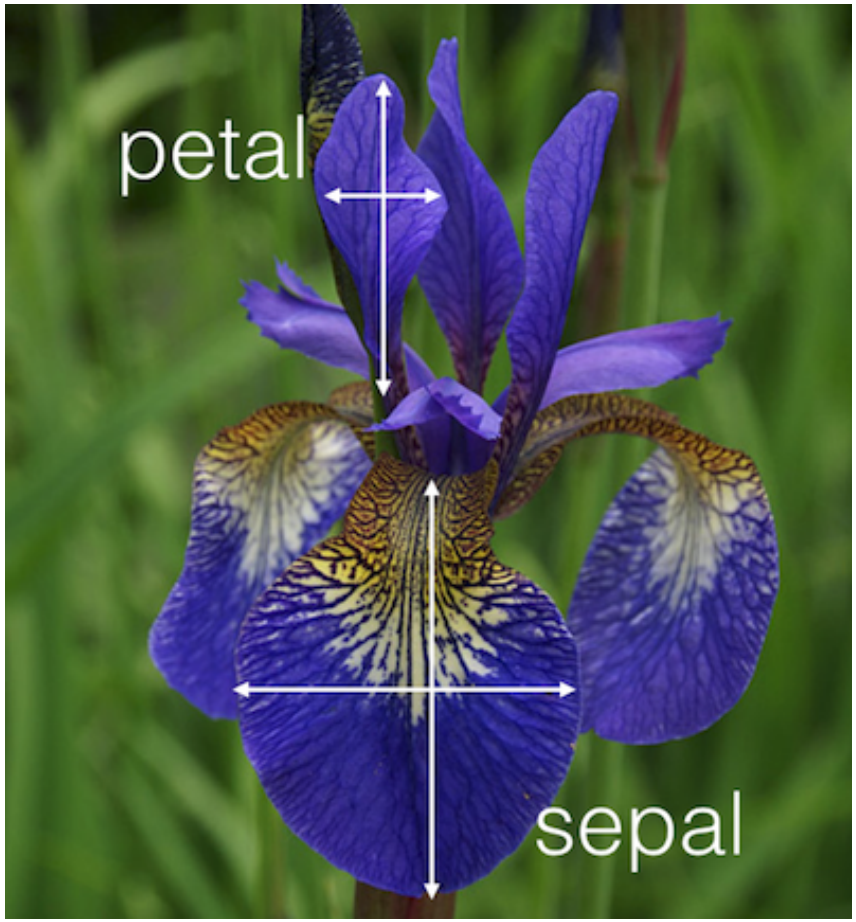

Figura 2: Variáveis usadas para classificação.

Exemplo - Iris dataset (Ronald Fisher)

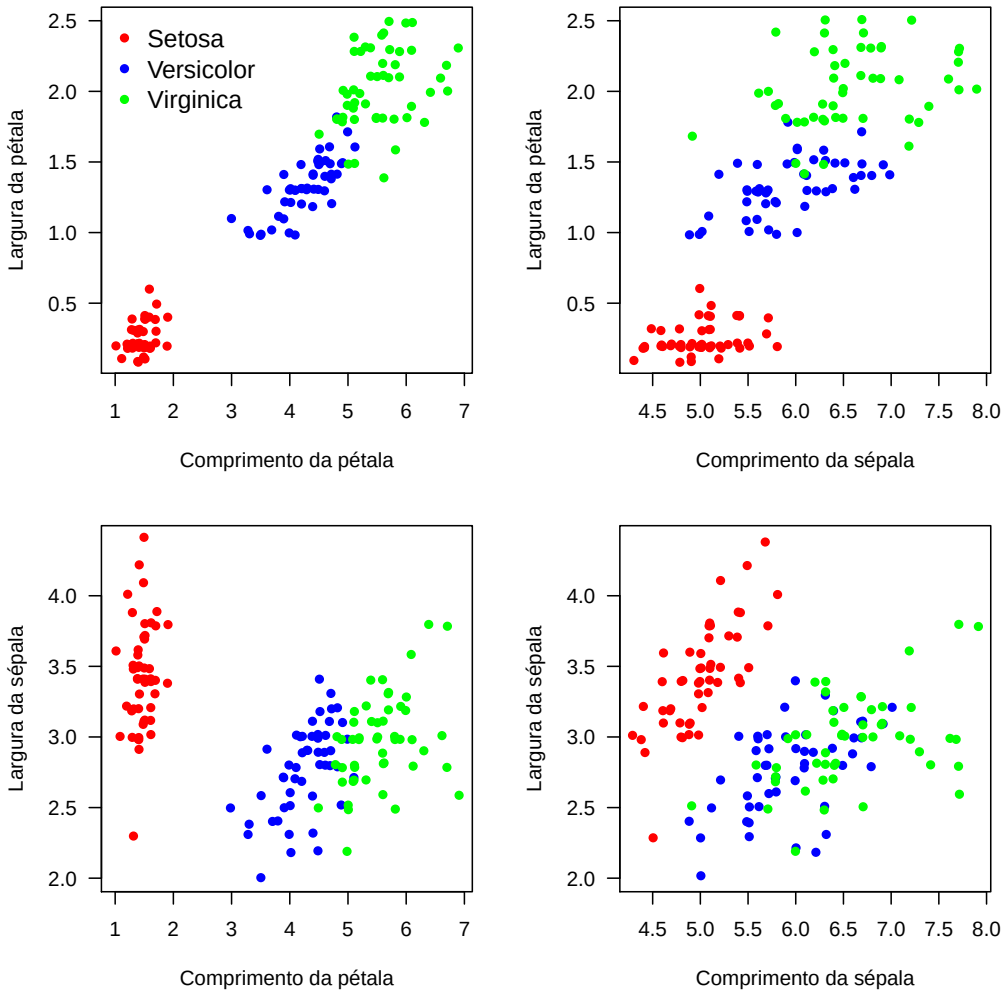

Figura 3: Gráficos para os dados de classificação de espécies de flores

Exemplo - Iris dataset (Ronald Fisher)

