Inteligência Artificial - Lista 03

Perguntas

Questão 1

Explique e exemplifique classificação e regressão de dados.

Questão 2

Explique por que o parâmetro k no algoritmo K-NN pode influenciar o seu desempenho em um problema de classificação.

Questão 3

Quando se executa o algoritmo k-médias:

- Quando executado iniciando os centróides aleatoriamente, os grupos finais serão sempre os mesmos? Por quê? Isso é um problema? Se sim, como resolvê-lo?
- O índice de validação de agrupamento “Erro Quadrático” é um bom parâmetro para determinar qual o número ideal de grupos em uma base de dados? Justifique.

Questão 4

Considere o seguinte conjunto de dados, representando o diagnóstico de uma determinada doença por meio do resultado de 3 sintomas. De acordo com o algoritmo de treinamento de Árvores de decisão, qual seria o primeiro atributo escolhido? Justifique com os cálculos de ganho de informação.

Paciente	Sintoma 1	Sintoma 2	Sintoma 3	Diagnóstico
João	1	1	1	1
Maria	1	1	0	1
José	1	0	1	0
Ana	0	0	1	0
Antonio	1	0	0	1

Respostas

Questão 1

O aprendizado de máquina possui como premissa a similaridade do reconhecimento humano de padrões: criar uma hipótese a partir de uma experiência. Tanto a regressão quanto a classificação são tarefas de aprendizado preditivas, isto é, objetos possuem entrada e saída e são ambas caracterizadas como aprendizado supervisionado, mais rápido do que o aprendizado por reforço.

- Classificação: Rótulos discretos e não ordenado de valores (nominais), seu estimador é chamado classificador.
    - {doente, saudável}
    - {bom pagador, mau pagador}
    - {iris setosa, iris versicolor, iris virginica}
- Regressão: Rótulos contínuos , seu estimador é chamado regressor.
    - peso
    - temperatura
    - vazão de água

Questão 2

K é o parâmetro do algoritmo, sendo o número de vizinhos mais próximos. Caso ele seja muito grande:

1. Os vizinhos podem ser muito diferentes
2. Predição tendenciosa para a classe majoritária
3. Custo computacional mais elevado

Caso ele seja muito pequeno:

1. Não usufruir de informações suficientes
2. Previsão instável

Logo, é perceptível que o valor de K pode influenciar os problemas de classificação tanto se ele for acima do ideal quanto abaixo do recomendado: frequentemente são usados K pequenos e ímpares pois valores pares podem causar empates. É necessário analisar e avaliar o algoritmo empregado.

Questão 3

Os centróides iniciais são frequentemente iniciados aleatoriamente e os grupos finais podem ser alterados dependendo da configuração de atribuição, uma vez que depende da partição que foi designada em cada tentativa e das proximidades dos grupos aos centróides. Isso pode constituir um problema sim, por isso é importante que sejam feitas múltiplas execuções e que essas sejam avaliadas, também podemos fazer uma seleção informada de centróides ou aplicar algoritmos de busca.

Ao usar os grupos obtidos pela soma dos erros quadráticos é preciso tomar cuidado: ela serve para a comparação, porém, ao aumentar o número de partições, já tende a apresentar uma diminuição do SEQ. O ideal é usar essa prática em partições que englobam o mesmo número de grupos.

Questão 4

Árvore de decisão: Estratégia dividir para conquistar, isto é, repartir um problema complexo em problemas mais simples. É, formalmente, um grafo direcionado acíclico.

Os objetos são divididos pelo seu tipo de atributo e do número de divisões que o algoritmo suporta. Como os dados são qualitativos, o número de ramos é o número de possíveis valores.

Dividindo uma classificação: medida goodness of split ou, em casos onde a divisão não é tão visível, entropia em AD.

1. D = {3 +, 2-}
    a. $H(D) = -(3/5)*\log_2(3/5) -(2/5)*\log_2(2/5) = 0.9709505$

2. **Sintoma 1**:
    a. (Sim: Sim) = 3/4
       (Sim: Não) = 1/4
       $H = -(3/4)*\log_2(3/4) -(1/4)*\log_2(1/4) = 0,811278124$

    b. (Não: Não) = 1/1
       (Não: Sim) = 0/1
       H = 0

    c. 4/5 * 811278124 = 0,6490225

d. 0.9709505 - 0,6490225 = 0,321928

3. Sintoma 2:
    a. (Sim: Sim) = 2/2
       (Sim: Não) = 0/2
       H = 0

    b. (Não: Não) = 2/3
       (Não: Sim) = 1/3
       H = $-(2/3)*\log_2(2/3) -(1/3)*\log_2(1/3) = 0,923129023$

    c. 3/5 * 0,923129023 = 0,553877414
    d. 0,9709505 - 0,553877414 = 0,417073086

4. Sintoma 3:
    a. (Sim: Sim) = 1/3
       (Sim: Não) = 2/3
       H = $-(2/3)*\log_2(2/3) -(1/3)*\log_2(1/3) = 0,923129023$

    b. (Não: Não) = 0/2
       (Não: Sim) = 2/2
       H = 0

    c. 3/5 * 0,923129023 = 0,553877414
    d. 0,9709505 - 0,553877414 = 0,417073086

Por meio da análise dos cálculos de ganho de informação é possível perceber que o **Sintoma 1** seria o primeiro atributo a ser escolhido.