INSTITUTO DO EMPREGO E FORMAÇÃO PROFISSIONAL, IP
DELEGAÇÃO REGIONAL DE LISBOA E VALE DO TEJO
CENTRO DE EMPREGO E FORMAÇÃO PROFISSIONAL DE SINTRA

UFCD – 10810

Fundamentos do desenvolvimento de modelos analíticos em Python

5 – Árvores de Decisão e Florestas Aleatórias

Carga horária: 25 horas

Formador: Manuel Viana

1

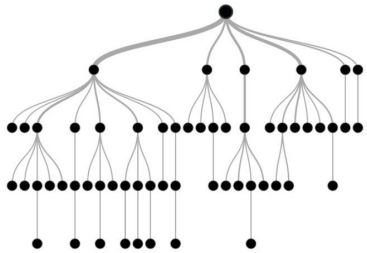
Introdução

- Os algoritmos de aprendizagem baseados em árvores de decisão são considerados dos melhores e mais utilizados métodos de aprendizagem supervisionada.
- Os métodos baseados em árvores dão-nos modelos preditivos de alta precisão, estabilidade e facilidade de interpretação.
- Podem ser adaptados para resolver vários tipos de problemas (classificação ou regressão).

2

Árvores de Decisão

- É um algoritmo de aprendizagem supervisionada (com uma variável alvo pré-definida), muito utilizado em problemas de classificação.
- Funciona para ambas as variáveis categóricas e contínuas de entrada e de saída.
- Na árvore de decisão, divide-se a população ou amostra em dois ou mais conjuntos homogéneos (ou sub-populações) com base nos divisores/diferenciadores mais significativos das variáveis de entrada.

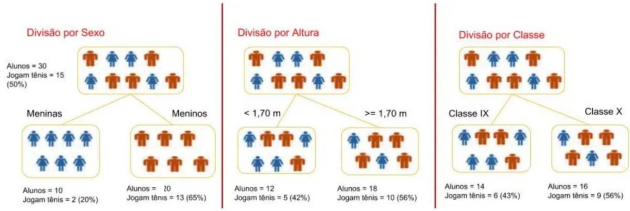


3

3

Árvores de Decisão – Exemplo 1

- Considere-se que uma amostra de 30 alunos tem três variáveis:
  - Sexo (menino ou menina)
  - Classe/Turma (IX ou X)
  - Altura (160 cm a 180 cm)
- E 15 dos 30 alunos, jogam ténis no recreio.
- Como é que podemos criar um modelo para prever quem vai jogar ténis durante o recreio?
  - Neste problema, precisamos dividir os alunos que jogam ténis no recreio com base nas três variáveis à disposição. É aqui que entra a Árvore de Decisão.

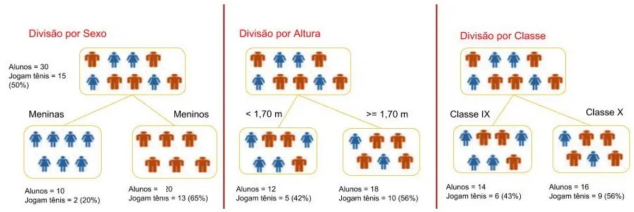


4

4

Árvores de Decisão – Exemplo 1

- A Árvore de Decisão divide os alunos com base nos valores das três variáveis e identifica a variável que cria os melhores conjuntos homogéneos de alunos (que são heterogêneos entre si).
- Neste exemplo, pode-se observar que a variável Sexo é capaz de identificar os melhores conjuntos homogéneos, comparativamente com as variáveis Altura e Classe/Turma.



5

5

Árvores de Decisão – Exemplo 2

- Considere-se que uma determinada pessoa joga Ténis aos Sábados e convida sempre um amigo para ir com ela.
- Às vezes o amigos vai, outras vezes não.
- Para o amigo, ir ao jogo depende de vários fatores:
  - Tempo
  - Temperatura
  - Humidade
  - Vento
- DataSet:

Temperatura	Tempo	Humidade	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
Quente	Nublado	74	Sim	Sim

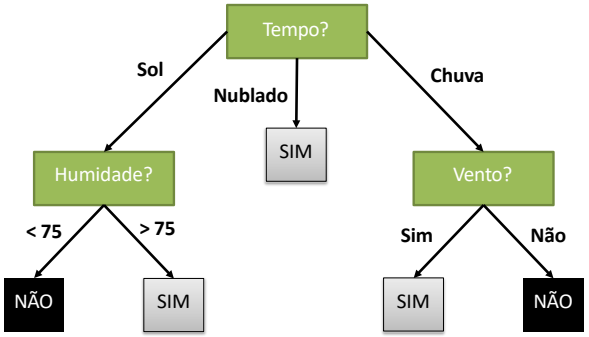
- Pretende-se prever se o amigo vai jogar ou não…

6

6

Árvores de Decisão – Exemplo 2

- Possível árvore de decisão (apenas para exemplo):

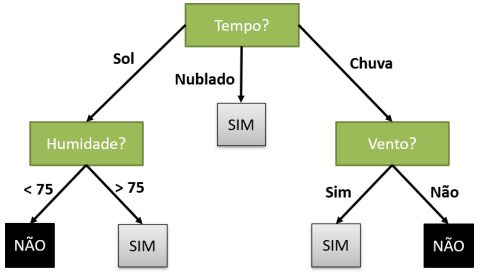


7

7

Árvores de Decisão – Exemplo 2

- **Terminologia:**
  - Nó Raíz – representa a população inteira ou amostra. É ainda dividido em dois ou mais conjuntos homogéneos.
  - Divisão – é o processo de dividir um nó em dois ou mais sub-nós.
  - Nó de Decisão – quando um sub-nó é dividido em sub-nós adicionais.
  - Folha ou Nó de Término – nós não divididos. Nó final que toma a decisão.
  - Poda – processo de remover sub-nós de um nó de decisão. Processo oposto à divisão.



8

8

Floresta Aleatória

- Na floresta aleatória, utilizam-se múltiplas árvores, em vez de uma única árvore.
- Para classificar um novo objeto baseado em atributos, cada árvore dá uma classificação, que é como se a árvore desse "votos" para essa classe.
- A floresta escolhe a classificação que tiver mais votos (de todas as árvores da floresta) e, em caso de regressão, considera a média das saídas por árvores diferentes.

9

9

Floresta Aleatória – modo de funcionamento

- Cada árvore é plantada e cultivada da seguinte forma:
  1. Assuma-se que o número de casos no conjunto de treino é **N**. Então, a amostra desses **N** casos é escolhida aleatoriamente, mas com substituição. Esta amostra será o conjunto de treino para o cultivo da árvore.
  2. Se houver **M** variáveis de entrada, um número **m < M** é especificado de modo que, em cada nó, **m** variáveis de **M** sejam selecionadas aleatoriamente. A melhor divisão nestes **m** é usada para dividir o nó. O valor de **m** é mantido constante enquanto crescemos a floresta.
  3. Cada árvore é cultivada na maior extensão possível e não há poda.
  4. Prever novos dados agregando as previsões das árvores ntree (ou seja, votos maiores para classificação, média para regressão).

10

10