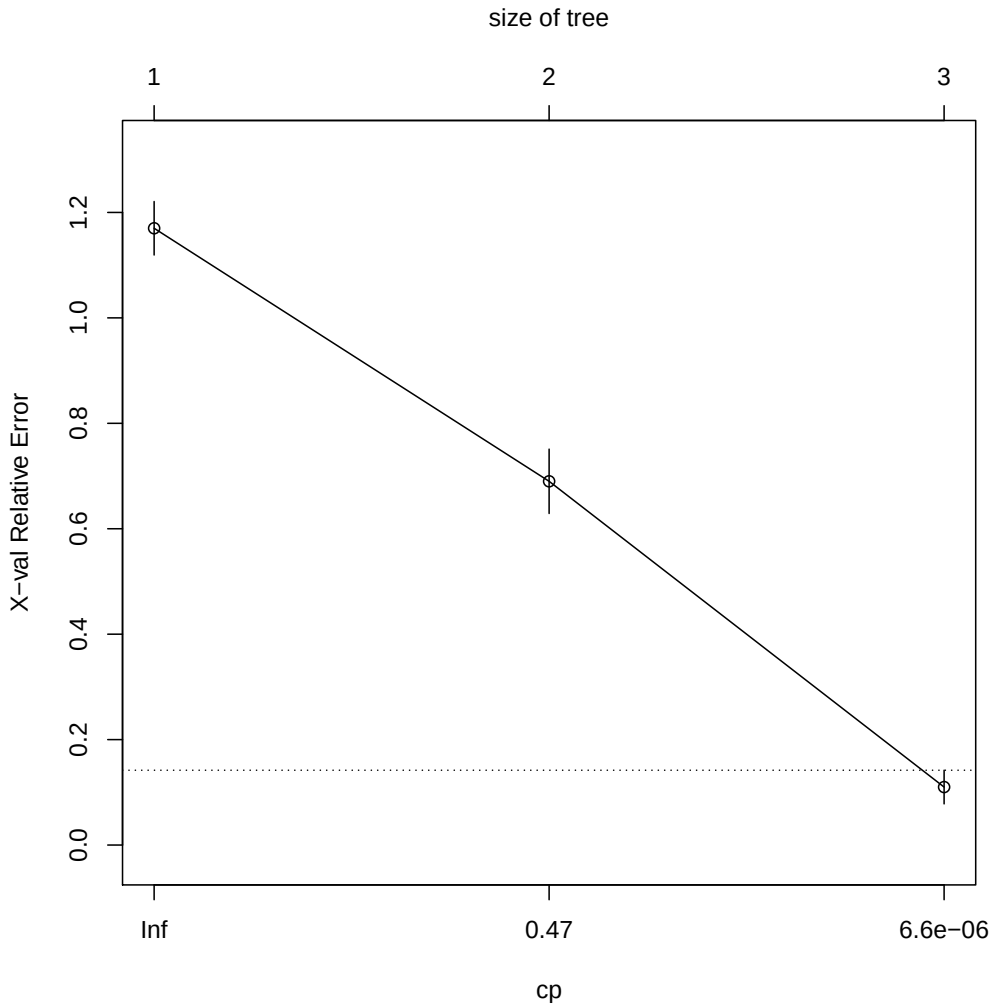

Figura 4: Gráfico de custo complexidade

Exemplo - Iris dataset (Ronald Fisher)

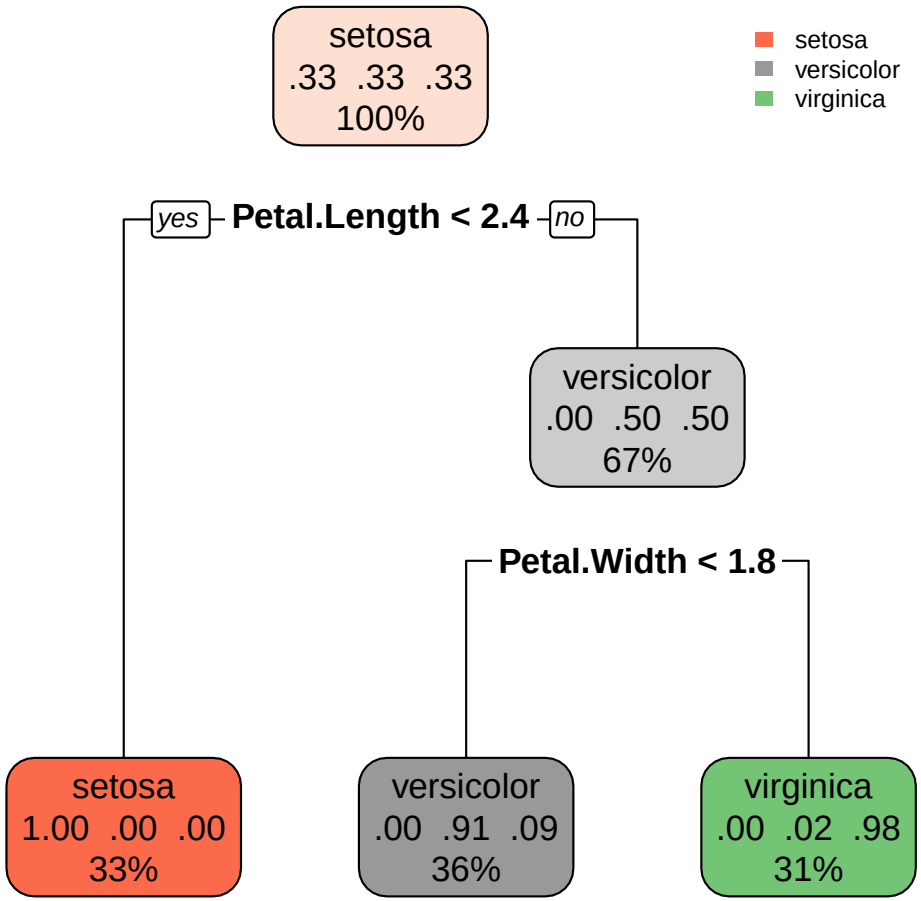

Figura 5: Árvore de classificação

Exemplo - Iris dataset (Ronald Fisher)

