O treinamento da Decision Tree

Ciclo	Ciclo 06: Algoritmos baseado em árvores
# Aula	42
Created	@March 9, 2023 11:54 AM
Done	☑
Ready	☑

Objetivo da Aula:

- [ ] A visão geral do treinamento
- [ ] A impureza de uma árvore
- [ ] A escolha do atributo na separação
- [ ] Os 6 passos do treinamento da Decision Tree
- [ ] Resumo
- [ ] Próxima aula

Conteúdo:

▼ 1. A visão geral do treinamento

Os algoritmos de árvore de decisão exigem que o conjunto de dados seja dividido em subconjuntos menores em cada nó da árvore. Para isso, é necessário escolher o melhor atributo para dividir o conjunto de dados. A escolha é feita com base em uma medida de impureza.

O objetivo é minimizar a impureza do conjunto de dados nos subconjuntos resultantes. As medidas de impureza mais comuns são a entropia, o índice Gini e o ganho de informação.

O atributo que resulta na maior redução de impureza é escolhido para a divisão. Esse processo é repetido recursivamente em cada nó da árvore até que os subconjuntos resultantes sejam puros ou até que um critério de parada seja atingido.

2. A impureza de uma árvore

A impureza que determina quão misturados estão os dados em relação à classe de saída. Por exemplo, um conjunto de dados com 100 elementos, 50 da classe A e 50 da classe B, são subdivididos subsequentemente ao ponto dos 100 elementos originais serem distribuídos entre as folhas.

Um dessas folhas possui 10 elementos da Classe A e 0 elementos da classe B. Essa folha possui um grau de impureza de zero, pois todos os seus elementos pertencem a única classe.

Por outro lado, uma dessas folhas possui 3 elementos da Classe A e 5 elementos da classe B. Essa folha tem um certo grau de impureza, pois 37.5% dos seus elementos pertencem a Classe A e 62.5% pertencem a Classe B.

Para determinar o grau de impureza de uma folha, são utilizados algumas medidas, como por exemplo:

1. O critério de Gini
2. Entropia
3. Ganho de informação.

2.1 O critério (índice) Gini

$$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^2$$

$$p_{i,k} \text{ é a probabilidade de cada classe do } i_{th} \text{ nó}$$

2.2 Entropia

A entropia é uma medida de impureza que quantifica a incerteza associada à distribuição de probabilidade de um conjunto de dados. Quanto maior a entropia, maior a incerteza. A fórmula da entropia para o cálculo da impureza de uma folha da Decision Tree é:

$$H_i = -\sum_{k=1}^{n} p_{i,k} \log_2 p_{i,k}$$

$$onde: p_{i,k}\ é\ a\ probabilidade\ de\ cada\ classe\ do\ i-ésimo\ nó$$

▼ 2.3 Information Gain

O ganho de informação é uma medida de quanta informação é ganha sobre a classe de saída ao dividir um conjunto de dados de acordo com um atributo específico. A fórmula do ganho de informação para o cálculo da impureza de uma folha da Decision Tree é:

$$IG(D_p, a) = I(D_p) - \sum_{j=1}^{v} \frac{N_j}{N_p} I(D_j)$$

$$onde: I(D_p)\ é\ a\ impureza\ de\ D_p$$
$$N_p\ é\ o\ número\ total\ de\ exemplos\ em\ D_p$$
$$a\ é\ o\ atributo\ considerado$$
$$v\ é\ o\ número\ de\ valores\ distintos\ de\ a$$
$$D_j\ é\ o\ subconjunto\ de\ D_p\ com\ valores\ distintos\ de\ a\ para\ j$$
$$N_j\ é\ o\ número\ total\ de\ exemplos\ em\ D_j$$
$$I(D_j)\ é\ a\ impureza\ de\ D_j$$

▼ 3. A escolha do atributo na separação

O algoritmo Decision Tree realiza o treinamento separando o conjunto de dados em duas partes, segundo um único atributo (k) e um valor (t_k), por exemplo, k=petal length e t_k < 2.45 cm, ou seja, “petal length < 2.45 cm”.

Ao realizar as subdivisões do conjunto de dados, o algoritmo busca minimizar o valor resultante da seguinte função:

$$J(k, t_k) = \frac{m_{left}}{m} G_{left} + \frac{m_{right}}{m} G_{right}$$

$$G_{left/right}\ mede\ a\ impureza\ subconjuntos\ a\ direita\ a\ esquerda$$
$$m_{left/right}\ é\ o\ número\ de\ exemplos\ no\ conjunto\ dados\ da\ direita\ a\ esquerda$$

J(idade, 20) = (20/150)*0.5 + (80/150)*0.0 = 0.1

J(salario, 15000) = (54/150)*0.168 + (46/150)*0.043 = 0.01

▼ 4. Os 6 passos do treinamento da Decision Tree

1. Escolha uma atributo (coluna) do conjunto de dados.
2. Para cada possível valor do atributo selecionado, use a função de custo para encontrar o valor da impureza da separação.
3. Repita os passos 1 e 2 para todas as combinações de atributo e valores, a fim de encontrar a combinação atributo-valor que retorne o menor valor da função custo da Decision Tree.
4. Uma vez definido o para atributo-valor, faça a separação do conjunto de dados em dois nós filhos.
5. Repita os passos de 1 a 3, para encontrar a segunda combinação atributo-valor para causar uma nova separação dos dados.
6. Repita o processo 5 até os valores dos parâmetros serem atendidos.

▼ 5. Resumo

1. O processo de treinamento da Decision Tree se resume a subdividir os dados originais em nós filhos, a fim de minimizar a impureza dos nós filhos (folhas).
2. A impureza de uma Decision Tree pode ser medida utilizando os critérios de Gini, Entropia ou Ganho de Informação.
3. A Decision Tree possui 6 passos para o seu treinamento.

▼ 6. Próxima aula

Decision Tree - prática