Lista de exercícios #3

1. Levando-se em conta a matriz de distâncias entre 5 tuplas abaixo, esboçar o dendrograma correspondente para um método hierárquico aglomerativo em que a distância entre dois clusters é dada pela menor distância entre duas tuplas (uma de cada cluster). Apresente as matrizes intermediárias obtidas em cada passo do algoritmo.

$$D = \begin{bmatrix} 0 & & & & \\ 1 & 0 & & & \\ 5 & 4 & 0 & & \\ 8 & 7 & 3 & 0 & \\ 7 & 6 & 4 & 2 & 0 \end{bmatrix}$$

2. Considere que os seguintes vetores representam tuplas de um banco de dados: [1,1]; [1,2]; [2,1]; [2,2]; [5,1]; [6,1]; [5,2]. Simular a execução de uma iteração do algoritmo k-médias, para k=2, inicializando o algoritmo nos pontos [3,0] e [5,0]. Utilize a distância euclidiana ao quadrado e apresente os cálculos. Quais são os centróides obtidos?

3. Obtenha os *itemsets* frequentes usando o algoritmo *Apriori* na base de dados abaixo considerando suporte mínimo de 30%, ou seja, todo *itemset* que ocorre menos de 3 vezes na base de dados é considerado infrequente. Indique claramente o passo-a-passo do algoritmo. **Não** é necessária a construção das árvores hash para a contagem de suporte. Desenhe o reticulado (*lattice*) de *itemsets* e rotule cada vértice com as seguintes letras:

   - **M**: se o *itemset* é frequente maximal;
   - **C**: se o *itemset* é frequente fechado;
   - **N**: se o *itemset* não for considerado durante a execução do algoritmo, *i.e.*, se o *itemset* não for gerado ou for removido na etapa de poda por ter subconjunto infrequente;
   - **F**: se o *itemset* é frequente mas nem M nem C;
   - **I**: se o *itemset* é identificado como infrequente depois da contagem de suporte.

Transação	Itens
1	{a,b,d,e}
2	{b,c,d}
3	{a,b,d,e}
4	{a,c,d,e}
5	{b,c,d,e}
6	{b,d,e}
7	{c,d}
8	{a,b,c}
9	{a,d,e}
10	{b,d}

4. Considere os seguintes *itemsets* de tamanho 3:

$$\{1,2,3\}, \{1,2,6\}, \{1,3,4\}, \{2,3,4\}, \{2,4,5\}, \{3,4,6\}, \{4,5,6\}$$

a. Construa a árvore hash para estes candidatos. Como função hash utilize: $h(p) = p \mod 2$. Considere o algoritmo de construção como discutido em aula, *i.e.,* a cada nível i é checado o item na posição i. Considere como número máximo de *itemsets* em um nó folha como 2. Caso tenha mais que 2 *itemsets* no 3º nível da árvore não faça mais divisões, em outras palavras, **nesse nível** o número máximo de *itemsets* no nó folha deve ser ignorado.

b. Considere a transação $\{1, 2, 3, 5, 6\}$. Utilizando a árvore hash da questão (a), indique quais nós folha serão checados para esta transação?

5. Você recebeu um conjunto de dados com 500 objetos. Você executou o algoritmo k-médias variando o número de grupos para todos os valores de interesse ($k \in \{2, \ldots, 499\}$). Em todas as execuções apenas um grupo não-vazio foi encontrado. Explique a razão para isso. Explique qual seria o resultado obtido usando os algoritmos *single-linkage* e DBSCAN na mesma base de dados.

6. Indique qual matriz de **similaridade** está relacionada com qual base dados nas figuras abaixo. Note que cada matriz de similaridade está ordenada de acordo com rótulos de grupos obtidos por um algoritmo de agrupamento. Todas as bases de dados possuem 100 objetos e três grupos.

