

# Exercício Teoria - 4o bimestre - ECM514 Ciência de Dados

Nome: \_\_\_\_\_ RA: \_\_\_\_\_

Q1. (PCA) Considere as seguintes afirmativas sobre Análise de Componentes Principais (PCA):

- i. PCA projeta os dados em componentes ortogonais que maximizam a variância explicada
- ii. PCA é sensível à escala das variáveis; portanto a padronização (z-score) costuma ser recomendada
- iii. PCA pode ser aplicado a modelos de classificação, mas não em modelos de clusterização

Estão corretas:

- ☐ Somente i, ii
- ☐ Somente ii, iii
- ☐ Somente iii
- ☐ Somente i
- ☐ Todas as alternativas

Q2. (Cluster Hierárquico) Considere as seguintes afirmativas sobre métodos hierárquicos de agrupamento:

- i. O linkage single tende a produzir cadeias (clusters alongados) devido à ligação pelo par de ponto mais próximos
- ii. Distâncias entre observações devem ser métricas euclidianas para se aplicar linkage methods
- iii. Considere dois esquemas de clusterização, com 4 clusters, uma com o linkage ward e outra com o linkage complete. A métrica de silhueta pode ser empregada para escolha da melhor técnica, empregando o modelo de silhueta maior.

Estão corretas:

- ☐ Somente i, iii
- ☐ Somente ii, iii
- ☐ Somente iii
- ☐ Somente i
- ☐ Nenhuma das alternativas

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

Q3. (Cluster Hierárquico) Considere a matriz de distâncias acima, quais os 3 clusters criados empregando-se um linkage complete?

- ☐ (1, 3, 5), (2), (4)
- ☐ (1, 3), (2), (4, 5)
- ☐ (1), (4), (2,3,5)
- ☐ (1) (2,4) (3,5)
- ☐ Nenhuma das alternativas

Q4. (TF-IDF, Embedding) Considere as seguintes afirmativas sobre o esquema de representação vetorial de textos:

- i. Termos, como stopwords, tendem a ter um valor pequeno na posição de sua representação TF-IDF
- ii. Representação TF-IDF tendem a ser vetores bastante maiores (~1000-30000, da ordem do vocabulário de termos) que representações do tipo Word Embedding (300-500).
- iii. TF-IDF captura, através dos n-gramas, a semântica dos termos, do mesmo modo que o Word Embedding.

Estão corretas:

- ☐ Somente i, ii
- ☐ Somente ii, iii
- ☐ Somente iii
- ☐ Somente i
- ☐ Todas as alternativas

Q5. (Ganho de Informação) Considere o conjunto de dados abaixo. Quais atributos apresentam respectivamente maior Ganho de Informação e menor Entropia?

A	B	C	Class
1	R	Y	Yes
0	R	Y	No
1	B	Y	No
1	B	Y	Yes

- ☐ A e B
- ☐ A e C
- ☐ B e T
- ☐ B e C
- ☐ B e A

Q6. (Projetos) O que é o SHAP no contexto do Aprendizado de Máquina?

- ☐ Uma técnica de Clusterização
- ☐ Uma métrica de Agrupamentos
- ☐ Uma métrica de Classificação
- ☐ Um modelo de aprendizado não Supervisionado não implementado no sklearn mas por uma biblioteca p...
- ☐ Uma técnica que busca explicar/apresentar o quanto cada variável contribui para uma previsão do mode...