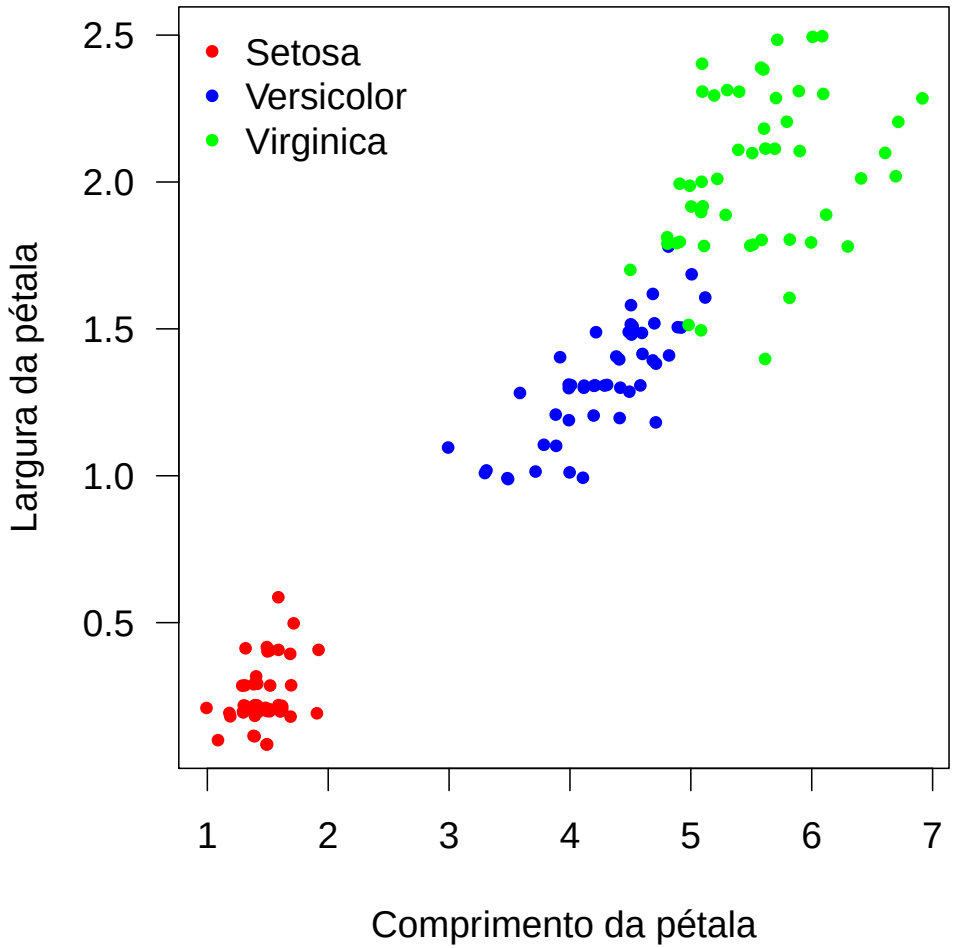

Figura 6: Largura vs comprimento das pétalas

Exemplo - Iris dataset (Ronald Fisher)

