Árvores de Decisão

10 de julho de 2023

1 O que é uma árvore de decisão?

Uma árvore de decisão é um algoritmo de aprendizagem de máquina supervisionado utilizado para resolver problemas de classificação e regressão. Ela é chamada de "árvore"porque tem uma estrutura hierárquica semelhante a uma árvore invertida, com um único nó raiz, nós intermediários (também conhecidos como nós internos) e nós folha.

A construção de uma árvore de decisão envolve a divisão recursiva do conjunto de dados de treinamento com base em características (ou atributos) relevantes. O objetivo é criar partições que sejam o mais puras possível em relação à classe alvo, ou seja, que tenham exemplos de uma única classe. Essa divisão é feita com base em critérios de divisão, como a entropia ou o índice Gini, que medem a impureza dos dados em uma determinada partição.

Uma vez construída a árvore, ela pode ser utilizada para fazer previsões em dados de teste ou em novos exemplos. Cada exemplo percorre a árvore, seguindo os caminhos definidos pelas decisões tomadas nos nós intermediários, até chegar a um nó folha, que corresponde à classe prevista para aquele exemplo.

As árvores de decisão possuem várias vantagens, como a capacidade de lidar com dados numéricos e categóricos, interpretabilidade e facilidade de visualização. No entanto, elas também podem ser suscetíveis a problemas como overfitting (ajuste excessivo aos dados de treinamento) e sensibilidade a pequenas variações nos dados de entrada. Diversas técnicas, como a poda da árvore e o uso de conjuntos de árvores, como o Random Forest, são empregadas para mitigar esses problemas e melhorar o desempenho das árvores de decisão.

2 Quais são os componentes de uma árvore de decisão?

Uma árvore de decisão é composta por três componentes principais: nós, arestas e rótulos.

1. Nós:
   (a) Nó Raiz: É o ponto de partida da árvore e representa o conjunto completo de dados de treinamento. Ele é dividido em nós intermediários (nós internos) com base num critérios de divisão que deve ser determinado pelo algoritmo de treinamento.
   (b) Nós Intermediários: Representam as decisões tomadas durante a construção da árvore. Cada nó intermediário também é associado a um critérios de divisão.
   (c) Nós Folha: Representam as classes ou valores previstos pela árvore de decisão. São as folhas da árvore, onde não há mais subdivisões. Cada nó folha é rotulado com a classe ou valor de destino correspondente.
2. Critérios de decisão:
   (a) Atributo de decisão: É o atributo escolhido para realizar a decisão no nó atual da árvore. O critério de decisão seleciona o atributo que melhor separa os dados com base em alguma medida de impureza ou ganho de informação.

(b) Valor de decisão: É o valor ou limiar do atributo escolhido que define a condição de decisão. Os exemplos cujo valor do atributo é maior ou igual ao valor de decisão são direcionados para um ramo, enquanto aqueles cujo valor do atributo é menor são direcionados para o outro ramo.

(c) Regra de decisão: É a regra ou condição definida pela combinação do atributo de decisão e valor de decisão. Essa regra especifica como os exemplos são divididos em ramos filhos. Por exemplo, "atributo idade >= 30"é uma regra de decisão que divide os exemplos em dois ramos com base no valor do atributo idade.

3. Arestas:

   - As arestas conectam os nós e representam as decisões tomadas com base nos valores dos atributos.

4. Rótulos:

   - Os rótulos são atribuídos aos nós folha e representam as classes ou valores previstos pela árvore de decisão para um exemplo de entrada específico. Eles indicam a decisão final tomada pela árvore com base nos atributos do exemplo.

Durante a construção da árvore, o algoritmo de aprendizagem de máquina divide recursivamente o conjunto de dados com base em critérios de decisão, como a entropia ou o índice Gini. Essa divisão é feita nos nós intermediários, de modo a criar partições puras em relação à classe alvo. A estrutura hierárquica da árvore permite que ela seja facilmente interpretada e seguida para fazer previsões em novos exemplos, percorrendo os caminhos definidos pelos nós intermediários até chegar a um nó folha com a classe prevista.

