Lista de Exercícios #2

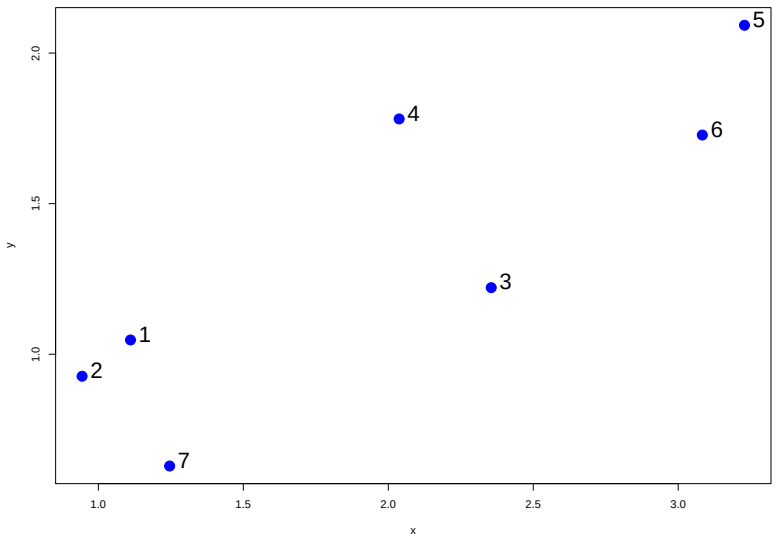
Mineração de Dados - QS 2020

Instruções

1. A lista deve ser feita individualmente.
2. Desconsiderando as questões de implementação, **todas** as respostas devem ser manuscritas. O material entregue deve consistir da cópia digitalizada da folha, por exemplo, uma foto **legível** da folha de soluções.
3. Os exercícios de implementação podem ser feitos na linguagem de programação que quiser, desde que para executar o seu código não seja necessária a instalação de **nenhum** software com custo maior que zero, isso inclui o Windows :).
4. Qualquer tentativa de fraude identificada será punida de acordo com o código de honra.
5. Além das respostas de cada pergunta no formato textual, você deve escolher um dos exercícios marcados com $\star$ e enviar o link para *download* de um vídeo descrevendo a solução do exercício. O vídeo deve ter no máximo 5 minutos e no início deve aparecer o rosto e carteirinha do aluno. A falha no recebimento do vídeo fará a lista ser desconsiderada.

Exercício 1 Execução manual

A Figura abaixo mostra uma coleção de objetos em um espaço de 2 dimensões (X, Y).



A matriz de distâncias (Euclidiana) entre os exemplos é mostrada abaixo:

	1	2	3	4	5	6	7
1	0.00	0.21	1.26	1.18	2.36	2.09	0.44
2	0.21	0.00	1.44	1.39	2.56	2.28	0.42
3	1.26	1.44	0.00	0.64	1.23	0.89	1.26
4	1.18	1.39	0.64	0.00	1.23	1.05	1.40
5	2.36	2.56	1.23	1.23	0.00	0.39	2.46
6	2.09	2.28	0.89	1.05	0.39	0.00	2.14
7	0.44	0.42	1.26	1.40	2.46	2.14	0.00

1. Aplique o procedimento de agrupamento hierárquico ao conjunto de exemplos e desenhe o dendograma correspondente. Use a ligação por distância mínima (**single-linkage**). Apresente o passo-a-passo do algoritmo, incluindo as matrizes de distâncias intermediárias (conforme feito no vídeo).

2. Aplique o procedimento de agrupamento hierárquico ao conjunto de exemplos e desenhe o dendograma correspondente. Use a ligação por distância máxima (**complete linkage**). Apresente o passo-a-passo do algoritmo, incluindo as matrizes de distâncias intermediárias (conforme feito no vídeo).

3. Execute o algoritmo k-medóides. Considere como medóides iniciais os objetos 3 e 4. Apresente o passo-a-passo do algoritmo, incluindo as matrizes de distâncias intermediárias (conforme feito no vídeo).

4. Execute o algoritmo k-médias. Considere como centróides iniciais os objetos 3 e 4. Os dados estão dispostos na tabela abaixo

	Atributo1	Atributo2
1	1.11	1.05
2	0.94	0.93
3	2.36	1.22
4	2.04	1.78
5	3.23	2.09
6	3.08	1.73
7	1.25	0.63

Exercício 2 k-Médias, *single-linkage* e DBSCAN

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.

Exercício 3 $\star$ Critério de validação

Implemente o algoritmo k-médias e execute-o com $k = 9$, 100 vezes com diferentes inicializações na base de dados 9Gauss. 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 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.

Exercício 4 Tendência de agrupamento

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. Justifique suas escolhas.

