Reconhecimento de Padrões

Trabalho 5: Agrupamento e PCA

- Trabalho Individual

- Apenas simulações, sem trabalho escrito

-Enviar os códigos DEVIDAMENTE COMENTADOS, juntamente com a base de dados, para o email: alexandrefernandes@ufc.br

- Os códigos devem estar bem organizados e comentados, para que seja possível entendê-los e corrigi-los. Códigos que estejam desorganizados ou sem os devidos comentários explicativos terão penalização na nota.

- Não usar funções prontas para o K-means e PCA. É permitido usar uma função existente para o cálculo dos autovalores e autovetores no PCA.

- Prazo para entrega: 13/07 às 23:59

Prática: Agrupamento com PCA

- A base dados usada é denominada “Shop Customer Data”, que faz uma análise dos clientes de uma loja, ajudando a empresa a melhor entender seus clientes. A base de dados está detalhada em https://www.kaggle.com/datasets/datascientistanna/customers-dataset.

- Você deve numerizar os atributos *Gender* e *Profession*.

- Fazer agrupamento de dados usando K-means (sozinho) e usando PCA + K-means. Usar o K-means com a distância Euclidiana.

- Testar os seguintes valores de K: 2, 3, 4 e 5. No final, o algoritmo escolhe o caso que fornecer o melhor resultado em termos de Largura Média de Silhueta (SWC).

- Usar inicialização aleatória das sementes. Rodar o algoritmo 10 vezes e tirar uma média dos resultados.

- Como critério de convergência do *K-means*, pare o algoritmo quando as atribuições das classes não mudarem mais de uma iteração para a outra ou, de forma equivalente, quando os centroides não de alterarem de uma iteração para a outra.

- Você deve testar diferentes valores para o número de componentes do PCA e, no final, o algoritmo escolhe o caso que fornecer o melhor resultado em termos de SWC.

- O código deve fornecer como saída: número de componentes do PCA e a Largura Média de Silhueta (SWC) para ambos os casos. Se você desejar, pode usar a função existente para auxiliar no cálculo da SWC.

- Atenção, neste caso, não usar k-*fold* nem qualquer outra estratégia de partição de dados para treinamento e validação, pois não há treinamento (caso não supervisionado).