Introdução aos Métodos de Árvore de Decisão

Tarefa de leitura

Capítulo 8 do livro
Introduction to Statistical Learning
por Gareth James, et al.

Métodos de Árvore de Decisão

Vamos começar com um experimento mental para dar alguma motivação por trás do uso de um método de árvore de decisão.

Métodos de Árvore de Decisão

Imagine que eu jogo tênis todos os sábados e sempre convido um amigo para vir comigo.

Às vezes meu amigo aparece, às vezes não.

Para ele depende de uma variedade de fatores, como: clima, temperatura, umidade, vento etc.

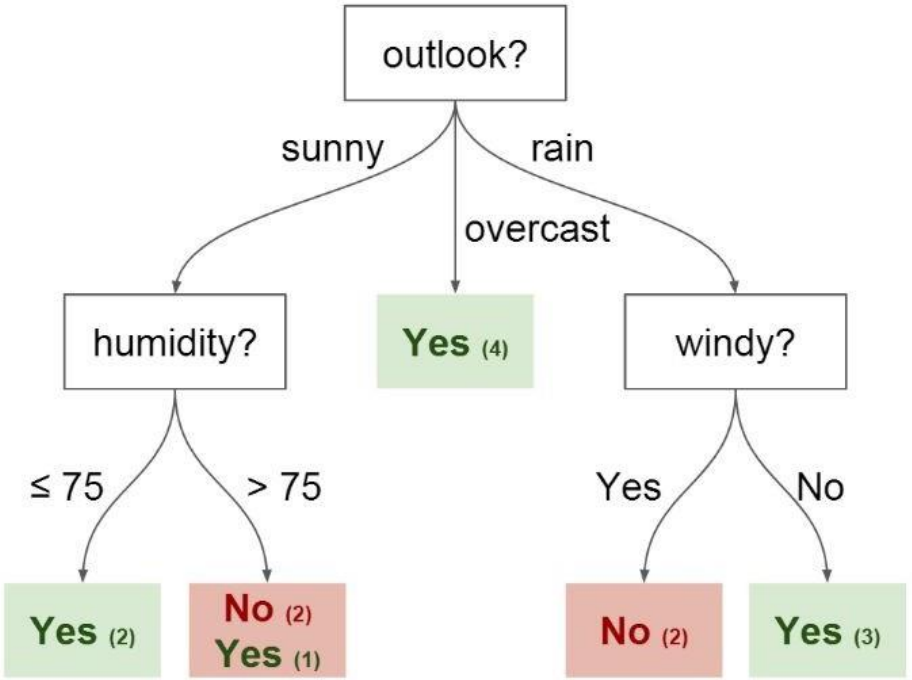
Começo a acompanhar essas características e se ele apareceu ou não para jogar comigo.

Temperature	Outlook	Humidity	Windy	Played?
Mild	Sunny	80	No	Yes
Hot	Sunny	75	Yes	**No**
Hot	Overcast	77	No	Yes
Cool	Rain	70	No	Yes
Cool	Overcast	72	Yes	Yes
Mild	Sunny	77	No	**No**
Cool	Sunny	70	No	Yes
Mild	Rain	69	No	Yes
Mild	Sunny	65	Yes	Yes
Mild	Overcast	77	Yes	Yes
Hot	Overcast	74	No	Yes
Mild	Rain	77	Yes	**No**
Cool	Rain	73	Yes	**No**
Mild	Rain	78	No	Yes

Métodos de Árvore de Decisão

Eu quero usar esses dados para prever se ele vai ou não aparecer para jogar.