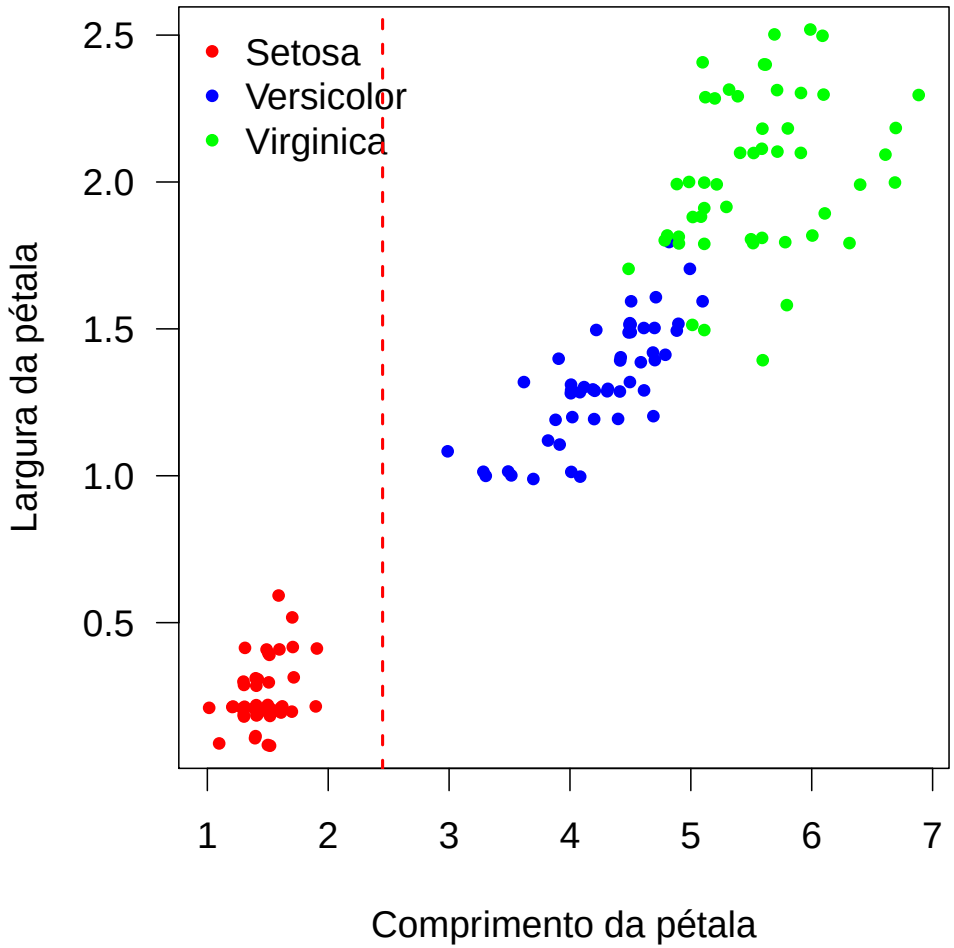

Figura 7: Partição 1

Exemplo - Iris dataset (Ronald Fisher)

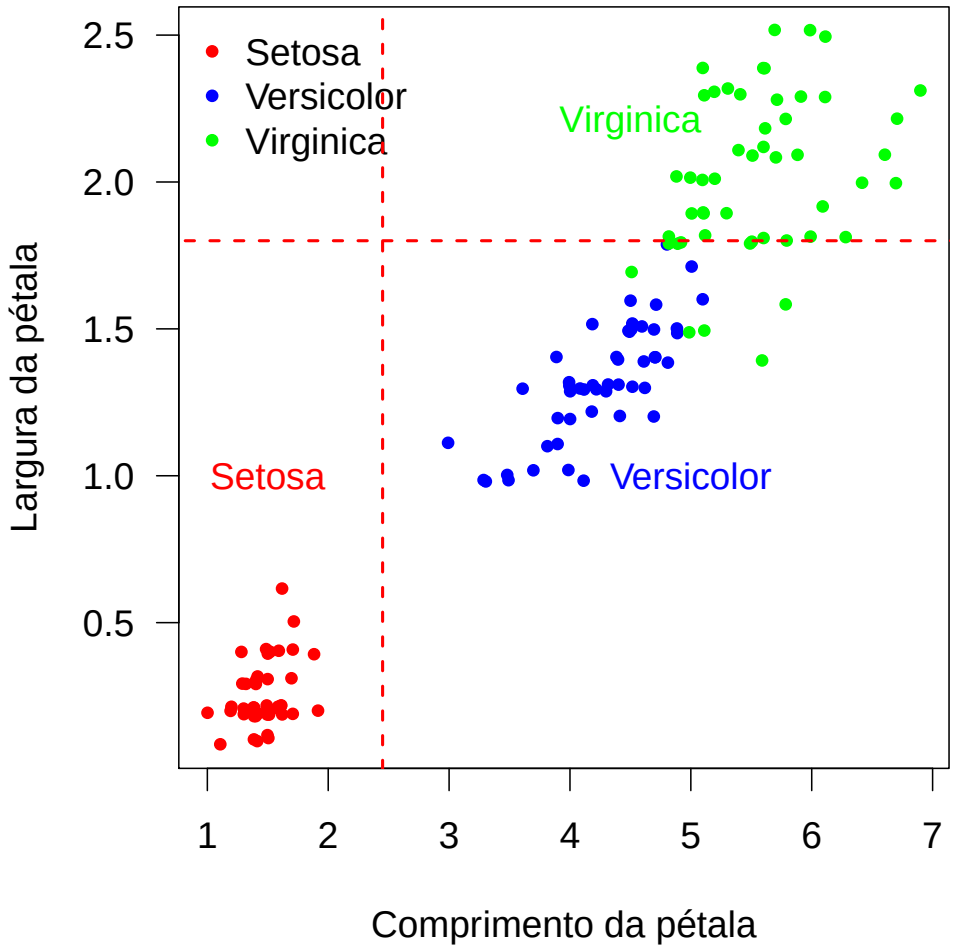

Figura 8: Partição 2 e classificação final

Exemplo - Iris dataset (Ronald Fisher)

Tabela 1: Resultado da classificação

Class/Real	Setosa	Versicolor	Virginica
Setosa	50	0	0
Versicolor	0	49	5
Virginica	0	1	45

Árvores de classificação e regressão

- Seja y a variável resposta e $\boldsymbol{x} = (x_1, x_2, ..., x_p)$ o vetor de covariáveis. Considere n observações de y e $\boldsymbol{x}$.

- O método CART inicia com a partição da amostra original em duas, segundo alguma regra do tipo

$$x_k \leq c \mid x_k > c,$$

para alguma covariável x_k numérica e c algum valor amostrado de x_k, ou

$$x_k \in A \mid x_k \notin A,$$

para uma variável x_k categórica e A uma particular categoria (ou um subconjunto de categorias) de x_k.

Árvores de classificação e regressão

- Exemplos de partição:

  * Idade $\leq$ 30 anos *vs* idade $>$ 30 anos;

  * Renda per capita $\leq$ 2 s.m. *vs* renda per capita $>$ 2 s.m;

  * Sexo **masculino** *vs* sexo **feminino**;

  * Estado civil **solteiro** *vs* estado civil **casado, viúvo ou divorciado**;

  * Estado civil **solteiro ou divorciado** *vs* estado civil **casado ou viúvo**. . .

Árvores de classificação e regressão

- Uma vez efetuada uma partição, temos o espaço das covariáveis (e, consequentemente, os dados) divididos em duas regiões, R_1 e R_2.

