COMP5318/COMP4318 Machine Learning and Data Mining

Week 10 Tutorial exercises
Clustering 2

Exercise 1. DBSCAN clustering

Use the DBSCAN algorithm to cluster the items A1, A2, …, A8. The distance matrix is given below. Assume that Eps=2 and MinPts=2.

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	5	6	3.6	7	7.2	8	2.2
A2		0	6.1	4.2	5	4.1	3.2	4.5
A3			0	5	1.5	1.5	7.5	6.5
A4				0	3.6	4.1	7.2	1.5
A5					0	1.4	6.7	5
A6						0	5.4	5.5
A7							0	7.5
A8								0

Solution:

1) Find the number of points in the neighborhood of each point based on Eps<=2:

N(A1)={A1} as the distance from A1 to all the other points is >2
N(A2)= {A2}
N(A3)= {A3, A5, A6}
N(A4)= {A4, A8}
N(A5)= {A5, A3, A6}
N(A6)= {A6, A3, A5}
N(A7)= {A7}
N(A8)= {A8, A4}

2) Label each example as core, border or noise using MinPts>=2

- A3, A4, A5, A6 and A8 are core as their neighborhood contains 2 or 3 points which is =>2
- A1, A2 and A7 are noise as their neighborhood contains only 1 point (themselves) which is <2 and they are not in the neighborhood of a core point

3) Find the clusters

-A3 is core, and its neighborhood contains A5 and A6 which are also core => A3, A5 and A6 form a cluster
-A4 is core and its neighborhood contains A8 which is core => A4 and A8 form a cluster
Final clustering: K1={A3, A5, A6}, K2={A4, A8}

***Exercise 2.* Evaluating clustering quality using the silhouette coefficient**

Given are 4 items P1, P2, P3 and P4. They were clustered using a clustering algorithm. The cluster labels and the distance matrix are shown below. Evaluate the quality of the clustering by computing the silhouette coefficient for each point, each of the 2 clusters and the overall clustering.

Distance matrix:

	P1	P2	P3	P4
P1	0	0.1	0.65	0.55
P2		0	0.7	0.6
P3			0	0.3
P4				0

Cluster labels:

point	cluster label
P1	1
P2	1
P3	2
P4	2

Solution:

The algorithm has found 2 clusters K1={P1, P2} and K2={P3, P4}. The silhouette coefficient measures both cluster cohesion and separation.

a_i = the average distance from example *i* to all examples in the same cluster
b_i = the minimum of the average distance of *i* to all examples in other clusters
Silhouette coefficient $s_i=(b_i-a_1)/\max(a_i,b_i)$

For each point:

For P1: $a_1=0.1$, $b_1=(0.65+0.55)/2$, $s_1=0.833$
For P2: $a_2=0.1$, $b_2=(0.7+0.6)/2$, $s_2=0.846$
For P3: $a_3=0.3$, $b_3=(0.65+0.7)/2$, $s_3=0.556$
For P4: $a_4=0.3$, $b_4=(0.55+0.6)/2$, $s_4=0.478$

For each cluster:

For cluster K1, the averaged silhouette coefficient $s_1=(0.833+0.846)/2=0.84$
For K2, $s_2=(0.556+0.478)/2=0.52$

Overall for the clustering:

s=(0.84+0.52)/2=0.68, relatively good (positive and close to 1)

***Exercise 3.* Evaluating clustering quality using correlation**

For the data from the previous exercise, evaluate the clustering quality using the correlation between the similarity matrix derived from the distance matrix (given below) and the similarity matrix derived from the clustering results (i.e. the matrix whose ij entry is 1 if two objects belong to the same cluster and 0 otherwise).

The similarity matrix derived from the distance matrix is given below. It was computed from the distance matrix as *s =1 - (d - dmin)/(dmax - dmin),* where *dmin* and *dmax* are the minimum and maximum distances in the matrix: *dmin*=0.1 and *dmax*=0.7).

Similarity matrix:

	P1	P2	P3	P4
P1	1	1	0.08	0.25
P2		1	0	0.17
P3			1	0.67
P4				1

Solution:
Given that there are two clusters K1={P1, P2} and K2={P3, P4}, the ideal similarity matrix is:

	P1	P2	P3	P4
P1	1	1	0	0
P2		1	0	0
P3			1	1
P4				1

Thus, we need to compute the correlation between the vectors
x=[1, 0.08, 0.25, 0, 0.17, 0.67] and
y=[1,0,0,0,0,1]

Recall that:

$$\text{corr}(\mathbf{x},\mathbf{y}) = \frac{\text{covar}(\mathbf{x},\mathbf{y})}{\text{std}(\mathbf{x})\text{std}(\mathbf{y})} \qquad std(\mathbf{x}) = \sqrt{\frac{\sum_{k=1}^{n}(x_k - mean(\mathbf{x}))^2}{n-1}} \qquad mean(\mathbf{x}) = \frac{\sum_{k=1}^{n} x_k}{n}$$

$$\text{co var}(\mathbf{x},\mathbf{y}) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - mean(x))(y_k - mean(y))$$

mean(x)=0.36, std(x)=0.39
mean(y)=0.33, std(y)=0.52
covar(x,y)=0.16

covar(x,y)=[(1-0.36)(1-0.33)+(0.08-0.36)(-0.33)+(0.25-0.36)(-0.33)+(-0.36)(-0.33)+(0.17-0.36)(-0.33)+(0.67-0.36)(1-0.33)] / (6-1) = 0.189

=>corr(x,y)=0.189/(0.39*0.52)=0.93 => high correlation (close to 1) => items that are close to each other are in the same cluster => good clustering