Decision Tree e Random Forest

Prof. Gustavo Willam Pereira

INSTITUTO FEDERAL
Sudeste de Minas Gerais

Introdução

- O algoritmo *Decision Tree* constrói modelos de regressão ou classificação na forma de uma estrutura em árvore. Veja a figura 1.
- Ele divide um conjunto de dados em subconjuntos menores.
- Espera-se que o novo subconjunto de dados tenha resultados mais homogêneos em relação ao conjunto original, ou seja, tenha menor impureza.

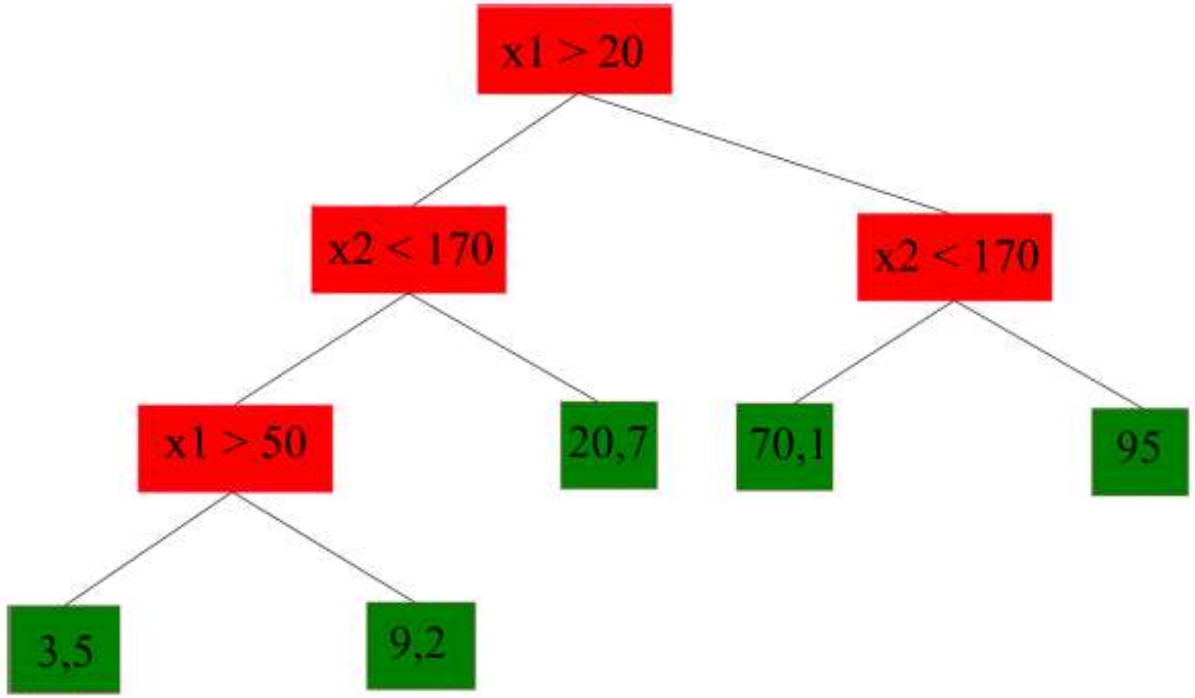


Figura 1. Arvore de decisão gerada pelo algoritmo *Decision Tree*

INSTITUTO FEDERAL
Sudeste de Minas Gerais

Decision Tree

- Na Figura 2 é apresentado a divisões realizadas nas variáveis X1 e X2 que deram origem à arvores da Figura 1.

