CSE 881: Data Mining (Fall 2016) Homework 9

Due date: Nov 22, 2016

1. Use the distance matrix shown in the table below to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged and the y-axis show the distance between pairs of clusters being merged at each iteration.

	p1	p2	p3	p4	p5
p1	0	0.5840	0.1955	0.3815	0.1127
p2	0.5840	0	0.6132	0.4956	0.5733
p3	0.1955	0.6132	0	0.2390	0.3067
p4	0.3815	0.4956	0.2390	0	0.4694
p5	0.1127	0.5733	0.3067	0.4694	0

2. Consider the data set shown in Figure 1. Suppose we apply DBScan algorithm with Eps = 0.15 (in Euclidean distance) and MinPts = 3.

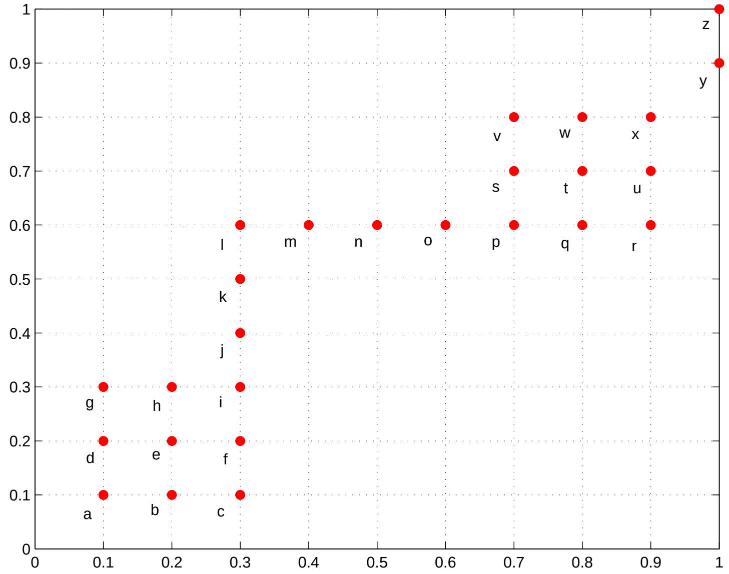

Figure 1: Data set for DBScan clustering.

(a) List all the core points in the diagram (you can use the labels of the data points in the diagram). Note: a point is considered a core point if there are **more than MinPts** number of points (including the point itself) within a neighborhood of radius Eps.

(b) List all the border points in the diagram.

(c) List all the noise points in the diagram.

(d) Using the DBScan algorithm, how many clusters will be obtained from the data set?

3. Consider the graph data shown in Figure 2. Assume the weights for all the links are equal to 1.

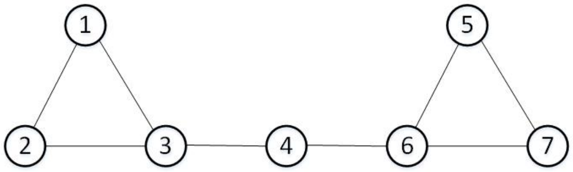

Figure 2: Graph data

(a) Compute the Laplacian matrix for the graph. Use the node indices shown in Figure 2 to order the rows and columns of the matrix.

(b) Compute the first three smallest eigenvalues of the graph Laplacian matrix.

(c) Compute the eigenvectors that correspond to the three smallest eigenvalues given in part (b).

(d) Apply k-means on the eigenvector matrix to generate 3 clusters. List the three clusters found.

(e) Calculate the normalized cut obtained for the 3 clusters found. Let V denote the set of all the nodes in a graph and $W = [w_{ij}]$ denote its adjacency matrix. Suppose V is partitioned into 3 disjoint subsets, V_1, V_2, and V_3, where $V_1 \cup V_2 \cup V_3 = V$. The normalized cut for the partitions can be computed as follows:

$$\text{Ncut}(V_1, V_2, V_3) = \sum_{i=1}^{3} \frac{\text{Cut}(V_i, V - V_i)}{d(V_i)} \tag{1}$$

where

$$\begin{aligned} d(V_i) &= \sum_{k \in V_i, j \in V} w_{ij}, \\ \text{Cut}(A, B) &= \sum_{i \in A, j \in B} w_{ij} \end{aligned} \tag{2}$$

(f) Suppose the 3 clusters found are as follows:

$$(1, 2), \quad (3, 4, 6), \quad (5, 7)$$

Compute the normalized cut of the clusters. Is the normalized cut smaller, larger, or equal to the solution found in part (d)?