PUC Minas - Inteligência Artificial

Aluno: Otávio Augusto de Assis Ferreira Monteiro

Atividade: Lista 2

Questão 01

1) Uma árvore de decisão é gerada por meio de um algoritmo recursivo que busca dividir o conjunto de dados em subconjuntos cada vez mais puros. O atributo que está na raiz da árvore é aquele que possui o maior ganho de informação entre todos os atributos disponíveis no início do processo. Isso significa que ele é o atributo que, sozinho, melhor consegue separar os dados do conjunto de treinamento em suas respectivas classes.

2) Uma árvore de decisão gerada a partir de uma base de dados pode ser usada para tarefas de classificação e de regressão. Essencialmente, a árvore funciona como um modelo preditivo que infere uma "regra". Depois de treinada, ela pode ser usada para classificar novas instâncias de dados nunca vistas antes. Para isso, percorre-se a árvore a partir da raiz, realizando os testes indicados nos nós e seguindo os ramos correspondentes aos valores da nova instância, até chegar a um nó folha, que fornecerá a classificação (a predição) para essa nova instância.

3) Vantagens:
    - Interpretabilidade: São simples e fáceis de interpretar, pois as decisões podem ser visualizadas como uma série de regras lógicas.
    - Flexibilidade: São métodos não paramétricos, ou seja, não fazem suposições sobre a distribuição dos dados.

- Seleção de Atributos Embutida: O próprio processo de construção da árvore seleciona os atributos mais relevantes, tornando o modelo robusto contra atributos irrelevantes.
- Eficiência: A construção da árvore utiliza um algoritmo guloso, cuja complexidade de tempo é linear com o número de exemplos, sendo eficiente para grandes bases de dados.

Desvantagens:

- Instabilidade: Pequenas variações nos dados de treinamento podem gerar árvores muito diferentes.
- Tratamento de Atributos Contínuos: Exigem a ordenação dos dados, o que pode consumir até 70% do tempo de indução em grandes conjuntos de dados com muitos atributos contínuos. Algoritmos como o ID3 não lidam nativamente com eles, exigindo uma discretização prévia.
- Valores Ausentes: Algoritmos mais simples como o ID3 não tratam valores ausentes, exigindo mecanismos especiais ou algoritmos mais avançados como o C4.5.
- Confiabilidade das Folhas: Inferências feitas em nós próximos às folhas tendem a ser menos confiáveis do que aquelas feitas perto da raiz.

4) A avaliação da qualidade de uma árvore de decisão depende do tipo de tarefa para a qual ela foi construída. Para problemas de classificação, a principal ferramenta é a Matriz de Confusão. A partir dela, podemos calcular diversas métricas para avaliar o desempenho do modelo para cada classe.

   Agora, para problemas de regressão, utilizam-se métricas como o Erro Quadrático Médio (MSE) e o Erro Absoluto Médio (MAE).

5) As regras podem ser extraídas diretamente da estrutura da árvore. Cada caminho, da raiz até um nó folha, representa uma regra de classificação. A regra é formada pela conjunção (E lógico) de todos os testes encontrados ao longo do caminho.

Questão 2

1. Para determinar a raiz da árvore de decisão, o ganho de informação foi calculado para todos os dez atributos. O atributo Cliente apresentou o maior ganho, com um valor de 0.541, e por isso foi escolhido como a raiz da árvore.

2. A partir do nó raiz Cliente, a árvore se divide. Para os valores "Nenhum" e "Alguns", os subconjuntos de dados são puros, resultando diretamente em nós folha com as classificações "Não" e "Sim", respectivamente. Para o valor "Cheio", o subconjunto de dados é impuro e precisa de uma nova divisão. Ao calcular o ganho de informação para este subconjunto, foi encontrado um empate entre os atributos Fome, Tempo e Preço, todos com um ganho de 0.251. Escolhendo Fome como critério de desempate, este se torna o nó de decisão no segundo nível da árvore para o ramo onde o Cliente é "Cheio".