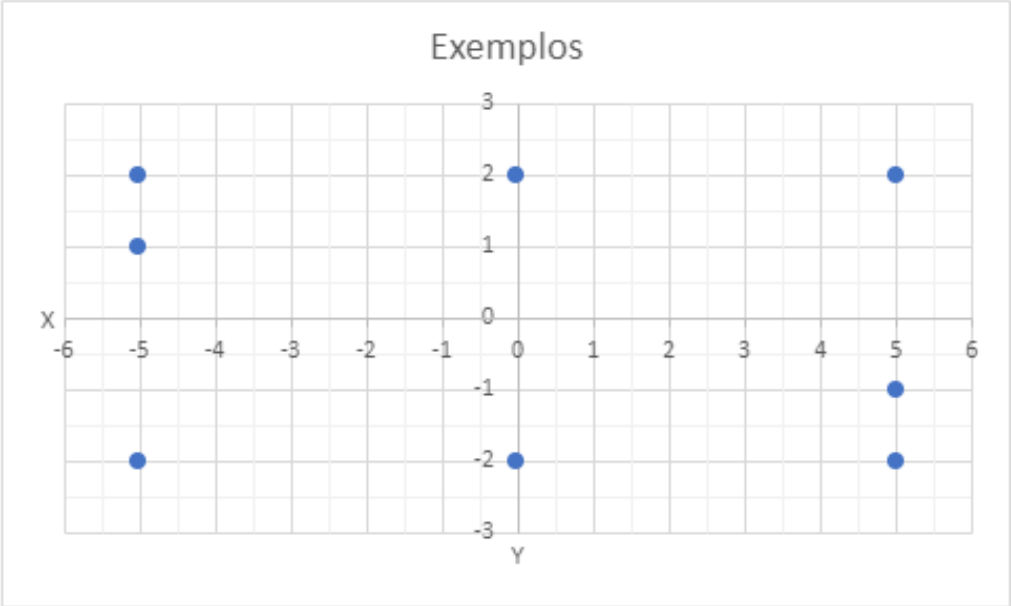
Lista 05 - IA

Luiza Ávila

01)

K=2

Centróides: (0,2);(0,-2)



Cálculo das distâncias:

D(Ex1,Ex3) = |-5-0| + |2-2| =5

D(Ex1,Ex4) = |-5-0| + |2-(-2)| =9

D(Ex2,Ex3) = |5-0| + |-2-2| =9

D(Ex2,Ex4) = |5-0| + |-2-(-2)| =5

D(Ex5,Ex3) = |-5-0| + |1-2| =6

D(Ex5,Ex4) = |-5-0| + |1-(-2)| =8

D(Ex6,Ex3) = |-5-0| + |-2-2| =9

D(Ex6,Ex4) = |-5-0| + |-2-(-2)| =5

D(Ex7,Ex3) = |5-0| + |2-2| =5

D(Ex7,Ex4) = |5-0| + |2-(-2)| =9

D(Ex8,Ex3) = |-5-0| + |(-1)-2| =8

D(Ex8,Ex4) = |-5-0| + |(-1)-(-2)| = 6

Agrupamentos: Ex3 --> Ex2, Ex6, Ex8

Ex4 --> Ex1, Ex5, Ex7

02)

B- Acontece erro pois a classe de dados é uma base que é classificada. Ele realizou o agrupamento perfeito, que não é ideal.

C- Considerando a matriz de confusão (existente só pois já existe classificação), o algoritmo errou 3 iris-versiculor e 14 iris-virginica. Ele também calcula que errou 11% das instâncias. Comparado com anteriormente, o algoritmo piorou suas classificações, mas ficou mais real com o que geralmente é.

03)

B- Centróides iniciais.

C- Média dos atributos em todos os grupos e no total. Ou seja, mostra o total e os centróides.

04) Tem problemas quando os clusters se diferem em tamanho, densidade, formas não esféricas e problemas com valores atípicos.

05)

A- Silhouette Index: avalia a coesão e a separação dos clusters. Baseia-se na diferença da diferença média dos pontos pertencentes ao cluster mais próximo para os pontos de um grupo. Para cada ponto da base de dados, Xi, é calculado o valor do silhouette index, Si , de acordo com a equação:

$$S_i = \frac{\mu_{out}^{min}(x_i) - \mu_{in}(x_i)}{max\{\mu_{out}^{min}(x_i), \mu_{in}(x_i)}$$

O sillhouette index é definido como o valor médio Si entre todos os pontos, dado pela equação:

$$SilhouetteIndex = \frac{1}{n}\sum_{i=e}^{n} S_i$$

Davies- Bouldin Index: feito utilizando quantidades e características inerentes ao conjunto de dados.

É formulado como a máxima razão entre a homogeneidade interna e a separação de clusters de acordo com o valor de K que minimiza DB(K).

$$DB=\frac{1}{N}\sum_{i=1}^{N} D_i \qquad D_i = \max_{j:i\neq j} R_{i,j} \qquad R_{i,j} = \frac{S_i + S_j}{M_{i,j}}$$

$$\propto \quad DB = \frac{1}{N}\sum_{i=1}^{N} \max_{i\neq j}\left(\frac{S_i + S_j}{d(c_i, c_j)}\right)$$

B – O grafo está na segunda métrica, apresentando então, uma estrutura razoável. Resultados foram os mesmos da questão anterior.

06) O problema de rotulação é estudado há muito tempo e visa escolher os melhor rótulo para cada cluster com base nas suas principais características, sendo as mais marcantes referentes aos centróides. Enquanto algoritmos de automatização ainda são estudados, a forma mais segura de rotular ainda á analisando os centróides de cada cluster.