Arvore de Decisão e Floresta Aleatória

Instituto Federal de São Paulo

Árvore de Decisão

- Imagine que jogo futebol sempre aos sábados e convido um amigo para vir comigo.
- As vezes ele aparece e as vezes ele não aparece.
- Para ele, ir ao jogo depende de um conjunto de fatores tais como: clima, temperatura, umidade, vento, etc.
- Eu começo a tomar nota dos dias em que ele aparece e como as variáveis se comportam naquele dia.

Árvore de Decisão

Tempo	Temperatura	Umidade	Vento	Jogou
Médio	Sol	80	Não	Sim
Quente	Sol	75	Sim	Não
Quente	Nublado	77	Não	Sim
Frio	Chuva	70	Sim	Não
Frio	Nublado	72	Sim	Sim
Médio	Sol	77	Não	Não
Frio	Sol	70	Não	Sim
Médio	Chuva	69	Não	Sim
Médio	Sol	65	Sim	Sim
Médio	Nublado	77	Sim	Sim

Árvore de Decisão

- Por meio dos dados que obtive vou usá-lo para predizer se o meu colega jogará ou não.
- Uma forma intuitiva de se fazer isso é por meio da criação de uma árvore de decisão.
- Pego as variáveis e faço uma divisão dela para saber se meu colega participará ou não.

Árvore de Decisão

- Uma arvore de decisão é formada por: nós, ramos, raiz e folha.
- Nós são os pontos de interseção onde temos uma pergunta.
- Folhas são os nós finais que determinam a tomada de decisão.
- Raiz é o nó que origina a primeira divisão.
- Ramos são as saídas de um nó para outro nó.

Árvores de Decisão

Imagine um conjunto de dados com 3 parâmetros (X,Y e Z) com duas classes possíveis:

X	Y	Z	Classe
1	1	1	A
1	1	0	A
0	0	1	B
1	0	0	B

Nesse caso o conjunto Y nos dá a separação perfeita dos dados. As outras variáveis não são tão eficientes.

Árvore de Decisão

- Neste caso matematicamente vamos buscar dentre as variáveis a que tem mais entropia e ganho de informação para escolher quais dados farão parte da árvore de decisão.
- O algoritmo fará a divisão sempre na classe que apresentar o maior ganho de informação.

Florestas Aleatórias

- As árvores de decisão podem apresentar muito overfit.
- Para resolver isso podemos criar árvores usando divisões diferentes do conjunto de dados.
- Para melhorar o desempenho das árvores de decisão, podemos também usar amostras aleatórias de parâmetros de divisão:
  - Um novo conjunto de parâmetros é escolhido aleatoriamente para cada árvore a cada divisão da mesma.
  - O algoritmo escolhe 1 dentre ‘m’ parâmetros.

Exemplo com Python

- Vamos explorar os dados de pacientes com cifose e vamos tentar predizer se uma cirurgia será ou não suficiente para cada caso.