- A variável responsável pela partição e o "ponto de corte" são determinados de forma a compor duas regiões R_1 e R_2 tais que:
  - Dentro de cada região os indivíduos sejam homogêneos quanto à resposta;
  - Indivíduos de regiões diferentes sejam heterogêneos;

- Na sequência, o processo de partição é repetido em R_1 e em R_2, novamente buscando, em cada região, a variável e respectivo ponto de corte que proporcionem melhor ajuste.

Árvores de classificação e regressão

- O processo é repetido em R_1 e R_2, e assim sucessivamente nas subamostras formadas. Ao final, teremos M regiões delimitadas no espaço das covariáveis, que denotaremos por $R_1, R_2, ..., R_M$.

- O resultado da aplicação do método CART pode ser representado por um diagrama contendo as partições e os grupos constituídos (**nós**), que denominamos **árvore**.

Árvores de classificação e regressão

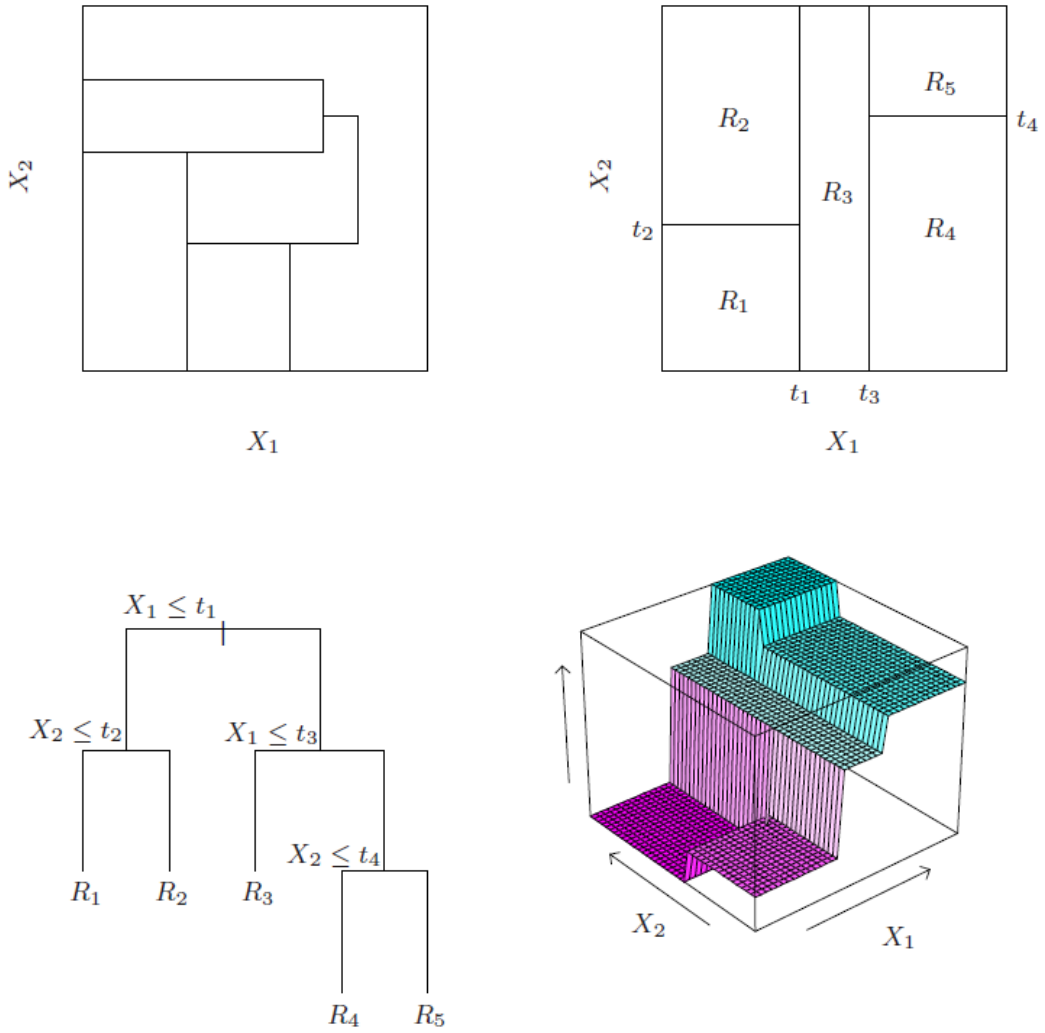

Figura 9: Ilustração - partições e árvores de regressão

Árvores de regressão - Seleção das partições

- Para árvores de regressão, é usual considerar a soma de quadrados de resíduos como critério de minimização para a partição das amostras (nós):

$$SQR = \sum_{i=1}^{n} (y_i - \hat{f}_m(\boldsymbol{x}_i))^2, \tag{1}$$

em que $\hat{f}_m(\boldsymbol{x}_i)$ é a predição para o nó m em que a observação i é alocada ($m \in \{1, 2, ..., M\}$).

Árvores de regressão - Seleção das partições

- No caso de árvores de regressão, consideramos:

$$\hat{f}_m(\boldsymbol{x}_i) = \frac{1}{n_m} \sum_{i \in m} y_i, \tag{2}$$

em que n_m é o número de observações em m.

- Assim, a predição é definida pela média dos valores de y para as observações alocadas ao nó m.

Árvores de regressão - Seleção das partições

- Suponha a partição de um nó (O) em dois novos nós (L e R) segundo uma particular regra (variável e ponto de corte). A avaliação da partição se baseia na redução da soma de quadrados de resíduos:

$$\Delta SQR = SQR_O - \left(\frac{n_L}{n_O} SQR_L + \frac{n_R}{n_O} SQR_R \right), \tag{3}$$

sendo n_O, n_L e n_R os números de observações nos respectivos nós.