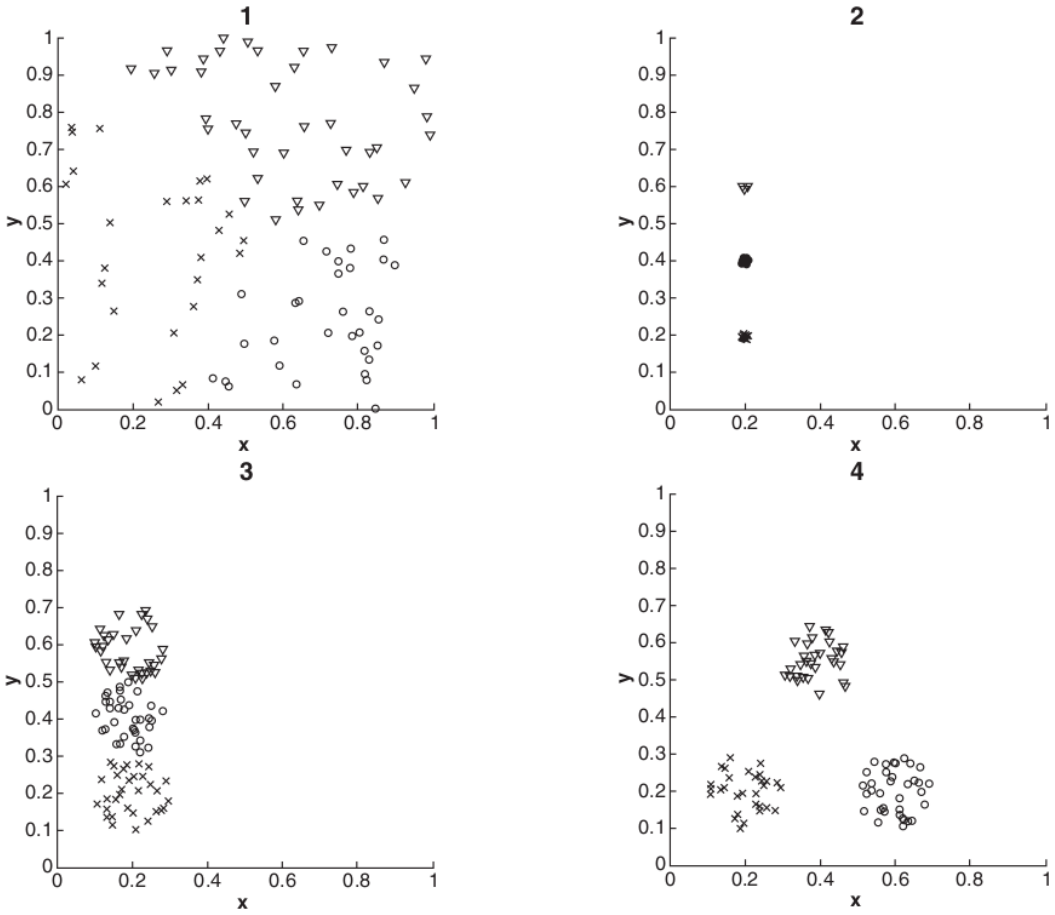


Figure 1: Exercício 6 - bases de dados

Exercício 5 Regras de associação

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.

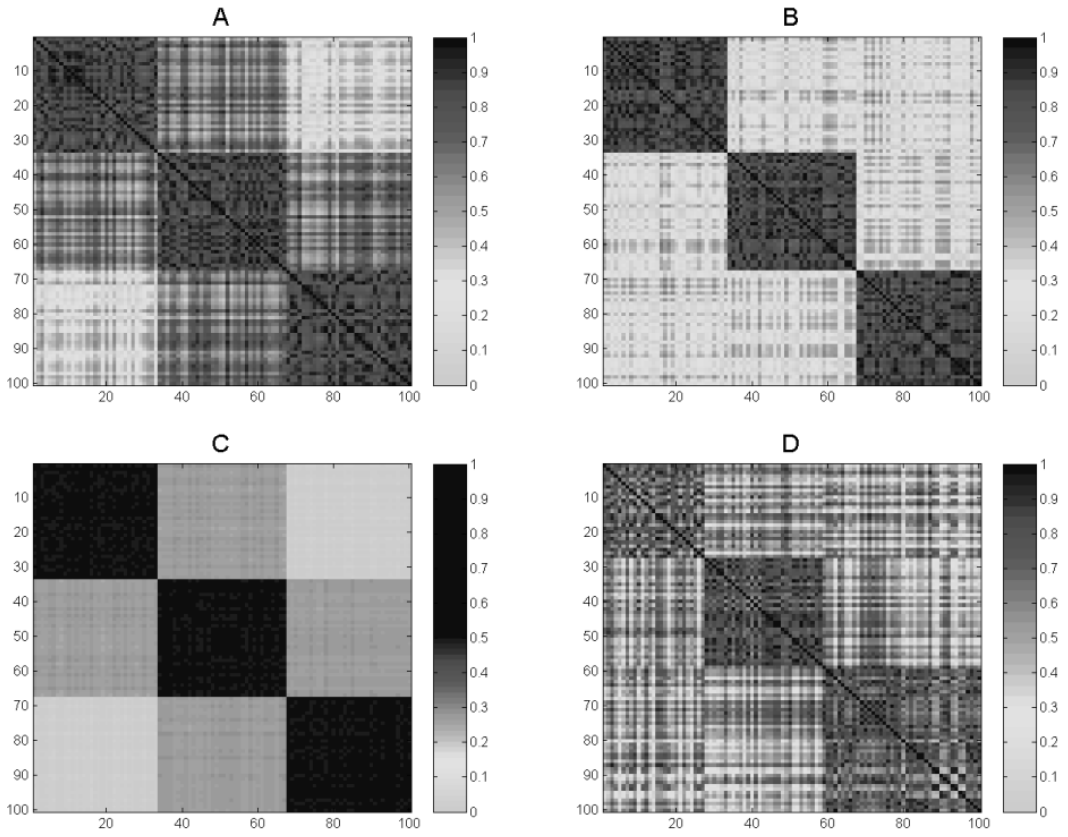


Figure 2: Exercício 6 - matrizes de similaridade

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