2KDBSCAN

tempo limite de execução: 1s

Neste problema você precisará implementar um programa para uma versão adaptada do algoritmo DBSCAN (*Density-based spatial clustering*) que denominaremos de 2KDBSCAN.

Diferente do DBSCAN que precisa de dois parâmetros para responder quais são os grupos nos dados, o 2KDBSCAN precisará de apenas um parâmetro K. Para um determinado valor K, a versão deverá responder **quantos grupos existem**, qual o valor **máximo de** ε (*epsilon*) para que tais grupos possam existir e o **número de observações consideradas ruídos**. Vejamos os detalhes de implementação da versão 2KDBSCAN:

1 - O primeiro passo é obter um grafo completo a partir das observações da base de dados. Nesse grafo, cada observação é um vértice e para cada vértice haverá uma aresta ligando-o às demais observações. Cada aresta deverá ser ponderada pelo valor de uma métrica. Por padrão, considera-se a distância Euclidiana a métrica para essa versão:

$$w_{ij}(\mathbf{v}_i, \mathbf{v}_j) = \sqrt{\sum_{a=1}^{A}(v_{ia} - v_{ja})^2}, \tag{1}$$

onde w_{ij} é o peso da aresta que conceta os vértices $\mathbf{v}_i$ e $\mathbf{v_j}$ e v_{ia} e v_{ja} correspondem ao valor do atributo a para cada um dos vértices (observações).

2 - Em seguida, deverá ser obtida, a partir do grafo completo, uma árvore geradora mínima (*minimum spanning tree* - MST). Essa árvore, uma vez obtida, define um grupo que contém todas as observações da base de dados.

3 - O próximo passo é remover K arestas de maior peso da MST. Para cada K removido poderão ser obtidos:

- novos grupos e/ou;
- conforme a definição do DBSCAN, observações consideradas ruídos (*noises*).

4 - Para determinar o que é grupo e o que é ruído será preciso contar o número de componentes conexas do grafo. Dessa forma, se:

- uma componente conexa tiver mais de um vértice, então ela define um grupo;
- caso contrário, o vértice é definido como ruído.

5 - Para cada aresta removido o programa deverá imprimir o respectivo peso. Esse peso corresponde ao valor máximo de ε para que os possíveis grupos e/ou ruídos passem a existir. Todas as arestas de mesmo peso deverão ser removidas simultaneamente. Por último, também deverá ser apresentado o número de grupos e o número de ruídos após remover K arestas.

Entrada:

A entrada é formada por uma linha contendo 3 inteiros separados por espaço:

1 O primeiro inteiro $2 \leq N \leq 5000$ indica a quantidade de observações na base de dados.

2 O segundo $1 \leq A \leq 10$ indica a quantidade de atributos das observações.

3 O terceiro $1 \leq K \leq N - 1$ indica a quantidade de arestas que serão removidas.

As próximas N linhas correspondem às observações. Cada observação em uma linha apresenta R valores reais, $-100 \leq r_i \leq 100$, separados por espaços.

Saída:

Como saída o programa deverá imprimir o valor do peso da(s) aresta(s) removida(s) com a precisão de duas casas decimais após a vírgula; seguido do número de grupos e do número de ruídos. Se não houverem ruídos ou grupos formados, então deverá ser impresso o valor zero (0).

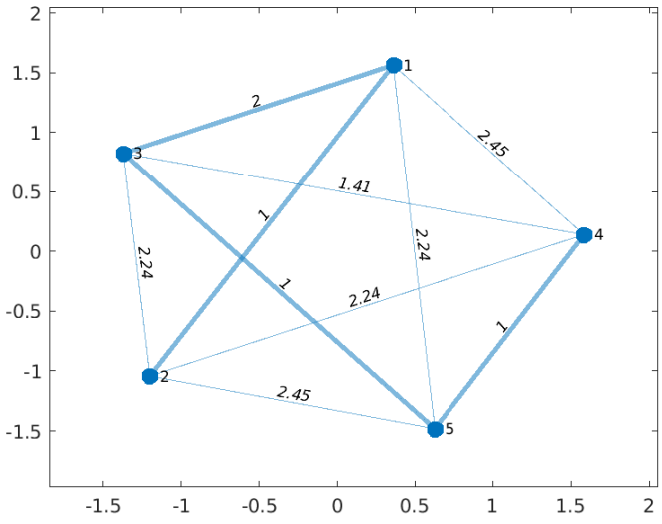
Exemplos:

Entrada:	Saída:
5 3 1 1 1 2 1 0 2 1 1 0 2 0 0 2 1 0	2.00 Numero de grupos = 2 Numero de ruidos = 0

Entrada:	Saída:
5 3 4 1 1 2 1 0 2 1 1 0 2 0 0 2 1 0	2.00 1.00 Numero de grupos = 0 Numero de ruidos = 5

Notas sobre o exemplo 1:

O grafo completo para o exemplo 1 é apresentado pela figura a seguir:



Na figura, os vértices estão numerados de 1 à 5 na forma em que foram lidos da entrada. A MST é apresentada pelas arestas destacadas.

Ao remover a aresta de maior peso da MST (aresta que liga o vértice 1 ao 3, cujo peso é 2) obtém-se duas componentes conexas, isto é, 2 grupos:

- um grupo formado pelas observações 3, 5 e 4;
- o outro grupo formado pelas observações 1 e 2.

O valor da aresta removida é 2.00. Esse é o valor máximo de ε para que os dois grupos passem a existir, isto é, um valor $\Delta > 0$ somado ao valor 2.00 resultaria em um grupo contendo todas as observações. Esse resultado pode ser obtido aplicando o DBSCAN para $minPts = 2$ e configurando ε com qualquer valor pertencente ao intervalo $(1.00, 2.00)$. O valor 2 no nome 2KDBSCAN se deve ao fato da versão se relacionar com o resultado do DBSCAN para $minPts = 2$.