nº: nome:

Aprendizagem Automática 2022

Exame - 9 de Janeiro 2022

1. Considere o seguinte conjunto de dados, com flores de duas espécies, onde cada exemplo é caracterizado por 3 atributos A1, A2, e A3, e o valor que se pretende prever é a espécie (A ou B). Se o conjunto for fornecido ao algoritmo de classificação **KNN K-vizinhos mais próximos**, com K=3 e usando a distância taxicab, ou Manhattan (equivalente à norma de Minkowski com p=1) como métrica de distância, qual o valor previsto para a espécie para o exemplo A1=1.0, A2=3.0, A3=0.2? Justifique a resposta apresentando os cálculos realizados(em caso de empate dê prioridade aos atributos pela seguinte ordem: A1;A2;A3)

ID	A1	A2	A3	Y
1	0.6	3.2	0.4	A
2	1.3	3.7	0.5	A
3	1.0	3.3	0.4	A
4	2.2	2.9	2.3	B
5	1.1	2.5	2.0	A
6	1.7	2.8	1.1	B
7	2.3	2.5	2.0	B
8	2.5	3.0	2.2	B
9	2.2	3.4	3.4	B

2. Sobre o algoritmo KNN de regressão explique de que modo o parâmetro K influencia os resultados do modelo, e como decidir o valor de K mais adequado.

3. Considere o seguinte conjunto de dados, onde cada exemplo é caracterizado por 3 atributos e pertence a uma de 2 classes. O conjunto de dados reflete decisões de um jogador de ténis quanto a ir ou não jogar sabendo se treinou na semana anterior, como se prevê que o tempo esteja, e se tem ou não uma lesão. Se o conjunto for fornecido ao algoritmo **naïve de Bayes**, com estimador de laplace (N_Casos_positivos + 1) / (N_total + k)),

ID	Treino	Tempo	Lesão	Jogar
1	S	Sol	S	Sim
2	N	Sol	S	Não
3	S	Sol	N	Sim
4	N	Sol	N	Sim
5	S	Nuvens	N	Sim
6	N	Nuvens	N	Sim
7	N	Chuva	N	Não
8	S	Chuva	S	Não
9	N	Chuva	S	Não

3.a) determine as expressões que permitem calcular a que classe pertence o exemplo {N, Nuvens, S }?

3.b) Indique como decidiria a classe em função do resultado das expressões calculadas.

Justifique as respostas apresentando os cálculos realizados, ou as expressões matemáticas.

4. Indique como se justifica a vantagem de usar o classificador suavizado de Laplace em vez do valor do rácio = *casos_considerados/num_casos_totais*

5. Considere o conjunto de dados da pergunta 3. Qual o atributo escolhido para a raíz da árvore de decisão quando é apresentado o conjunto anterior e a função de impureza é o erro de classificação? Justifique apresentando os cálculos, e/ou as expressões matemáticas.

6. Na definição duma árvore de decisão podem ser usados vários ìndices de impureza.

   6a. Indique justificando em que situações o índice Gini é mínimo.

   6b. Indique justificando em que situações o índice da entropia é máximo.

7. Calcule a média ponderada da precisão, para a seguinte matriz de confusão:

		True/Actual		
		Cat (🐱)	Fish (🐟)	Hen (🐔)
Predicted	Cat (🐱)	4	6	3
	Fish (🐟)	1	2	0
	Hen (🐔)	1	2	6

8. O gradiente descendente estocástico é um método iterativo para otimizar uma função objetivo com propriedades de suavidade, que é usado frequentemente na otimização de redes neuronais. Pode ser considerado como uma aproximação estocástica da otimização gradiente descendente, uma vez que substitui o gradiente real por uma estimativa do mesmo. Explique a diferença entre gradiente estocástico, e gradiente descendente, e considere com e sem *batches*.

9. O algoritmo de backpropagation permite otimizar os pesos de uma rede neuronal iterativamente. Imagine que se treina uma rede várias vezes com os mesmos critérios de terminação com pesos inicializados aleatoriamente. O modelo gerado é sempre o mesmo? Justifique a resposta.

10. Num algoritmo de comité existem alguns pressupostos sobre cada modelo individualmente. Indique os principais pressupostos que eventualmente permitem que o conjunto dos modelos tenha um desempenho superior a cada modelo individual.

11. O algoritmo *Random Forest* introduz aleatoriedade (daí o nome) no processo de construção das árvores que constituem o comité. Indique como é introduzida essa aleatoriedade.

12. O seguinte *heatmap* apresenta o desempenho do algoritmo Random Forest sobre um conjunto de dados para diferentes valores de dois parâmetros: profundidade das árvores (*max_depth*), e número de árvores(*n_estimators*). Considera que o intervalo de valores testado é adequado? (se responder afirmativamente justifique a sua resposta, se responder negativamente proponha gamas de valores alternativas para realizar a pesquisa)