UNIVERSIDADE ESTADUAL DE CAMPINAS
Curso de Aperfeiçoamento – Fundamentos de Mineração de Dados
INF-1020 Fundamentos de Mineração de Dados
Prof. Julio Cesar dos Reis

LISTA - AGRUPAMENTO

Nome Aluno 1: Diego Coimbra

RG Aluno 1: 14358389-12

Nome Aluno 2: Nara Guimarães

RG Aluno 2: 44662603-x

1	0,75	
2	0,75	
3	0,75	
4	0,75	
5	0,75	
6	0,75	
7	0,75	
8	0,75	
9	2,0	
10	2,0	
Total	10,0	

Instrução: Esta lista de exercícios deve ser resolvida em dupla. A nota será no intervalo [0, 10].

Questã	Valor	Nota

1. Qual a diferença entre técnicas de classificação de dados e agrupamentos?

A diferença principal entre agrupamento (também chamado de clustering) e classificação é que o agrupamento é uma técnica de aprendizado de máquina não supervisionado que agrupa instâncias semelhantes com base em recursos, enquanto a classificação é uma técnica de aprendizado supervisionado que atribui tags predefinidas a instâncias com base em recursos.
De modo mais detalhado:

- Clustering é um método de agrupar objetos de forma que objetos com características semelhantes se juntem e objetos com características diferentes se separem.
- Classificação é um processo de categorização que usa um conjunto de dados de treinamento para reconhecer, diferenciar e compreender objetos. A classificação é uma técnica de aprendizagem supervisionada em que um conjunto de treinamento e

observações definidas corretamente estão disponíveis.

2. O que é um cluster? Por que análise de clusters pode ser relevante para aplicações de software modernas intensivas em dados?

Os clusters são agrupamentos de elementos de um mesmo grupo, de forma que elementos dentro de um mesmo cluster sejam muito parecidos, e os elementos em diferentes clusters sejam distintos entre si.

Permite encontrar nos dados uma estrutura de agrupamento natural, avaliando hipóteses acerca da estrutura de relações. Ajuda também a maximizar a homogeneidade de indivíduos dentro de grupos, e maximiza a heterogeneidade entre os grupos/conjuntos, de modo que, em muitos aspectos, podem ser considerados como um único sistema.

3. Explique os diferentes tipos de agrupamento.

- **Agrupamento Particional:** divisão dos dados em subconjuntos de forma que não exista sobreposição (grupos) tal que cada dado (objeto) está em exatamente um subconjunto.
- **Agrupamento Hierárquico:** é criada uma decomposição hierárquica dos dados, de forma que um conjunto de grupos aninhados são organizados em uma
- estrutura hierárquica chamada de 'árvore'.
- **Exclusivo vs não exclusivo:** no método exclusivo o dado pertence somente a um grupo, já no método não exclusivo o dado pode pertencer a vários grupos. Podem representar várias classes ou pontos de "fronteira".
- tem um grau de pertinência para cada grupo.
- **Fuzzy vs não fuzzy:** Agrupamento Fuzzy (difuso) é uma forma de agrupamento em que cada elemento pode pertencer a mais de um grupo (cluster). Já o agrupamento não-fuzzy (não difuso ou rígido) os dados são divididos em grupos distintos, onde cada ponto de dados só pode pertencer a exatamente um grupo.
- **Parcial vs Completo:** No agrupamento completo, todos os objetos são atribuídos necessariamente a um cluster, ao contrário do agrupamento parcial, onde os objetos não são necessariamente atribuídos a um cluster. Alguns objetos podem não ser bem definidos, e então, não pertencem a nenhum grupo.

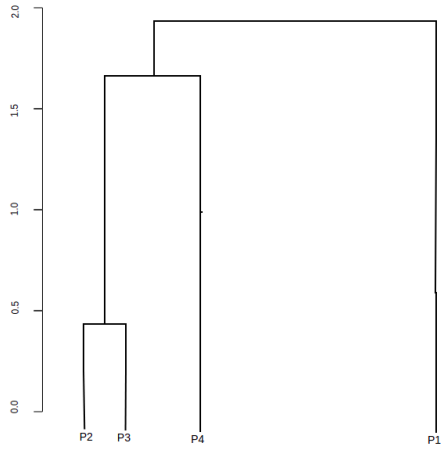
4. Qual a diferença entre algoritmos de agrupamento baseado em centróides e algoritmos de agrupamento baseado em conectividade.