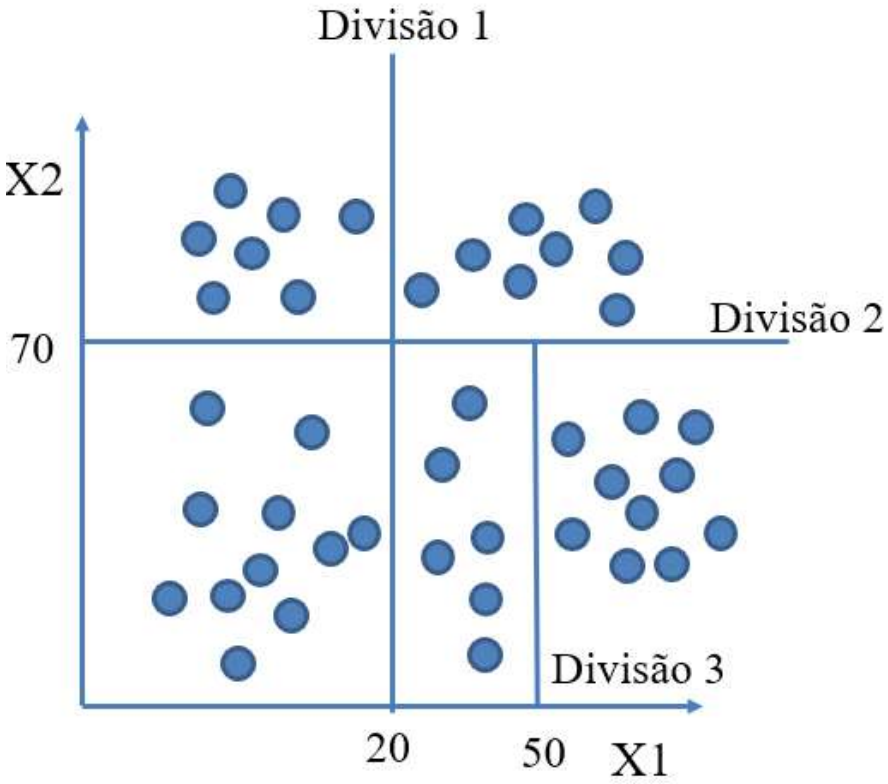


Figura 2: Subdivisões realizadas nas variáveis X1 e X2

INSTITUTO FEDERAL
Sudeste de Minas Gerais

Decision Tree

- Um nó de decisão X1 > 20 possui duas ramificações (ramos), sim e não.
- Cada nó poderá ter dois ou mais ramos.
- A partir do nó sim, por exemplo, outra decisão poderá ser tomada, X2 < 170.
- O nó folha representa uma decisão sobre o destino numérico final da variável y (retângulo em verde).
- O nó de decisão mais alto em uma árvore corresponde ao melhor preditor chamado nó raiz.
- As árvores de decisão podem manipular dados categóricos e numéricos.

Decision Tree

- Um nó de decisão X1 > 20 possui duas ramificações (ramos), sim e não.
- Cada nó poderá ter dois ou mais ramos.
- A partir do nó sim, por exemplo, outra decisão poderá ser tomada, X2 < 170.
- O nó folha representa uma decisão sobre o destino numérico final da variável y (retângulo em verde).
- O nó de decisão mais alto em uma árvore corresponde ao melhor preditor chamado nó raiz.
- As árvores de decisão podem manipular dados categóricos e numéricos.