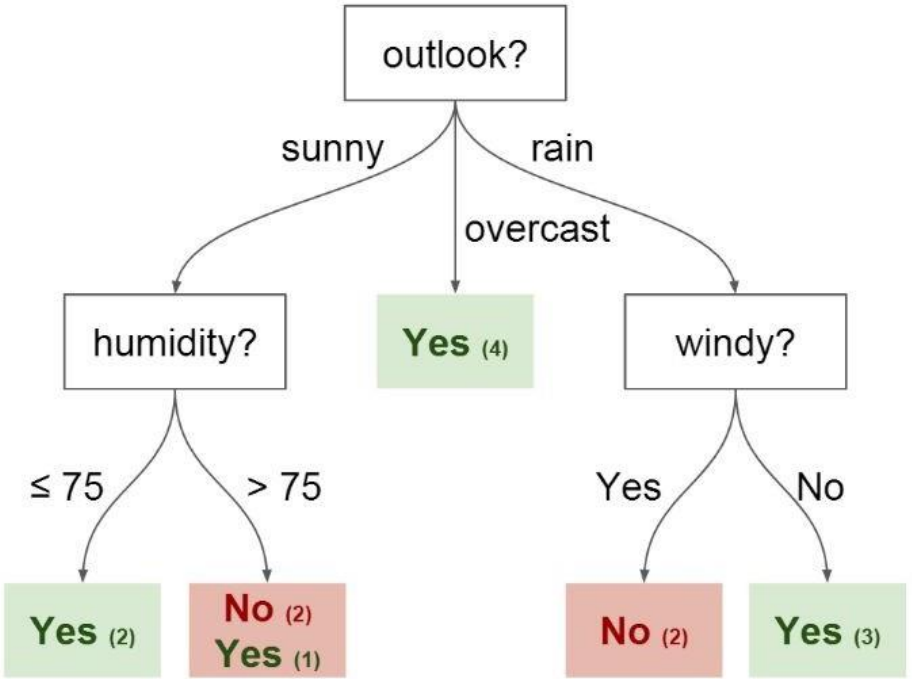
Uma maneira intuitiva de fazer isso é por meio de uma árvore de decisão



Métodos de Árvore de Decisão

Nesta árvore temos:

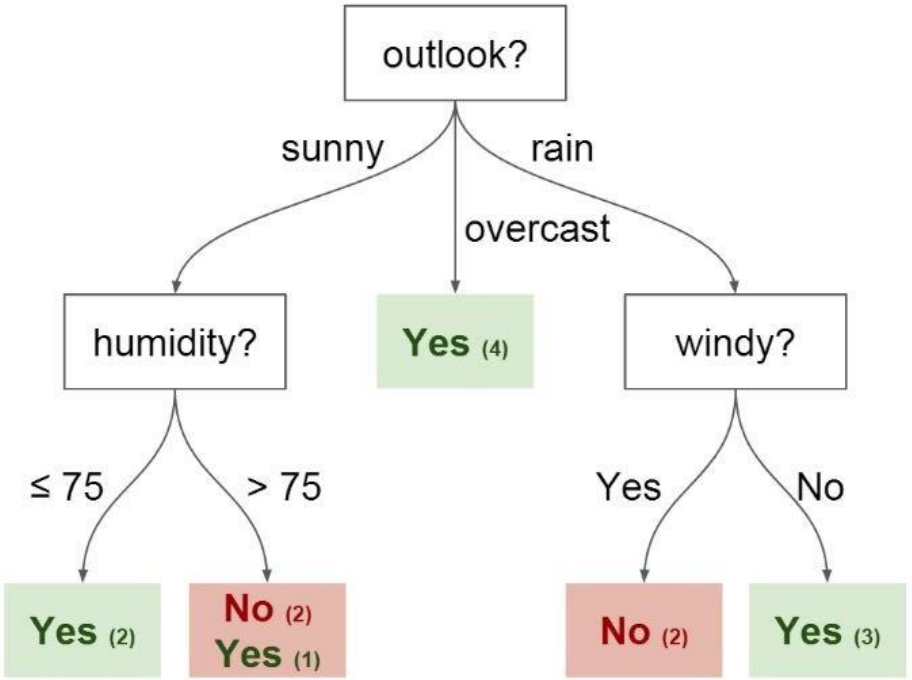
- Nós (Nodes)
  - Dividir pelo valor de um determinado atributo
- Vértices (Edges)
  - Resultado de uma divisão para o próximo nó



Métodos de Árvore de Decisão

Nesta árvore temos:

- Raiz (Root)
  - O nó que executa a primeira divisão
- Folhas (Leaves)
  - Nós terminais que preveem o resultado



