Universidade Federal de São Carlos
Aprendizado de Máquina 1 – 01/2022
Prof. Murilo Naldi

Lista de exercícios 4 – Agrupamento

1. Explique detalhadamente a diferença entre classe (rótulo), usando para classificação, e grupo (*cluster*), resultado do processo de agrupamento.

2. Explique porque o agrupamento de dados pode ser útil em aplicações sobre conjuntos de dados reais. Porque?

3. O que é um agrupamento particional? O que é um agrupamento hierárquico? É possível obter uma partição a partir de uma hierarquia? E o contrário?

4. O que define um método como sendo relacional ou não relacional? Cite exemplos.

5. Agrupamento de dados é muito sensível e totalmente dependente da medida de similaridade utilizada. Tais medidas podem ser sensíveis à características dos dados, como sua dimensionalidade e a natureza de seus atributos. Explique como isso influencia diretamente no agrupamento.

6. Baseado nas respostas da pergunta anterior, como fazer adequadamente o agrupamento de textos/documentos? Busque referências.

7. Qual é a diferença entre o conceito clássico de grupo baseado em similaridade do agrupamento por densidade? Se o último também utiliza medidas de similaridade, ambos não deveriam ser equivalentes? Quais as diferenças?

8. O algoritmo k-médias otimiza a distância entre os objetos dos grupos e seus centroides, também conhecido como média da soma dos erros quadrados (do inglês *Mean Squared Error – MSE*) ou *J*. Para isso, ele necessita que o parâmetro *k* seja corretamente definido. Quais são as implicações de uma escolha incorreta? Como melhorar a escolha sem saber quantos grupos os dados possuem?

9. O resultado *k*-médias também é muito sensível a escolha dos protótipos (centroides) iniciais dos grupos. Quais as implicações que uma escolha ruim pode ter no resultado final do algoritmo? Explique como funciona esse processo. O que se pode fazer para tentar evitá-lo?

10. Os algoritmos hierárquicos podem montar uma hierarquia *top-down* (divisivos) ou botton-up (aglomerativos). Qual a diferença entre eles? Cite nomes de algoritmos dos dois tipos.

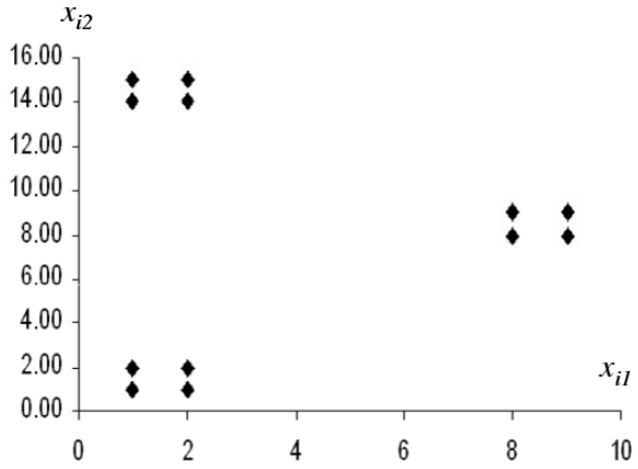
11. Algoritmos hierárquicos usam medidas clássicas para calcular a similaridade/dissimilaridade entre grupos. A escolha da medida influencia nos resultados obtidos. Explique as principais características do agrupamento feito com cada uma delas, vantagens e desvantagens.

12. Considerando os objetos de uma base de dados nós de um grafo completo conectado por arestas em que os pesos são definidos pela dissimilaridade entre os nós conectados, existe uma relação de equivalência entre aplicar o algoritmo hierárquico de ligação simples nesses dados e encontrar uma árvore geradora mínima desse grafo (Minimum Spanning Tree – MST). Explique porque.

13. O dendrograma resultante de um algoritmo de agrupamento hierárquico pode ajudar a entender a organização dos dados, em especial seus grupos. Explique como isso ocorre e porque é especialmente vantajoso para dados com mais de três atributos.

14. Considere o seguinte conjunto de dados abaixo:

Objeto $\mathbf{x}_i$	x_{i1}	x_{i2}
1	1	2
2	2	1
3	1	1
4	2	2
5	**8**	**9**
6	**9**	**8**
7	**9**	**9**
8	**8**	**8**
9	1	15
10	2	15
11	1	14
12	2	14



Neste exercício, utilize a distância de Manhattan (também conhecida como bloco cidade), cuja fórmula é dada a seguir:

$$d(\mathbf{x}_i, \mathbf{x}_j) = (\sum_{z=1}^{n_a} |\mathbf{x}_i(z) - \mathbf{x}_j(z)|)$$

em que $x_i(z)$ é o z-ésimo atributo do objeto x_i e n_a é o número de atributos do conjunto de dados.

Execute k-médias sobre os seguintes protótipos iniciais por até 5 iterações ou a convergência (o que ocorrer primeiro).

a) [6,6], [4,6], [5,10]
b) [2,3], [4,4], [2,5]
c) [4,2], [5,2], [3,18]
d) [5,2], [4,13], [10,10]

Em seguida, monte uma matriz de dissimilaridade entre todos os objetos e aplique o algoritmo hierárquico de ligação simples sobre estes dados. Com o resultado obtido, monte o dendrograma. Como ficaria a partição com três grupos desta hierarquia?

15. O que são algoritmos de agrupamento por densidade?

16. Considerando os dois parâmetros ε e m_{pts}, explique como funciona o algoritmo DBSCAN usando principais conceitos relativos ao agrupamento por densidade.

17. O conceito de externo (*outlier*) é natural para os algoritmos de agrupamento por densidade. Explique-o sobre a visão deste tipo de algoritmo. Porque os algoritmos particionais mais conhecidos não são capazes de lidar com esse tipo de dados? E os hierárquicos, como o algoritmo de ligação simples, porque não consegue bons resultados quando existe esse tipo de objeto no conjunto de dados?