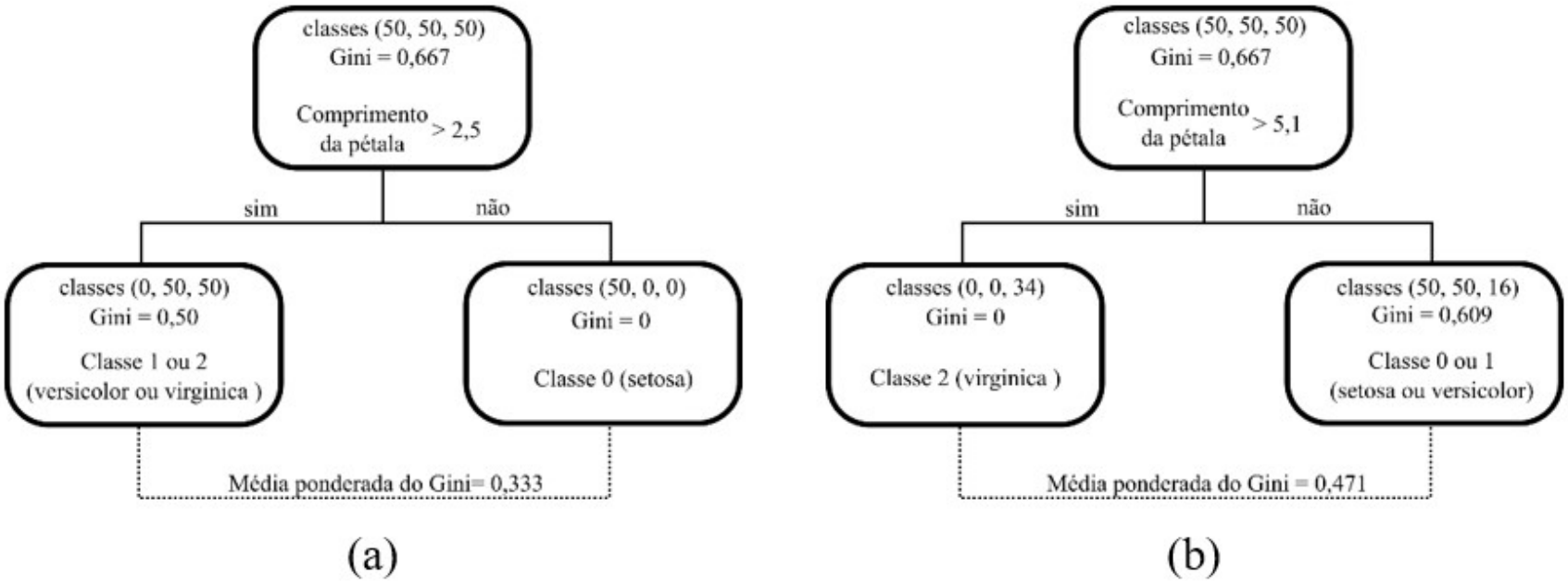
Decision Tree

- Basicamente, o algoritmo seleciona um atributo e determina um limiar para o atributo que melhor separa o conjunto de dados.
- O atributo que gerar o subconjunto com menor impureza será colocado no topo da árvore (nó raiz).
- Assim, irá ocorrer a subdivisão dos nós da árvore até que se chegue em uma decisão final (nó folha), na qual, não haverá mais subdivisões.
- A subdivisão poderá ser interrompida quando o resultado gerado tenha uma impureza maior que os dados originais.
- Outro critério de parada na subdivisão poderá ser pela definição de algum parâmetro, por exemplo, máxima profundidade da árvore (max_depth=5).

Decision Tree

- O resultado final do algoritmo Decision Tree é uma árvore com nós de decisão e folhas.
- Um exemplo de um nó de decisão é apresentado na Figura abaixo.
- Nesse exemplo foi utilizado o banco de dados flor de íris, e o atributo comprimento da pétala.
- O limiar ótimo para o comprimento da pétala é definido com base numa pontuação de impureza.
- Existem algumas formas de calcular a impureza, como por exemplo, coeficiente Gini, entropia, ID3, ou CART.
- Na Figura a seguir é apresentado o coeficiente Gini calculado para duas situações, com limiar do comprimento da pétala igual a 2,5 (Figura a) e limiar do comprimento da pétala igual a 5,1 (Figura b).

Decision Tree



Subdivisão do nó para o atributo comprimento da pétala para dois limiares (a) 2,5 cm e (b) 5,1 cm.