- A partição que produzir maior valor para ΔSQR deve ser executada.

- A regra de partição é aplicada sucessivamente aos nós originados até atingir algum critério de parada (número mínimo de observações por nó ou nos nós a serem partidos, número máximo de níveis na árvore...).

Árvores de regressão - O processo de poda

- Após obtida uma grande árvore, inicia-se o processo de poda, em que as partições são sucessivamente desfeitas até voltar à amostra original.

- O processo de poda baseia-se na seguinte função de custo-complexidade:

$$R_\alpha(T) = R(T) + \alpha|T|, \tag{4}$$

em que T representa uma árvore, $|T|$ o número de nós finais (complexidade) e $R(T)$ a soma de quadrados de resíduos da árvore:

$$R(T) = \sum_{m=1}^{M} \frac{n_m}{n} SQR_m. \tag{5}$$

Árvores de regressão - O processo de poda

- O parâmetro α na função de custo-complexidade controla a complexidade do modelo.
- Para diferentes valores de α tem-se árvores de diferentes tamanhos que minimizam $R_\alpha(T)$.
- Tomando $\alpha = 0$ tem-se como solução a maior árvore possível (sem qualquer poda), uma vez que não se penaliza sua complexidade.

Árvores de regressão - O processo de poda

- Para $\alpha \to \infty$ tem-se penalização máxima para a complexidade e a solução é simplesmente a amostra original.
- Variando α a partir de zero tem-se uma sequência de árvores aninhadas, cada uma ótima para seu particular tamanho (número de nós finais).
- É usual representar a função de custo-complexidade por meio de uma curva (versus α e ou $|T|$).

Árvores de regressão - Seleção do modelo

- Uma vez definida a sequência de árvores aninhadas, deve-se identificar, nessa sequência, a árvore ótima (ou seja, a melhor escolha para α).

- Nesta etapa, é comum utilizar validação cruzada. Uma breve descrição da seleção por validação cruzada é descrita na sequência.

Árvores de regressão - Seleção do modelo

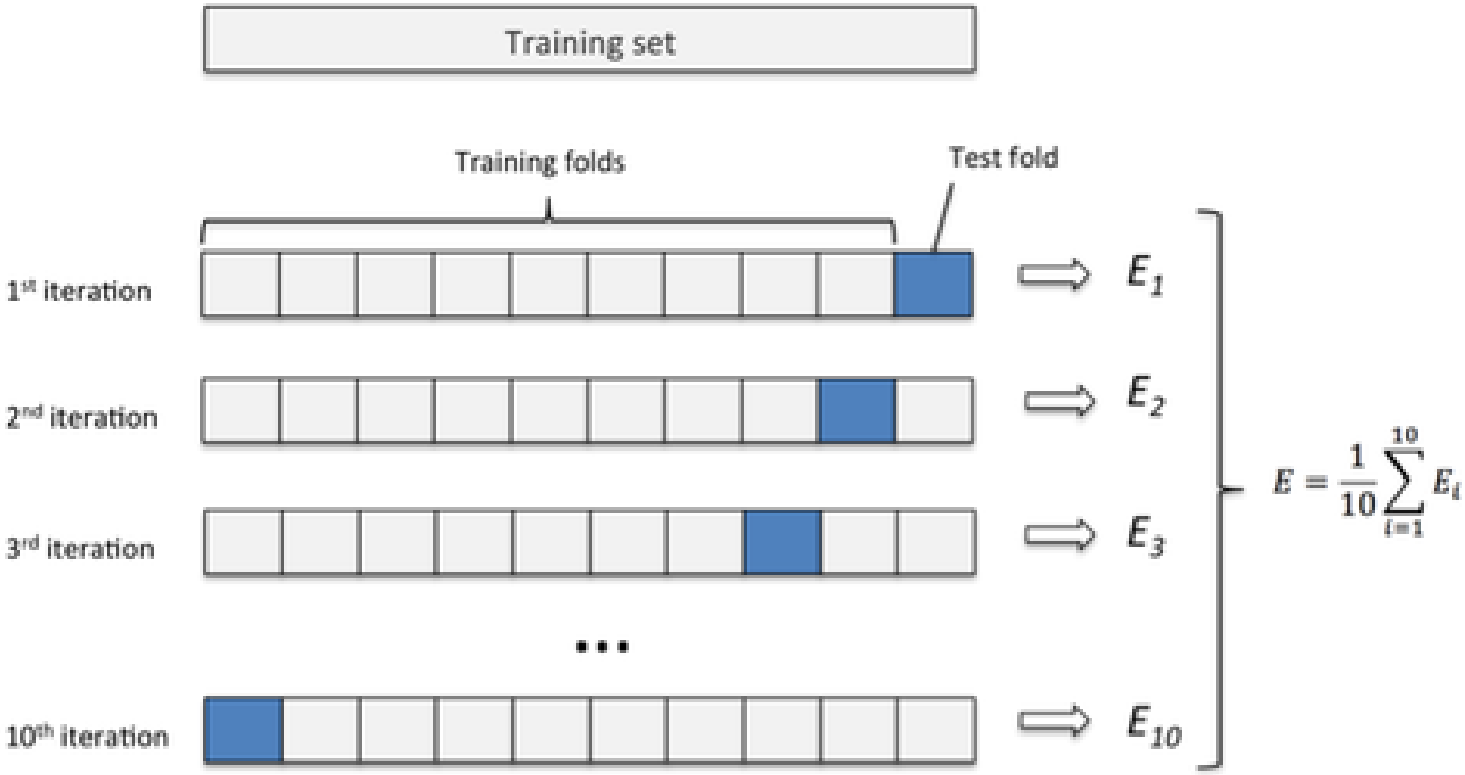

Figura 10: Ilustração - validação cruzada (10 fold)