- **baseado no centróide:** tem como ponto de partida descobrir o centro de cada cluster/agrupamento, e os novos elementos são classificados dependendo da menor

distância entre os centróides. A cada novo elemento, o centróide do conjunto é recalculado. Os números de clusters existentes são definidos desde o início

- **baseado na conectividade:** pontos e subgrupos próximos entre si devem pertencer ao mesmo supergrupo. Resultando assim em um conjunto de grupos aninhados organizados em uma árvore hierárquica

5. Apresente um dendograma para o agrupamento apresentado na Figura 1.



Dendograma desenhado em: https://online.visual-paradigm.com/drive/#diagramlist:proj=0&dashboard

6. Explique como funciona o algoritmo k-means. Apresente algumas limitações desta técnica de agrupamento.

É um algoritmo do tipo não supervisionado, ou seja, que não trabalha com dados rotulados, com o objetivo de encontrar similaridades entre os dados e agrupá-los conforme o número de cluster passado pelo argumento k, utilizando um método simples e eficiente baseado no conceito de distância.

O algoritmo de forma iterativa atribui os pontos de dados ao grupo que representa a menor distância, ou seja, ao grupo de dados que seja mais similar. O processo executado pelo K-Means é composto por quatro etapas:

1 - **Inicialização**: o algoritmo gera de forma aleatória k centroids, onde o número de centróides é representado ao parâmetro k. Estes centroids são pontos de dados que serão utilizados, como o nome sugere, de pontos centrais dos clusters.

2 - **Atribuição ao Cluster**: é calculado a distância entre todos os pontos de dados e cada um dos centróides. Cada registro será atribuído ao centroid ou cluster que tem a

menor distância, utilizado o método de distância Euclidiana. Esta etapa é finalizada com os dados divididos conforme o número de centróides estipulado pelo argumento k.

3 - **Movimentação de Centroids**: Assim que os pontos de dados são atribuídos aos clusters conforme sua distância, o próximo passo é recalcular o valor dos centróides, onde é calculada a média dos valores dos pontos de dados de cada cluster e o valor médio será o novo centróide. O termo movimentação se refere a alteração da localização do centróide em um plano se pensarmos em um gráfico.

4 - **Otimização do K-médias**: Nessa fase final da execução do K-means as fases Atribuição ao Cluster e Movimentação de Centroids são repetidas até o cluster se tornar estático ou algum critério de parada tenha sido atingido. O cluster se torna estático quando nenhum dos pontos de dados alteram de cluster. Um critério de parada pode ser o número de iterações máximas que o algoritmo irá fazer durante a fase de otimização. Por fim o K-means chega ao fim da sua execução dividindo os dados no número de clusters especificado pelo argumento k.

Limitações do método de k-means:

- precisa de separabilidade linear dos clusters
- precisa especificar o número de clusters
- Algoritmia: O procedimento de Loyds não converge para o verdadeiro máximo global, mesmo com uma boa inicialização quando há muitos pontos ou dimensões

7. Explique o problema da inicialização dos centróides no algoritmo k-means e as possíveis soluções.

Uma das principais fraquezas do k-means está na primeira etapa, com a escolha aleatória de k dos pontos de dados para serem os centróides iniciais.

O processo de inicialização do algoritmo de meios K padrão é totalmente aleatório. Por ser aleatório, às vezes os centróides iniciais são uma ótima escolha e levam a um agrupamento quase ideal. Outras vezes, os centróides iniciais são uma escolha razoável e levam a um bom agrupamento. Mas às vezes - novamente, porque eles são escolhidos aleatoriamente - às vezes os centróides iniciais são ruins, levando a um agrupamento não ideal.

Não é difícil adivinhar que isso cria problemas, especialmente em conjuntos de dados maiores e mais complexos que é executado por muito tempo e pode não atingir o clustering perfeito. É aí que entra k means ++. O algoritmo k-means ++ corrige esse defeito mudando a maneira como escolhemos os centróides iniciais. Tudo o mais sobre k-means permanece o mesmo.