3 Treinamento de Árvore de Decisão

3.1 Função CalcularImpureza

1. Calcular a impureza ou a métrica de qualidade usando os rótulos/classes dos exemplos
2. Retornar o valor da impureza ou métrica de qualidade

Medidas comuns de impureza são:

- Gini

$$Gini(p_1, p_2, ..., p_k) = 1 - \sum_{i=1}^{k} p_i^2 \tag{1}$$

onde $p_1, p_2, ..., p_k$ são as probabilidades de pertence à cada classe e k é o número de classes. A fórmula do índice de Gini é utilizada para medir a impureza ou heterogeneidade de um conjunto de dados. Quanto menor o valor do índice de Gini, mais puro e homogêneo é o conjunto, indicando uma separação mais clara entre as classes. Um valor próximo de 0 indica uma divisão perfeitamente pura, enquanto um valor próximo de 1 indica uma divisão impura ou heterogênea.

- Entropia

$$Entropia(p_1, p_2, ..., p_k) = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

Essa fórmula calcula a entropia com base nas probabilidades $p_1, p_2, ..., p_k$, que representam a proporção de exemplos pertencentes a cada classe em relação ao total de exemplos.

Quanto maior for a entropia, maior é a incerteza ou impureza do conjunto de dados, indicando uma distribuição mais equilibrada entre as classes. Por outro lado, uma entropia próxima de zero indica um conjunto de dados puro, onde todos os exemplos pertencem à mesma classe.

3.2 Função EncontrarMelhorDivisao

1. Inicializar a melhor impureza/métrica de qualidade como um valor alto (ou baixo, dependendo do critério)
2. Para cada atributo:
   (a) Para cada valor único do atributo:
      i. Dividir os exemplos em dois conjuntos (exemplos da esquerda e exemplos da direita) com base no valor do atributo
      ii. Calcular a impureza/métrica de qualidade dessa divisão
      iii. Se a impureza/métrica de qualidade for melhor do que a melhor impureza/métrica de qualidade atual:
         A. Atualizar a melhor impureza/métrica de qualidade
         B. Armazenar o atributo e valor do atributo que resultaram na melhor divisão
         C. Armazenar os exemplos da esquerda e da direita resultantes na melhor divisão
3. Retornar o atributo e valor do atributo que resultaram na melhor divisão, juntamente com os exemplos da esquerda e da direita

3.3 Função ConstruirArvore

1. Se todos os exemplos pertencerem a uma única classe:
   (a) Retornar um nó folha rotulado com a classe
2. Se não houver atributos restantes para divisão:
   (a) Retornar um nó folha rotulado com a classe mais frequente nos exemplos
3. Encontrar a melhor divisão dos exemplos usando a função EncontrarMelhorDivisao
4. Criar um nó intermediário com o atributo e valor do atributo da melhor divisão
5. Dividir os exemplos em dois conjuntos com base na melhor divisão
6. Recursivamente construir a subárvore da esquerda chamando ConstruirArvore(exemplosEsquerda)

7. Recursivamente construir a subárvore da direita chamando ConstruirArvore(exemplosDireita)
8. Adicionar as subárvores como filhos do nó intermediário
9. Retornar o nó intermediário

3.4 Chamada da função principal para construir a árvore de decisão

1. exemplos = conjunto de exemplos de treinamento
2. arvore = ConstruirArvore(exemplos)

4 Exemplo

Considere a base de dados a seguir.

