Universidade Federal de Alfenas - UNIFAL-MG - *campus* Varginha
Bacharelado Interdisciplinar em Ciência e Economia
Disciplina: Análise multivariada - Profa. Patrícia de Siqueira Ramos
Lista 4 - Análise de agrupamento

Seja o seguinte conjunto de dados ($n = 6$, $p = 2$):

Obs	X_1	X_2
0	1	4
1	1	3
2	0	4
3	5	1
4	6	2
5	4	0

Para auxiliar, a matriz de distâncias euclidianas entre as observações é:

$$\boldsymbol{D} = \begin{bmatrix} 0 & & & & \\ 1,00 & 0 & & & \\ 1,00 & 0,41 & 0 & & \\ 5,00 & 4,47 & 5,83 & 0 & \\ 5,39 & 5,10 & 6,32 & 1,41 & \\ 5,00 & 4,24 & 5,66 & 1,41 & 2,83 \end{bmatrix}$$

1 (manual) Agrupe os dados usando os métodos de agrupamento abaixo. Para cada letra mostre as matrizes de distâncias obtidas para todos os passos e os dendrogramas resultantes dos agrupamentos:
a) Método do vizinho mais próximo.
b) Método do vizinho mais distante.
c) Método da distância média.
d) Comente as principais diferenças identificadas entre os métodos.

2 (python) Faça o que foi pedido na questão 1, porém utilizando o `Python`. Além disso, também retorne os dendrogramas obtidos com os métodos:
a) centróide.
b) Ward.

3 (python) Aplique o método das k-médias no conjunto de dados:
a) Utilize o número de grupos k que você julgou melhor nas questões 1 e 2.
b) Obtenha o gráfico $SQDG\times$ número de grupos e verifique qual seria o número de grupos mais indicado para o método das k-médias.
c) Obtenha os *silhouette plots* para 2, 3 e 4 grupos. A conclusão é a mesma em relação a quantos grupos utilizar?

4 (python) No *notebook* da aula prática sobre agrupamento ('p8-aa.ipynb') complete as células marcadas como 'Questão 4.a)' etc.:
a) Nos dados do IMRS para MG, aplique o método da ligação média. Use a distância de Mahalanobis e também método da ligação média. O resultado foi muito diferente do método da ligação média usando a distância euclidiana?
b) *Silhouette plots* foram obtidos para diferentes valores de k para o método das k-médias. Como ficaria a interpretação desses gráficos? Qual o melhor número k?
c) Foram obtidos os agrupamentos dos municípios da mesorregião Sul/Sudoeste usando o método de

Ward e a distância euclidiana. Como ficaria a interpretação sobre os três grupos obtidos?
d) *Silhouette plots* foram obtidos para diferentes valores de k para o método das k-médias para os dados da mesorregião Sul/Sudoeste. Como ficaria a interpretação desses gráficos? Qual o melhor número k?