Uma outra abordagem é analisar o gráfico de cotovelo para identificar o número ideal de centróides iniciais.

Outra alternativa é escolher k-pontos mais longes uns dos outros, porém isso faz com que a estratégia fique muito sensível aos outliers.

8. Explique as medidas para validação de clusters e a diferença entre coesão e separação.

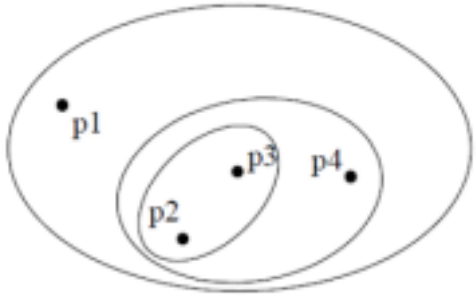


Figura 1: Clusters

Os algoritmos de validação de agrupamentos podem ser classificados em três grandes classes, sendo as duas principais baseadas em critérios externos e critérios internos, e a terceira baseada em um critério relativo.

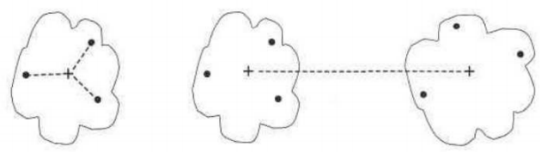
a) O critério ou índice externo, entre os clusters, avalia os resultados de um algoritmo de agrupamento baseado em uma estrutura pré-especificada, que nada mais é do que uma expectativa de resultados de agrupamentos para o conjunto de dados. Usado para medir a medida em que os rótulos de cluster correspondem a rótulos de classe fornecidos externamente.

b) O critério ou índice interno, dentro de cada cluster, avalia os agrupamentos com base nas próprias características dos objetos agrupados, como a matriz de similaridades e as listas de características dos objetos. Usado para medir o quão bom é uma estrutura de agrupamento (o quão separados e compactos estão os clusters) sem respeito à informação externa.

c) O critério ou índice relativo compara duas estruturas de agrupamentos de dados, fornecidas pelo mesmo algoritmo de agrupamento, porém com parâmetros diferentes.

Dentro do critério interno, tem-se duas métricas importantes:

- coesão: é a soma das similaridades dos objetos considerando o centróide de um agrupamento. A mesma valida a solidez dentro de um grupo, ou seja, o quão próximo os objetos estão dentro de um mesmo cluster.
- separação: é a proximidade, ou grau de afinidade, que centróides de diferentes agrupamentos apresentam. Validando o isolamento entre grupos e medindo o quão separado cada cluster está dos demais.



(a) Coesão (b) Separação

Em termos matemáticos:

- Coesão é medida pelo soma dos quadrados dentro do cluster (SSE)

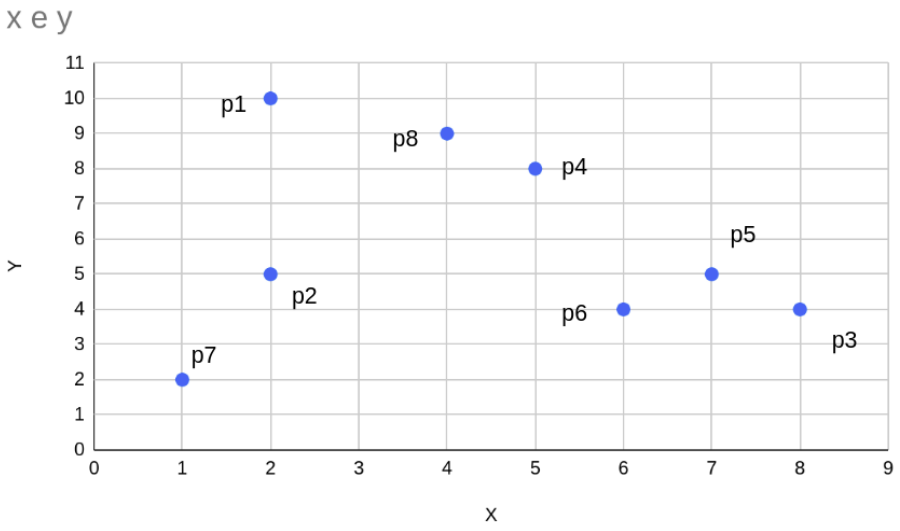
$$SSE = WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

