Aprendizado de Máquina e Reconhecimento de Padrões (UTFPR/CPGEI) - Lista de Exercícios 4

Tópicos: Clustering e Detecção de Novidades

1. Gere um conjunto de dados bidimensionais **X** com 400 exemplos. Esses exemplos formam quatro grupos igualmente distribuídos seguindo distribuições Gaussianas dadas pelos parâmetros:

$$m_1 = [0,\ 0]^T,\ m_2 = [10,\ 0],\ m_3 = [0,\ 9], \text{ and } m_4 = [9,\ 8]^T$$

$$S_1 = I, \quad S_2 = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1.5 \end{bmatrix}, \quad S_3 = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1.1 \end{bmatrix}, \quad S_4 = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.5 \end{bmatrix}$$

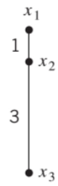
   a) Aplique o modelo k-means com m = 4 (número de centros). Utiliza a função “rand” do MATLAB para inicializar os parâmetros do k-means. Compare os centros obtidos pelo k-means com as médias das Gaussianas acima apresentadas. Plote os centros do k-means e os parâmetros **m** das Gaussianas e compare as diferenças.
   b) Repita o processo (a) para m = 3.
   c) Repita o processo (a) para m = 5.
   d) Repita o processo (a), agora inicializando o modelo k-means com $[-2.0, -2.0]^T$, $[-2.1, -2.1]^T$, $[-2.0, -2.2]^T$, $[-2.1, -2.2]^T$.
   e) Repita o processo (a), inicializando os três primeiros centros do k-means aleatoriamente (rand) e o quarto centro com $[20, 20]^T$.
   f) Repita o processo (a), (b), (c) e (d) utilizando o modelo Fuzzy C-Means e compare com os resultados obtidos para o k-Means.
   g) Comente os resultados.

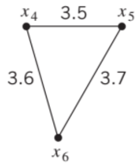
2. Gere um conjunto de dados bidimensionais **X** com clusters sem sobreposição e com diferentes formatos (distribuições). O primeiro cluster consiste de 600 pontos situados ao redor de um círculo centrado em (0, 0) e com raio igual a 6. O segundo cluster consiste de 200 pontos situados em torno de uma elipse centrada em (0, 0) com parâmetros $\text{raio}_{\text{maior}} = 3$ e $\text{raio}_{\text{menor}} = 1$. O terceiro cluster consiste de 200 pontos em um segmento de linha entre os pontos (8, -7) e (8, 7). O quarto cluster consiste em 100 pontos situados em torno de um semi-círculo centrado em (13, 0) com raio igual a 3 e ordenadas (eixo y) sendo inteiramente negativas. Aplique o modelo k-means nesse conjunto de dados a apresente conclusões.

3. Gere um conjunto de dados bidimensionais **X** com 400 pontos situados em torno de dois círculos concêntricos. Os dois primeiros 200 devem estar em torno de um círculo de raio 3 centrado em (0, 0). Os demais devem estar em torno de um círculo de raio 6 centrado em (1, 1). Aplique um algoritmo de clustering espectral com medida de similaridade Gaussiana considerando $\varepsilon=1.5$ e $\sigma=2$ e plote os resultados. Altere ε e σ e rode novamente. Qual o efeito de ε e σ nos resultados?

4. Considere o conjunto de dados {x_1, x_2, x_3, x_4, x_5, x_6}, sendo:

$$D = \begin{bmatrix} 0 & 1 & 4 & 20 & 22 & 23 \\ 1 & 0 & 3 & 22 & 24 & 25 \\ 4 & 3 & 0 & 23 & 25 & 26 \\ 20 & 22 & 23 & 0 & 3.5 & 3.6 \\ 22 & 24 & 25 & 3.5 & 0 & 3.6 \\ 23 & 25 & 26 & 3.6 & 3.7 & 0 \end{bmatrix}$$





A matriz que representa as distâncias d_{ij} entre pares de pontos entre os vetores $\mathbf{x}_i$ e $\mathbf{x}_j$, conforme ilustrado na figura ao lado da matriz. Aplique, passo-a-passo, o algoritmo single-link e avalie o resultado. Repita o processo usando o complete-link. Apresente ambos os dendogramas.

5. Gere um conjunto de dados bidimensionais **X** com 40 exemplos. Esses exemplos formam quatro grupos igualmente distribuídos seguindo distribuições Gaussianas dadas pelos parâmetros:

$$m_1 = [0,\ 0]^T,\ m_2 = [10,\ 0],\ m_3 = [0,\ 9],\ \text{and}\ m_4 = [9,\ 8]^T$$

$$S_1 = I, \quad S_2 = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1.5 \end{bmatrix}, \quad S_3 = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1.1 \end{bmatrix}, \quad S_4 = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.5 \end{bmatrix}$$

a) Aplique o algoritmo single-link e o complete-link e apresente os dendogramas resultantes. Determine os agrupamentos que melhor representam os dados de **X** e comente os resultados. Pesquise por métodos que fornecem uma métrica para justificar a escolha do melhor agrupamento determinado e apresente os resultados dessas métricas.

6. Gere um conjunto de dados bidimensionais **X** com 400 exemplos. Esses exemplos formam quatro grupos igualmente distribuídos seguindo distribuições Gaussianas dadas pelos parâmetros:

$$m_1 = [0,\ 0]^T,\ m_2 = [10,\ 0],\ m_3 = [0,\ 9],\ \text{and}\ m_4 = [9,\ 8]^T$$

$$S_1 = I, \quad S_2 = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 1.5 \end{bmatrix}, \quad S_3 = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1.1 \end{bmatrix}, \quad S_4 = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.5 \end{bmatrix}$$

a) Considere que esses clusters formam uma única classe. Avalie os métodos de detecção de novidades vistos em aula (utilizando a toolbox dd_tools do Matlab https://www.tudelft.nl/ewi/over-de-faculteit/afdelingen/intelligent-systems/pattern-recognition-bioinformatics/pattern-recognition-laboratory/data-and-software/dd-tools/) e verifique qual (ou quais) métodos melhor representam a superfície de decisão para esse conjunto de dados.
b) Repita o processo para os dados do exercício 2.

c) Varie os parâmetros para selecionar o melhor conjunto em todos os casos.

Obs.: essa toolbox não funciona corretamente no Octave.