Exemplo	Pelo	Som	Classe
1	Curto	Latido	Cachorro
2	Curto	Miado	Gato
3	Curto	Latido	Cachorro
4	Longo	Miado	Gato
5	Longo	Latido	Cachorro
6	Curto	Latido	Cachorro
7	Longo	Miado	Gato

Tabela 1: Base de dados fictícia para a classificação de animais

5 Exercícios

1. Exercício de Gini: Considere um conjunto de dados com 100 exemplos, dos quais 60 pertencem à classe A e 40 pertencem à classe B. Calcule o índice de Gini desse conjunto de dados.
2. Exercício de Gini: Em um conjunto de dados com 80 exemplos, dos quais 45 pertencem à classe X e 35 pertencem à classe Y. Calcule o índice de Gini desse conjunto de dados.
3. Exercício de Critério de Divisão: Suponha que um conjunto de dados seja dividido em dois subconjuntos, onde o subconjunto A contém 30 exemplos, dos quais 20 pertencem à classe P e 10 pertencem à classe Q, e o subconjunto B contém 70 exemplos, dos quais 40 pertencem à classe P e 30 pertencem à classe Q. Calcule o ganho de Gini para essa divisão com base no índice de Gini inicial do conjunto de dados.
4. Exercício de Construção de Árvore de Decisão: Considere um conjunto de dados com duas características, "Altura"(com valores "Alto"e "Baixo") e "Idade"(com valores "Jovem"e "Adulto"), e uma classe "Classe"(com valores "A"e "B").

   Considere a seguinte tabela de dados:

Altura	Idade	Classe
Alto	Jovem	A
Alto	Adulto	A
Baixo	Jovem	B
Baixo	Adulto	B
Alto	Jovem	B
Baixo	Adulto	A

Construa uma árvore de decisão para classificar os exemplos com base nessas características, usando o critério de Gini.

5. Exercício de Avaliação de Divisões: Dado um conjunto de dados com 100 exemplos, onde 70 pertencem à classe X e 30 pertencem à classe Y, avalie duas divisões possíveis com base no índice de Gini, e determine qual delas é mais preferível em termos de impureza.

6 Soluções

1. Resposta para o exercício de Gini:

   O índice de Gini para um conjunto de dados pode ser calculado usando a fórmula:

$$Gini(p) = 1 - \sum_{i=1}^{n} p_i^2$$

   onde p_i é a proporção da classe i no conjunto de dados, e n é o número de classes.

   Para o primeiro conjunto de dados, a proporção da classe A é 0.6 e da classe B é 0.4. Portanto, o índice de Gini é:

$$Gini(p) = 1 - (0.6^2 + 0.4^2) = 1 - (0.36 + 0.16) = 1 - 0.52 = 0.48$$

   Para o segundo conjunto de dados, a proporção da classe X é 0.5625 e da classe Y é 0.4375. Portanto, o índice de Gini é:

$$Gini(p) = 1-(0.5625^2+0.4375^2) = 1-(0.31640625+0.19140625) = 1-0.5078125 = 0.4921875$$

2. Resposta para o exercício de Critério de Divisão:

   Para calcular o ganho do índice de Gini para a divisão, primeiro precisamos calcular o índice de Gini para cada subconjunto.

   Para o subconjunto A, a proporção da classe P é $\frac{2}{3}$ e da classe Q é $\frac{1}{3}$. Portanto, o índice de Gini para A é:

$$Gini(A) = 1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right) = 1 - \left(\frac{4}{9} + \frac{1}{9}\right) = 1 - \frac{5}{9} = \frac{4}{9} \approx 0.44444$$

   Para o subconjunto B, a proporção da classe P é $\frac{4}{7}$ e da classe Q é $\frac{3}{7}$. Portanto, o índice de Gini para B é:

$$Gini(B) = 1 - \left(\left(\frac{4}{7}\right)^2 + \left(\frac{3}{7}\right)^2\right) = 1 - \left(\frac{16}{49} + \frac{9}{49}\right) = 1 - \frac{25}{49} = \frac{24}{49} \approx 0.4898$$

O ganho do índice de Gini para a divisão é a diferença entre o índice de Gini inicial do conjunto de dados e a soma ponderada dos índices de Gini dos subconjuntos. Se assumirmos que o índice de Gini inicial é $Gini(p)$, então o ganho é:

$$Gini(p) - \left(\frac{30}{100}Gini(A) + \frac{70}{100}Gini(B)\right)$$

3. Resposta para o exercício de Construção de Árvore de Decisão:

   Claro! Vamos considerar um conjunto de dados pequeno para os dois últimos exercícios.