- Separação é medida pela soma dos quadrados entre clusters

$$BSS = \sum_i |C_i| (m - m_i)^2$$

  - $|C_i|$ é o tamanho do cluster i

9. Considere o seguinte conjunto de pontos: p_1 = (2, 10), p_2 = (2, 5), p_3 = (8, 4), p_4 = (5, 8), p_5 = (7, 5), p_6 = (6, 4), p_7 = (1, 2) e p_8 = (4, 9). Suponha que os pontos p_1, p_4 e p_7 sejam usados como centróides iniciais do algoritmo de agrupamento K-means (k = 3). Usando a distância de Manhattan (i.e., dados pontos (x_1, y_1) e (x_2, y_2), *distância Manhattan* = $|x_1 - x_2| + |y_1 - y_2|$), calcule os três clusters para cada uma das primeiras iterações do algoritmo.



P.S.: serão recalculados os centróides

pontos	p1(2,10)	p4(5,8)	p7(1,2)
p2(2,5)	|(2-2)|+|(10-5)| = 0 + 5 = 5	|(5-2)|+|(8-5)| = 3 + 3 = 6	|(1-2)|+|(2-5)| = 1 + 3 = 4

Novo centróide p2,7 = (1.5, 3.5)

pontos	p1(2,10)	p4(5,8)	p2,7(1.5, 3.5)
p3(8,4)	|(2-8)|+|(10-4)| = 6 + 6 = 12	|(5-8)|+|(8-4)| = 3 + 4 = 7	|(1.5-8)|+|(3.5-4)| = 6.5 + 0.5 = 7

optamos por incluir o ponto 3 com o ponto 4. Novo centróide: (6.5, 6)

pontos	p1(2,10)	p3,4(6.5,6)	p2,7(1.5, 3.5)
p5(7,5)	|(2-7)|+|(10-5)| = 5 + 5 = 10	|(6.5-7)|+|(6-5)| = 0.5 + 1 = 1.5	|(1.5-7)|+|(3.5-5)| = 5.5 +1.5 = 9

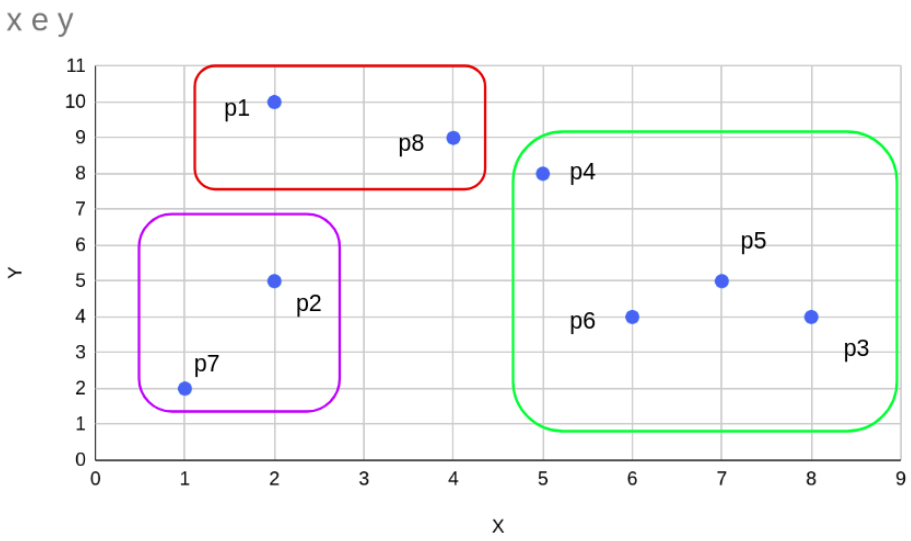
incluindo o ponto 5 junto dos pontos 3 e 4. Novo centróide: (7.25, 5)

