Setrem

Introdução à Inteligência Artificial

Pós-graduação em Cloud Computing: Infraestruturas, Plataformas e Serviços

Prof. João Paulo Aires

Trabalho 2: Questões sobre Aprendizado de Máquina

Setrem

Este trabalho tem o intuito de exercitar o que foi visto em aula durante o segundo fim de semana. Ele tem um valor de 60% na nota final, pode ser feito em dupla e deve ser enviado até o dia 06/04/2024 (23h59). Tanto dúvidas como a entrega do trabalho devem ser encaminhados para o e-mail do professor: jpaulo.aires@gmail.com

Questão 1

Considerando o conjunto de dados a seguir, responda às perguntas:

Febre	Idade	Mancha	Dor	Diagnóstico
sim	23	grande	sim	doente
não	9	pequena	não	saudável
sim	61	grande	não	saudável
sim	32	pequena	sim	doente
sim	21	grande	sim	saudável
não	48	grande	sim	doente
não	12	nenhuma	sim	saudável

a) Qual a frequência do atributo “Dor” = “não”?

b) Qual a moda para o atributo “Febre”?

c) Qual a média e a mediana do atributo “Idade”?

d) Converta o atributo “Mancha” em numérico usando o *one-hot encoding*.

Questão 2

Considerando as duas instâncias a seguir, responda às perguntas:

$x^{(1)} = [4\ 2\ 1\ 3\ 3]^T$

$x^{(2)} = [0\ 3\ 3\ 0\ 4]^T$

a) Calcule a distância euclidiana entre $x^{(1)}$ e $x^{(2)}$.

b) Calcule a distância do cosseno entre $x^{(1)}$ e $x^{(2)}$.

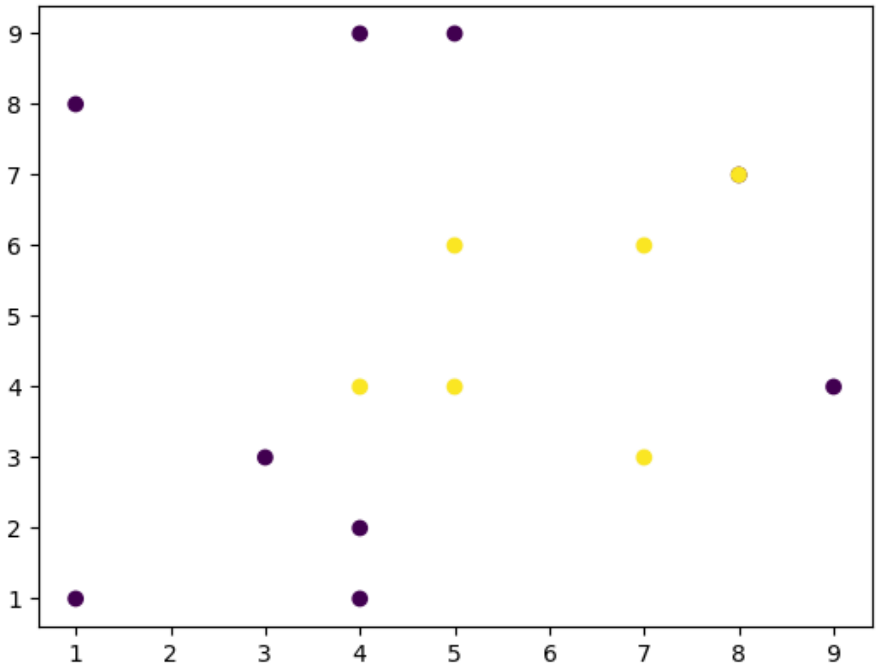
Questão 3

Considerando o algoritmo de aprendizado *k*-NN (*k*-Nearest Neighbors) e os dados a seguir, responda às questões utilizando a distância Euclidiana:

dados = [(4, 4, 1), (5, 9, 0), (4, 1, 0), (5, 6, 1), (7, 3, 1), (3, 3, 0), (7, 6, 1), (8, 7, 0), (8, 7, 1), (4, 2, 0), (1, 1, 0), (4, 9, 0), (5, 4, 1), (1, 8, 0), (9, 4, 0)]

Explicação dos dados (atributo_1, atributo_2, classe)

novo_dado = (9, 5, ?)



a) Qual é a classe do novo_dado quando *k* = 1?

b) Qual é a classe do novo_dado quando *k* = 3?

c) Considerando o contexto dos dados, qual opção de *k* é a mais adequada e por quê?

Questão 4

Considere o conjunto de atributos a seguir para uma tarefa de classificação de e-mails como *spam* ou não.

1. Frequência de palavras-chave relacionadas a spam no corpo do e-mail (alto (1)/baixo (0)).
2. Presença ou ausência de anexos no e-mail (sim(1)/não(0)).
3. Número de links externos no e-mail (> 3 (1), < 3 (0)).
4. Se o remetente é conhecido ou não (conhecido(1)/desconhecido(1)).

Freq	Anexos	Links	Rem.	Spam
1	0	1	1	-1 (Normal)
1	1	0	0	1 (Spam)
0	0	1	1	-1
0	1	1	1	-1
1	1	0	1	1
0	0	1	0	1

a) Treine um Perceptron com quatro entradas e uma única saída por uma época nos dados acima, começando com os pesos em 0. Informe os pesos finais para as quatro conexões e o bias, considerando um $\alpha = 0.4$

b) Com a sua rede treinada, classifique este novo e-mail:

Freq	Anexos	Links	Rem.	Spam
1	0	0	1	???

Questão 5

Dada estas três sentenças, represente-as usando Bag-of-Words. Lembre-se de normalizar o texto e remover as seguintes *stop-words*: ['o', 'a', 'e', 'eu', 'de', 'é', 'minha', 'para', 'que']. A resposta deve ser a definição do vocabulário e da representação vetorial de cada sentença.

1. "O cachorro late alto e eu gosto de gato."
2. "Eu gosto de pizza, pizza é minha comida favorita."
3. "O sol brilha forte durante o dia e a lua brilha para o cachorro que late."