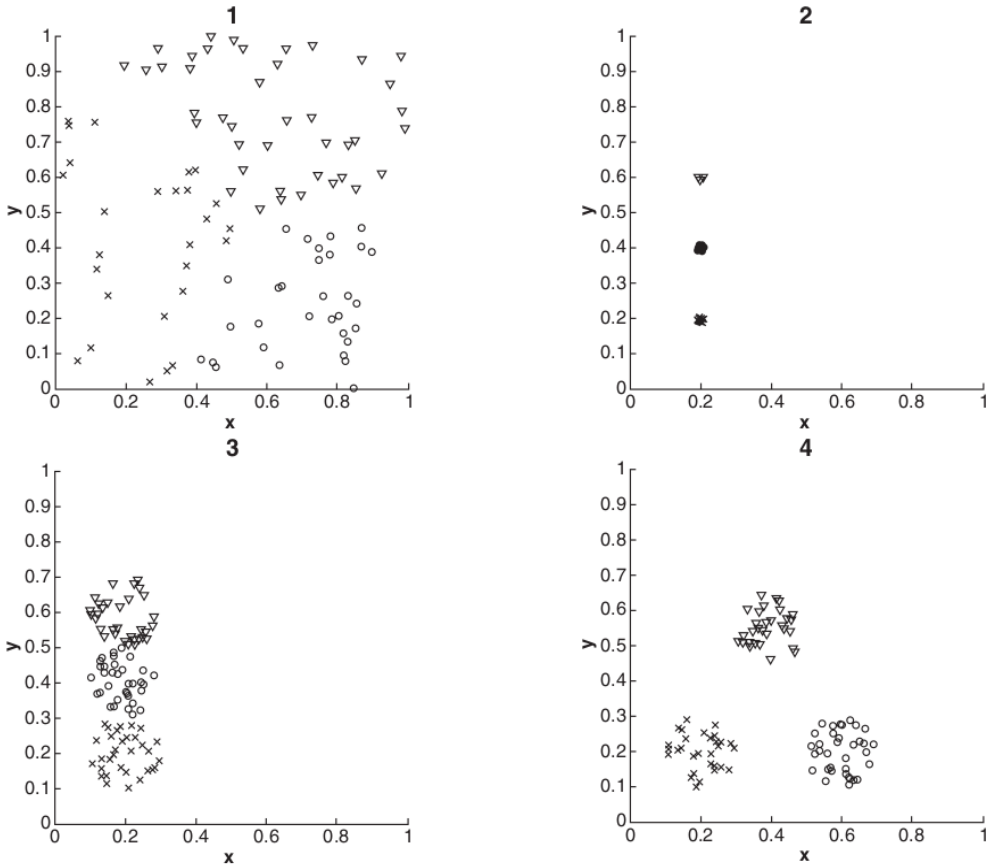

Figure 1: Exercício 6 - bases de dados

7. Descreva sob quais condições os grupos encontrados pelo algoritmo DBSCAN coincidem com os grupos encontrados pelo algoritmo k-médias.

8. Com o auxílio do computador, execute o algoritmo k-médias com $k = 9$, 100 vezes com diferentes inicializações na base de dados 9Gauss[1]. Considere o melhor valor de J obtido. Conforme visto em aula, é possível avaliar se o valor obtido é evidência de ter grupos na base de dados. Para isso, executa-se o algoritmo de agrupamento em bases de dados obtidas gerando-se pontos de forma uniforme no mesmo espaço. A partir dos valores de J obtidos nas bases artificiais é possível avaliar se o valor obtido na base de dados real era *improvável*. Apresente o código[2] para esta simulação, a distribuição de valores obtidos e a conclusão que você chegou com este experimento. Considere a geração de 100 bases de dados distintas.

[1] http://conteudo.icmc.usp.br/pessoas/campello/Sub_Pages/JH.htm

[2] k-médias implementado em R (https://stat.ethz.ch/R-manual/R-devel/library/stats/html/kmeans.html) e Python (http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html)

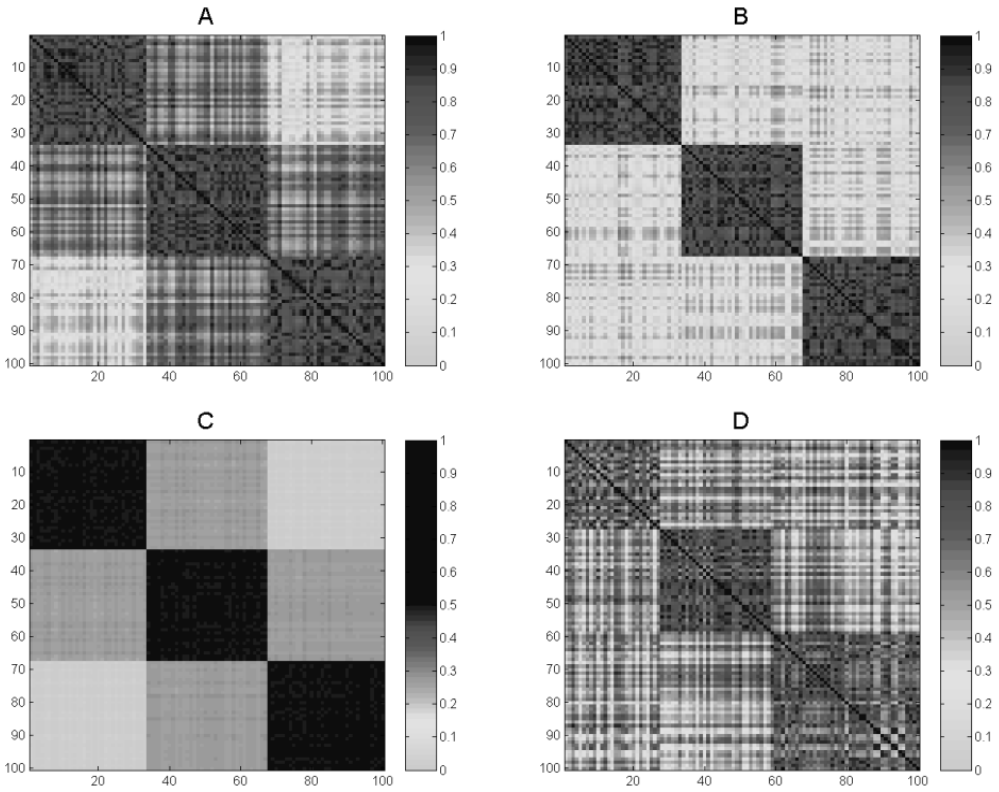

Figure 2: Exercício 6 - matrizes de similaridade