pontos	p1(2,10)	p3,4,5(7.25,5)	p2,7(1.5, 3.5)
p6(6,4)	|(2-6)|+|(10-4)| = 4 + 6 = 10	|(7.25-6)|+|(5-4)| = 0.75 + 1 = 1.75	|(1.5-6)|+|(3.5-4)| = 4.5 + 0.5 = 5

incluindo o ponto 6 junto dos pontos 3, 4 e 5. Novo centróide: (7.625, 4.5)

pontos	p1(2,10)	p3,4,5,6 (7.625, 4.5)	p2,7(1.5, 3.5)
p8(4,9)	|(2-4)|+|(10-9)| = 2 +	|(7.625-4)|+|(4.5-9)|	|(1.5-4)|+|(3.5-9)| =

	1 = 3	= 3.625 + 4.5 = 8.125	2.3 + 5.5 = 8



10. Considere os clusters presentes na Figura 2. Calcule a medida de coesão para cada um dos clusters representados pelos centróides c_i e c_r. Para esse fim, use a distância de Manhattan (*i.e.*, dados pontos (x_1, y_1) e (x_2, y_2), *distância Manhattan* = $|x_1 - x_2| + |y_1 - y_2|$)

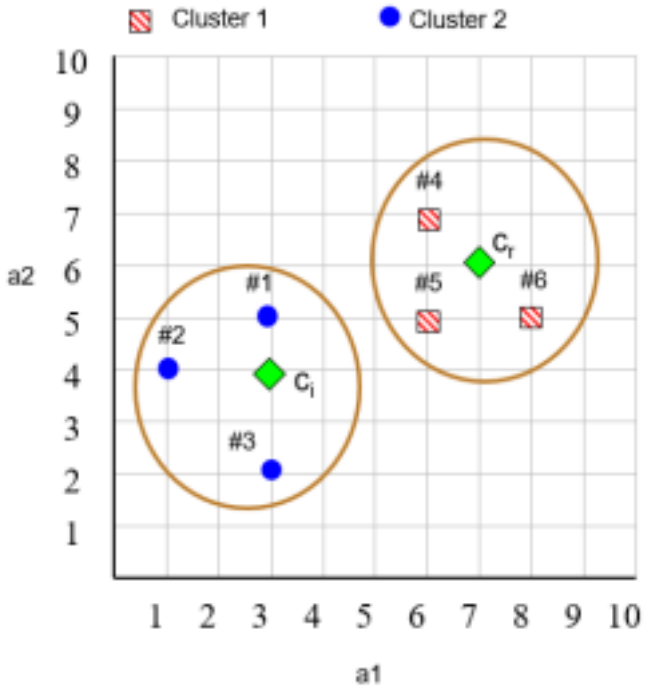


Figura 2: Clusters

Cálculo da coesão:

$$\text{Cluster SSE} = \sum_{\mathbf{x} \in C_i} dist(\mathbf{c}_i, \mathbf{x})^2 = \frac{1}{2m_i} \sum_{\mathbf{x} \in C_i} \sum_{\mathbf{y} \in C_i} dist(\mathbf{x}, \mathbf{y})^2$$

Ponto	a1	a2
1	3	5
2	1	4
3	3	2
ci	3	4

4	6	7
5	6	5
6	8	5
ci’	7	6

Para o cluster 1:

Pares de pontos	**Cálculo da distância**
ci-4	$(\|6-7\|+(\|7-6\|)^2 = (1+1)^2 = (2)^2 = 4$
ci-5	$(\|6-7\|+(\|5-6\|)^2 = (1+1)^2 = (2)^2 = 4$
ci-6	$(\|8-7\|+(\|5-6\|)^2 = (1+1)^2 = (2)^2 = 4$
SOMA	12

Para o cluster 2:

Pares de pontos	**Cálculo da distância**
ci-1	$(\|3-3\|+(\|5-4\|)^2 = (0+1)^2 = 1$
ci-2	$(\|1-3\|+(\|4-4\|)^2 = (2+0)^2 = 4$
ci-3	$(\|3-3\|+(\|2-4\|)^2 = (0+2)^2 = 4$
SOMA	9

O cluster 1 é mais coeso do que o cluster 2