18. Uma das limitações dos algoritmos particionais que utilizam densidade é a incapacidade de gerar soluções quando os dados possuem grupos com diferentes níveis de densidade bem separados ou aninhados (dentro de outros grupos). Porque isso acontece? Como superar essa limitação?

19. Explique quais são os passos do algoritmo HDBSCAN* (Dica: utilize teoria dos grafos).

20. O HDBSCAN* também é conhecido como algoritmo hierárquico de ligação simples robusto. Porque?

21. A representação gráfica mais natural de um algoritmo hierárquico é o dendrograma. Contudo, para grandes conjuntos de dados o dendrograma pode se tornar visualmente muito complexo e contraintuitivo. Uma opção mais compacta seria a árvore de grupos simplificada, que nada mais é do que a simplificação da hierarquia. Explique é o que é essa árvore e quais são os passos para se obter ela a partir de uma hierarquia de grupos.

22. Com uma hierarquia, temos diferentes partições aninhadas. No caso do algoritmos por densidade, cada nível da hierarquia representa uma partição com um nível de densidade mínimo. Sendo assim, diferentes níveis possuem diferentes densidades. Se quisermos extrair uma partição com um único nível de mínimo

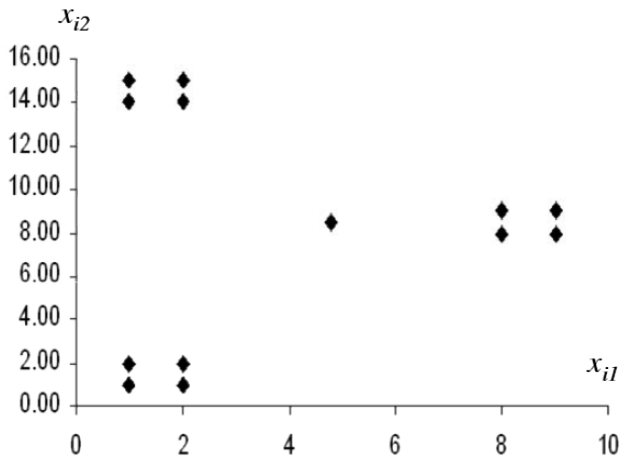
densidade, o que devemos fazer? E se quisermos considerar uma partição com grupos de diferentes níveis de densidade, como fazer?

23. Explique o conceito de estabilidade de grupo em uma hierarquia e como ele pode ser utilizado para avaliar grupos aninhados.

24. Em uma hierarquia de grupos temos a densidade mínima em que cada objeto se torna objeto de núcleo e deixa de ser externo (*outlier*). Como podemos usar essa informação para montar uma ordem (*rank*) de níveis de separabilidade (*outlierness*) utilizando GLOSH?

25. Considere o conjunto de dados da questão 14, adicionado de um ponto:

Objeto $\mathbf{x}_i$	x_{i1}	x_{i2}
1	1	2
2	2	1
3	1	1
4	2	2
5	**8**	**9**
6	**9**	**8**
7	**9**	**9**
8	**8**	**8**
9	1	15
10	2	15
11	1	14
12	2	14
13	4	**8**



Qual é a partição e externos dada pela aplicação do algoritmo DBSCAN com $\varepsilon = 2$ e $m_{pts} = 3$? Utilize a distância de Manhattan dada na questão 14.

Sabendo que a distância de núcleo $dcore_{mpts}(\mathbf{x})$ de um ponto **x** é a a distância de **x** até o seu seu m_{pts}-ésimo vizinho mais próximo, a distância de alcançabilidade mútua entre dois pontos **x** e **y** é dada por:

$$mrd_{mpts}(\mathbf{x},\mathbf{y}) = \max(dcore_{mpts}(\mathbf{x}), dcore_{mpts}(\mathbf{y}), d(\mathbf{x},\mathbf{y}))$$

onde $d(\mathbf{x},\mathbf{y})$ é a distância entre os pontos **x** e **y**. Neste exercício $d(\mathbf{x},\mathbf{y})$ é a distância Manhattan dada no exercício 14.

Usando $mrd_{mpts}(\mathbf{x},\mathbf{y})$, monte a matriz de alcançabilidade mútua entre todos os objetos do conjunto de dados e aplique o algoritmo hierárquico de ligação simples usando $mrd_{mpts}(\mathbf{x},\mathbf{y})$ ao invés da distância normal $d(\mathbf{x},\mathbf{y})$ (como você fez no exercício 14). O resultado será uma hierarquia de HDBSCAN*, que você poderá visualizando montando um dendrograma.

26. Quais as características de critérios de validação externos, internos e relativos? O critério J (do *k*-médias) é de que tipo(s)?

27. Aplique o critério silhueta nas partições obtidas nos exercícios 14 e 25 para determinar quais são as melhores partições. O critério para um objeto $\mathbf{x_i}$ é dado por:

$$silhouette(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{max\{a(\mathbf{x}_i), b(\mathbf{x}_i)\}}$$

em que $a(\mathbf{x_i})$ é a distância média de $\mathbf{x_i}$ para os outros objetos de seu grupo e $b(\mathbf{x_i})$ é a distância média de $\mathbf{x_i}$ para os objetos do grupo mais próximo que não é o seu grupo. A silhueta do grupo é a média das silhuetas dos objetos desse grupo e a silhueta da partição é a média das silhuetas de todos os objetos. Use distância de Manhattan.

28. Considerando os dados a seguir de forma que foram particionados por forma e também particionados por cor, calcule os valores dos índices de *Jaccard* e entre as duas partições (forma e cor).