Árvores de regressão - Seleção do modelo

- **Passo 1:** Identificação de uma sequência de valores α_1, α_2, ..., α_k para α, cada qual indicando uma das árvores na sequência aninhada que minimiza a função de custo-complexidade;

- **Passo 2:** Dividir a base de dados em s grupos de tamanho (aproximado) s/n: $G_1, G_2, ..., G_s$;

- **Passo 3:** Ajustar o modelo à base completa, exceto pelas observações em G_1, o qual denotamos T_1;

Árvores de regressão - Seleção do modelo

- **Passo 4:** Calcular a predição para cada observação i em G_j, sob cada modelo T_j, $j = 1, 2, ..., k$;

- **Passo 5:** Calcular a soma de quadrados dos erros de predição para o conjunto de observações em G_j :

$$\sum_{i \in G_j} (y_i - \hat{f}_{(j)}(\boldsymbol{x}_i))^2, \tag{6}$$

em que $\hat{f}_{(j)}(\cdot)$ denota a predição sob o modelo ajustado sem as observações em G_j.

Árvores de regressão - Seleção do modelo

- **Passo 6:** Os passos 3, 4 e 5 são repetidos para cada um dos demais grupos G_j. Ao término, para cada árvore $T_1, T_2, ..., T_k$ tem-se a respectiva soma de quadrados de predição obtida por validação cruzada:

$$SQVC = \sum_j \sum_{i \in G_j} (y_i - \hat{f}_{(j)}(\boldsymbol{x}_i))^2. \tag{7}$$

- Seleciona-se então a árvore que produz menor valor de SQVC.

- Na prática, usa-se a regra do erro padrão, em que se seleciona a menor árvore tal que seu SQVC não exceda o SQVC mínimo por mais de um erro padrão de SQVC (estimado também na validação cruzada).

Árvores de classificação e regressão - pontos fortes

- Facilmente (e fartamente!) interpretáveis;
- Predições obtidas de maneira bastante simples;
- Fácil identificação das variáveis mais associadas à resposta;

Árvores de classificação e regressão - pontos fortes

- Lida naturalmente com dados missing;
- Permitem acomodar interações de elevada ordem entre as variáveis;
- Implementados em diversos softwares, de maneira bastante eficiente.

Árvores de classificação e regressão - pontos fracos

- Árvores de classificação e regressão têm elevada variância, sendo instáveis frente a perturbações moderadas nos dados;
- Devido à instabilidade, usar uma árvore como modelo preditivo apresenta, em geral, baixa acurácia.
- A capacidade preditiva é aumentada de maneira significativa usando métodos baseados em múltiplas árvores.

Árvores de classificação

- Árvores de classificação se aplicam quando a variável resposta é categórica (binária ou politômica);
- O algoritmo de árvores de classificação é semelhante ao de árvores de regressão, com algumas modificações
- A diferença mais importante é a troca da soma de quadrados dos resíduos por alguma medida de heterogeneidade mais apropriada para dados categóricos.
- Dentre as alternativas, temos os critérios de Gini e da informação, conforme apresentados na sequência.

Árvores de classificação

- Vamos considerar um problema de classificação em que a resposta tenha r categorias, denotadas por $1, 2, ..., r$.
- Considere uma amostra (ou um nó) e $p_1, p_2, ..., p_r$ as proporções com que cada categoria é observada.
- A medida de informação (ou entropia) é definida por:

$$Inf = -2 \times \sum_{l=1}^{r} p_l \ln(p_l) \tag{8}$$

- A medida de Gini é definida por:

$$Gini = 1 - \sum_{l=1}^{r} p_l^2. \tag{9}$$

Árvores de classificação

- Para o caso de duas categorias, em que as proporções de casos em cada uma delas são p e $1-p$, as medidas de Informação e de Gini ficam dadas por:

$$Inf = -p\ln(p) - (1-p)\ln(1-p) \tag{10}$$

$$Gini = 1 - p^2 - (1-p)^2 = 2p(1-p). \tag{11}$$

- A Figura 1 apresenta o comportamento das medidas de informação e Gini para o caso de duas categorias.

