Parte 1 (12 valores)

Nome ____

Número ____________________Curso ____________________

1. Seja $\mathcal{A} = \mathbb{R}^3$ o espaço de admissível para eventos duma base de dados. Considere a função dissemelhança entre 2 eventos x, $y \in \mathcal{A}$ definida da seguinte forma,

$$d(x,y) = \frac{1}{3}\sum_{i=1}^{3} d_i(x,y)$$

onde,

$$d_i(x,y) = \begin{cases} 1, & \text{se} \quad x_i \neq y_i, \\ 0, & \text{se} \quad x_i = y_i. \end{cases}$$

a) Mostre que é uma função de dissemelhança.

---

b) Verifique se $d(x,y)$ goza da propriedade de simetria.

---

c) Verifique se $d(x,y)$ goza da propriedade da identidade de indiscerníveis (”definitness”).

---

d) Uma função de dissemelhança como esta é apropriada para ser usada num algoritmo de clusterização como o de Lloyd para uma base de dados com eventos $e \in \mathcal{A}$? Justifique.

---

2. Considere o conjunto $C = \{x^1, x^2, x^3, x^4, x^5, x^6, x^7\}$ onde $x^1 = (0,0)^T$, $x^2 = (1,1)^T$, $x^3 = (0,1)^T$, $x^4 = (1,0)^T$, $x^5 = (70,0)^T$, $x^6 = (0,2)^T$, $x^7 = (2,0)^T$. Nota: os elementos $x^i \in \mathbb{R}^2$.

a) Calcule o representante de C (do tipo centróide), usando a métrica Euclidiana.

---

b) Calcule o representante de C (do tipo centróide), usando a métrica de Manhattan.

---

c) O que podemos inferir relativamente aos dados de C tendo em conta os resultados das alíneas anteriores?

---

3. Considere que tem uma base de dados composta por eventos que são figuras de 25 pixels binários. Um evento é um vector com 25 componentes binárias. Nessa base de dados há 3 eventos:
$x = (1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1,1,0,0,1,1,1,0,0)$,
$y = (1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,0,1,1,1)$
$z = (0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)$.
representados abaixo.

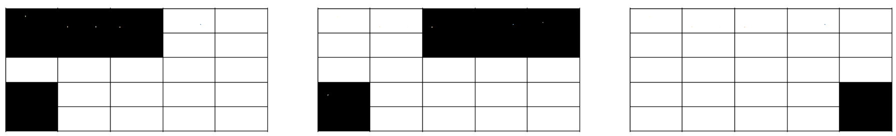

Figure 1: Evento x à esquerda; evento y no meio; evento z á direita

a) Calcule a semelhança entre x e y, entre x e z e entre y e z usando uma função de semelhança do tipo "overlap".

---

b) Calcule a semelhança entre x e y, entre x e z e entre y e z usando uma função de semelhança do tipo Jacard.

---

c) Atendendo aos resultados obtidos, qual das métricas lhe parece mais apropriada para medir a semelhança das figuras? Justifique.

---

4. Considere que tem uma base de dados $D = \{x^1, x^2, x^3, x^4, x^5\}$ onde $x^1 = (0,0)^T$, $x^2 = (0,1)^T$, $x^3 = (1,0)^T$, $x^4 = (10,10)^T$ e $x^5 = (8,10)^T$.
Vamos supor que se pretende aplicar o algoritmo de Lloyd com $k = 2$, usando a métrica euclidiana e representantes do tipo centroide. Para começar a correr o algoritmo considere que os representates iniciais são $\mathbb{M}(0) = \{(0,0)^T, (5,5)^T\}$. Nota: os elementos $x^i \in \mathbb{R}^2$.
Considere ainda que a condição de paragem é os representantes de duas iterações consecutivas serem os mesmos.
a) Apresente os calculos da primeira iteração do algoritmo de Lloyd.

---

b) Se os representantes iniciais fossem diferentes, obteria a mesma partição final após a convergência do algoritmo de LLoyd? Justifique?