INSTITUTO FEDERAL
Sudeste de Minas Gerais

Decision Tree

- O coeficiente Gini é calculado com base na seguinte equação.

$$Gini = 1 - \sum_{c=1}^{n} {p_c}^2$$

- Em que c representa uma dada classe e n é o número de classes. No exemplo utilizado temos três classes diferentes, e p é a probabilidade da ocorrência de cada classe no nó. Veja um exemplo de cálculo do coeficiente de Gini para o nó raiz.

$$Gini = 1 - \left(\frac{50}{150}\right)^2 - \left(\frac{50}{150}\right)^2 - \left(\frac{50}{150}\right)^2 = 0{,}667$$

Decision Tree

- Na figura foram apresentados apenas dois limiares diferentes para o comprimento da pétala: 2,5 cm (Figura a) e 5,1 cm (Figura b).
- O algoritmo deverá determinar o limiar ótimo para cada atributo.
- O limiar de 2,5 cm gerou subdivisões com Gini de 0,50 e 0.
- O coeficiente Gini da divisão é calculado com base na média ponderada. Para a figura a) o cálculo é determinado da seguinte forma:

$$Gini = 0{,}5\frac{100}{150} + 0\frac{50}{150} = 0{,}333$$

- O mesmo cálculo foi realizado para o limiar de 5,1 cm, e gerou um coeficiente Gini de 0,471. Nesse caso, o algoritmo irá decidir que o melhor limiar será de 2,5 cm.

Decision Tree

- O cálculo será realizado para todos os atributos, enfim, aquele atributo que apresentar o menor coeficiente Gini (menor impureza) será definido como nó raiz da árvore.
- Na sequência novas subdivisões poderão ser realizadas ou folhas definidas.
- Conforme já mencionado, durante a modelagem é sempre interessante ajustar os parâmetros para limitar o modelo e assim evitar super ajuste.
- Na árvore de decisão existem vários parâmetros. Um importante parâmetro é a profundidade da árvore, *max_depth*. A validação cruzada poderá ser utilizada para definir os melhores valores para os parâmetros.

Decision Tree

- Uma vantagem do algoritmo Decision Tree é a facilidade de interpretação da árvore gerada, ou seja, é fácil extrair conhecimento da base de dados.
- Uma desvantagem é que o algoritmo Decision Tree é guloso, isso significa que ele cria uma árvore colocando no topo o atributo que “melhor” separa o conjunto de dados.
- No entanto, essa nem sempre é a melhor solução.

Random Forest

- Geralmente algoritmos de *Machine Learning* apresentam melhor acurácia quando utilizamos mais de um algoritmo para tomada de decisão.
- Também pode-se conseguir uma melhor acurácia gerando-se vários modelos com um mesmo algoritmo modificando a fonte de dados.
- É com base nessa estratégia que funciona o algoritmo *Random Forest*.
- Basicamente ele gera vários modelos de *Decision Tree* com base na escolha aleatória de parte dos dados de treinamento.
- O algoritmo *Random Forest* apresenta os seguintes passos:

1) construir um subconjunto de dados aleatórios a partir dos dados de treinamento;
2) selecionar aleatoriamente um subconjunto de atributos a partir do subconjunto no passo 1;
3) criar uma árvore de decisão a partir do subconjunto;
4) repetir o passo 1, 2 e 3 para gerar várias árvores diferentes.

INSTITUTO FEDERAL
Sudeste de Minas Gerais

Random Forest

- O número de árvores geradas (n_estimator) é um parâmetro que pode ser definido.
- A classificação final será definida para a classe que obtiver o maior número de votos.
- Se for regressão, o resultado poderá ser obtido pela média das predições de cada árvore.
- O algoritmo *Random Forest* apresenta grande capacidade de aprendizagem (ajuste). Dessa forma, para evitar super ajuste, recomenda-se limitar o grau de liberdade do algoritmo.
- Para isso deve-se definir alguns parâmetros do algoritmo. Dois importantes parâmetros são o número de árvores geradas (*n_estimator*) e a profundidade máxima da árvore (*max_depth*).
- A escolha dos melhores parâmetros deverá ser feita com base na validação cruzada.

INSTITUTO FEDERAL
Sudeste de Minas Gerais