Árvores de classificação

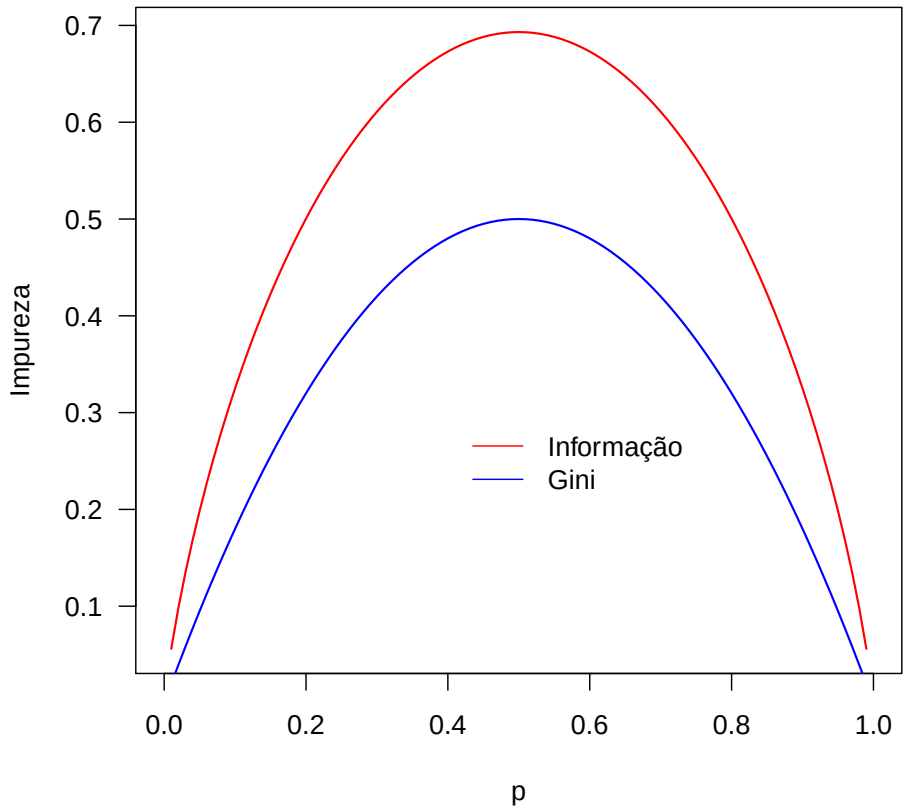

Figura 11: Comparação dos critérios de informação e de Gini para r=2 grupos

Árvores de classificação

- Como pode ser observado na Figura 11, as duas medidas são minimizadas quando os indivíduos da amostra pertencem a um mesmo grupo ($p \to 1$ ou $p \to 0$);
- Adicionalmente, as medidas de Gini e de informação são maximizadas quando as proporções são iguais nas r categorias da resposta ($p_1 = p_2 = ... = p_r$).
- Suponha a partição de um nó (O) em dois novos nós (L e R) segundo uma particular regra (variável e ponto de corte). A avaliação da partição se baseia na redução da medida de impureza:

$$\Delta H = H_O - \left(\frac{n_L}{n_O} H_L + \frac{n_R}{n_O} H_R \right), \tag{12}$$

em que H denota, genericamente, a medida de informação, de Gini ou qualquer outra medida de impureza.

Árvores de classificação

- O ajuste da árvore de classificação segue os mesmos passos de uma árvore de regressão, com o ajuste de uma grande árvore, poda e seleção da árvore por validação cruzada.

- Em árvores de classificação é comum classificar as observações pela categoria mais frequente no nó ao qual ela é alocada;

- Assim, para um problema de classificação binária, a predição é dada pela categoria k tal que $p_k > 0.5$, $k \in \{1, 2\}$.

Árvores de classificação

- As predições podem ser obtidas usando critérios alternativos, baseados nas proporções de indivíduos de cada categoria compondo o nó;
- Diferentes custos de má classificação podem ser especificados, de maneira a identificar uma regra de classificação que minimize os custos de má-classificação.
- A título de ilustração, podemos considerar, para um problema de classificação binária, em que:

$$c(1|2) = 5 \times c(2|1),$$

sendo $c(1|2)$ o custo de classificar como categoria 1 um indivíduo da categoria 2, que é cinco vezes o custo de classificar como categoria 2 um indivíduo da categoria 1.

Árvores de classificação

- A performance preditiva de árvores de classificação pode ser avaliada através de indicadores como acurácia, sensibilidade e especificidade, conforme estudamos antes;

- A separação da base original em duas (base de ajuste e de validação) é bastante recomendável para avaliar a performance do modelos em dados externos.

Conditional inference trees

Conditional inference trees (ctrees)

- HOTHORN, Torsten; HORNIK, Kurt; ZEILEIS, Achim. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics, v. 15, n. 3, p. 651-674, 2006.

- Implementação nos pacotes `party` e `partykit`.

- Metodologia flexível, aplicável na análise de respostas contínuas, discretas, ordinais, censuradas, multivariadas...

Conditional inference trees (ctrees)

- Duas limitações recorrentes para métodos de partição recursiva (particularmente CART):
  - Os modelos (árvores) não se baseiam na significância estatística das covariáveis. Não há distrinção entre partições significativas e não significativas;
  - **Viés de partição**: variáveis que proporcionam maior número de partições (como variáveis contínuas) tendem a ser "favorecidas", na execução de partição, em relação àquelas que proporcionam menor número de partições (por exemplo, covariáveis dicotômicas).

- *Conditional inference trees* baseiam-se na significância estatística das covariáveis e não apresentam viés de partição.

Conditional inference trees (ctrees)

- Assim como o algoritmo CART, o algoritmo ctree consiste na execução de partições binárias baseadas nos valores das covariáveis e partição do espaço das covariáveis em regiões (hiper) retangulares;

- O algoritmo *ctree* pode ser descrito, em linhas gerais, da seguinte forma:

Conditional inference trees (ctrees)

1. Considerando todos os elementos da amostra, Teste a hipótese nula global de independência entre a variável resposta e as variáveis explicativas.

- Se a hipótese nula não é rejeitada (para um nível de significância estabelecido), o processo é encerrado. Caso contrário, a variável com associação mais forte (menor p-valor) é selecionada para partição:

- Nesta etapa, algum procedimento pode ser aplicado para controlar o nível de significância global, devido aos múltiplos testes (como o fator de correção de Bonferroni);

Conditional inference trees (ctrees)

2. Para a variável selecionada no passo 2, verifique qual a partição ótima (buscando um ponto de corte, para covariáveis numéricas, ou algum arranjo de categorias, para covariáveis categoricas) e execute-a;

3. Repita os passos 1-2 recursivamente, até que a hipótese nula global de independência não seja rejeitada.

Conditional inference trees (ctrees)

- A metodologia cobre diversos testes largamente aplicados para avaliar associação entre as variáveis, dentre os quais os testes de Spearman, Wilcoxon, Kruskal-Wallis, Logrank, teste de associação linear para variáveis ordinais...

- A significância da associação entre as variáveis é avaliada usando aleatorização (permutação).

- Diferentemente do CART, conditional inference trees não utilizam poda, mas apenas um critério de parada estabelecido pelo nível de significância global considerado.