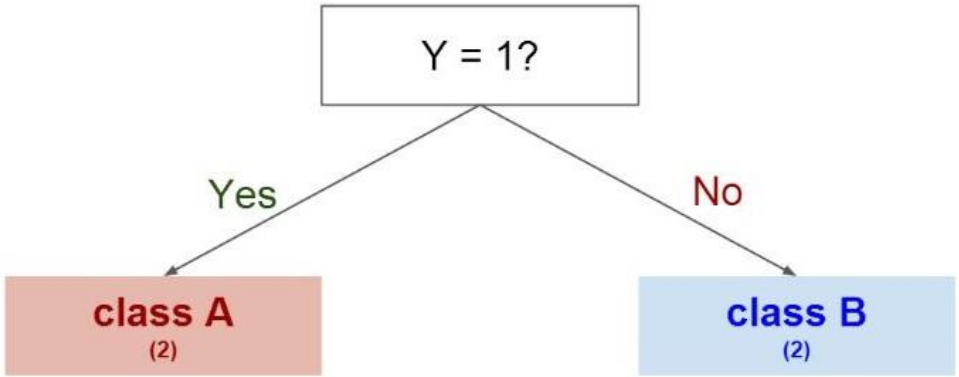
Intuição por trás das divisões

Dados imaginários com três características (X,Y e Z) e com duas classes possíveis (A e B).

X	Y	Z	Class
1	1	1	A
1	1	0	A
0	0	1	B
1	0	0	B

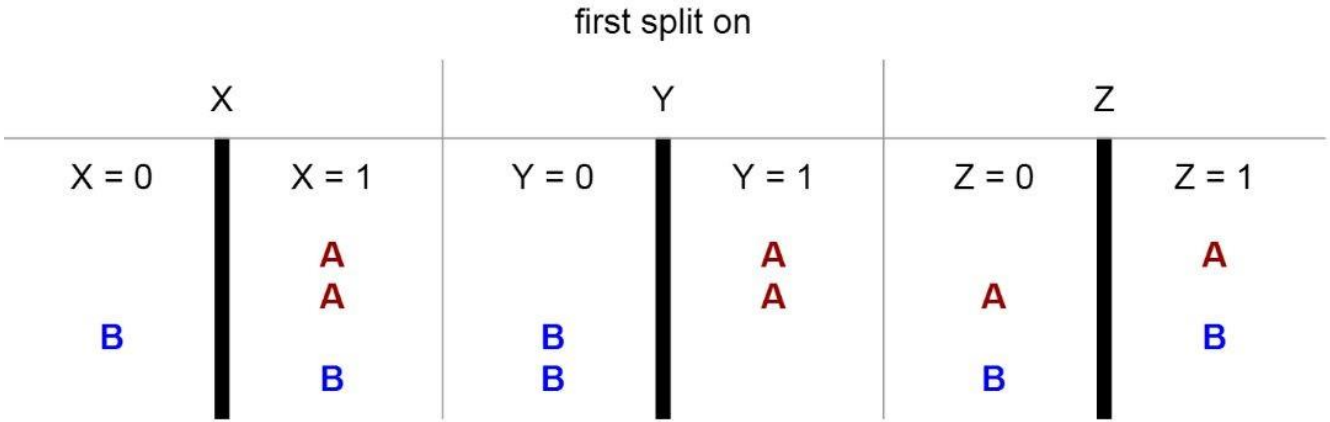
Intuição por trás das divisões

Dividir em Y nos dá uma separação clara entre as classes



Intuição por trás das divisões

Poderíamos ter tentado dividir em outra característica primeiro:



Intuição por trás das divisões

Entropia e Ganho de Informação são os métodos matemáticos para escolher a melhor divisão.

$$Entropy:$$

$$H(S) = -\sum_{i} p_i(S)\, log_2\, p_i(S)$$

$$Information\ Gain:$$

$$IG(S, A) = H(S) - \sum_{v \in Values(A)} \frac{|S_v|}{S} H(S_v)$$

Intuição por trás das divisões

Para melhorar o desempenho, podemos usar muitas árvores com uma amostra aleatória de recursos escolhidos como divisão.

- Uma nova amostra aleatória de características é escolhida **para cada árvore em cada divisão.**
- Para **classificação**, m é tipicamente escolhido como a raiz quadrada de p.

Florestas Aleatórias (Random Forests)

Qual é o objetivo?

- Suponha que haja **uma característica muito forte** no conjunto de dados. Ao usar árvores “ensacadas (bagged)”, a maioria das árvores usará esse recurso como a divisão superior, resultando em um conjunto de árvores semelhantes que são **altamente correlacionadas.**

Florestas Aleatórias (Random Forests)

Qual é o objetivo?

- A média de quantidades altamente correlacionadas não reduz significativamente a variância.
- Ao excluir aleatoriamente algumas características candidatas de cada divisão, o método Random Forests "descorrelaciona" as árvores, de modo que o processo de média pode reduzir a variação do modelo resultante.