4. Exercício de Construção de Árvore de Decisão:

   Considere a seguinte tabela de dados:

Altura	Idade	Classe
Alto	Jovem	A
Alto	Adulto	A
Baixo	Jovem	B
Baixo	Adulto	B
Alto	Jovem	B
Baixo	Adulto	A

   A árvore de decisão seria:

   1. Primeiro, dividiríamos os dados com base na característica "Altura". Isso porque a altura é capaz de separar as classes A e B mais do que a característica "Idade".

```
Raiz
  |--> [Altura = Alto]: A (2), B (1)
  |--> [Altura = Baixo]: A (1), B (2)
```

   2. Como ainda há impureza em ambas as divisões, precisaríamos continuar dividindo. Para a divisão [Altura = Alto], a característica "Idade"pode ser usada para uma separação perfeita.

```
[Altura = Alto]
  |--> [Idade = Jovem]: A (1), B (1)
  |--> [Idade = Adulto]: A (1)
```

   3. Para a divisão [Altura = Baixo], a característica "Idade"também pode ser usada para uma separação perfeita.

```
[Altura = Baixo]
  |--> [Idade = Jovem]: B (1)
  |--> [Idade = Adulto]: A (1), B (1)
```

4. No final, a árvore de decisão completa seria:

```
Raiz
  |--> [Altura = Alto]
  |      |--> [Idade = Jovem]: A (1), B (1)
  |      |--> [Idade = Adulto]: A (1)
  |--> [Altura = Baixo]
         |--> [Idade = Jovem]: B (1)
         |--> [Idade = Adulto]: A (1), B (1)
```

5. Resposta para o exercício de Avaliação de Divisões:

   Para o conjunto de dados com 100 exemplos, onde 70 pertencem à classe X e 30 pertencem à classe Y, vamos considerar duas possíveis divisões.

   1. Divisão 1: 50 exemplos em cada subconjunto, com subconjunto 1 contendo 35 exemplos da classe X e 15 da classe Y, e subconjunto 2 contendo 35 exemplos da classe X e 15 da classe Y.

   2. Divisão 2: 70 exemplos no subconjunto 1, com 50 exemplos da classe X e 20 da classe Y, e 30 exemplos no subconjunto 2, todos pertencentes à classe X.

   Agora, vamos calcular o índice de Gini para cada divisão.

   - Para a Divisão 1, os índices de Gini dos dois subconjuntos são iguais (como eles têm a mesma distribuição de classes):

$$Gini_1 = 1 - \left(\left(\frac{35}{50} \right)^2 + \left(\frac{15}{50} \right)^2 \right) = 1 - (0.49 + 0.09) = 0.42$$

   - Para a Divisão 2, o índice de Gini para o primeiro subconjunto é:

$$Gini_{21} = 1 - \left(\left(\frac{50}{70} \right)^2 + \left(\frac{20}{70} \right)^2 \right) = 1 - \left(\frac{25}{49} + \frac{4}{49} \right) = \frac{20}{49} \approx 0.4082$$

   E o índice de Gini para o segundo subconjunto é zero (já que todos os exemplos pertencem à mesma classe).

   O Gini médio ponderado para cada divisão é:

   - Para a Divisão 1:

$$Gini_{M1} = \frac{1}{2} Gini_1 + \frac{1}{2} Gini_1 = Gini_1 = 0.42$$

   - Para a Divisão 2:

$$Gini_{M2} = \frac{70}{100} Gini_{21} + \frac{30}{100} \cdot 0 = Gini_{21} = 0.4082$$

   Portanto, a Divisão 2 é preferível, pois resulta em um índice de Gini médio ponderado mais baixo.