Rhodes College

Name:_____

Topics in Computer Science: Data Mining **Assignment 5**

COMP 345 – CRN 19746/19839 Fall 2018

Clustering **Due Wed. Oct. 24th/Thurs. Oct. 25th (at the beginning of class)**

1. Read Chapter 10 in Han Book
2. Give the definition of a core point in DBSCAN.

3. Give the definition of a centroid in k-means.

4. The following is a set of one-dimensional points: {1; 1; 2; 3; 5; 8; 13; 21; 33; 54}. Perform two iterations of k-means on these points using the two initial centroids 0 and 11.

5. Assume that you have to explore a large data set of high dimensionality. You know nothing about the distribution of the data. How can k-means and DBSCAN be used to find the number of clusters in that data?

6.

	A_1	A_2
x_1	2	10
x_2	2	5
x_3	8	4
x_4	5	8
x_5	7	5
x_6	6	4
x_7	1	2
x_8	4	9

a. Suppose you want to cluster the eight points shown above using k-means. Assume that k = 3 and that initially the points are assigned to clusters as follows: C1 = {x1, x2, x3}, C2 = {x4, x5, x6}, C3 = {x7, x8}. What are the initial values for each centroid?

b. Consider the set of points given in Figure 1. Assume that eps = $\sqrt{2}$ and minpts = 3 (including the center point). Using Euclidian Distance find all the density-based clusters in the figure using the DBSCAN algorithm. List the final clusters (with the points in lexicographic order, i.e., from A to J) and outliers.

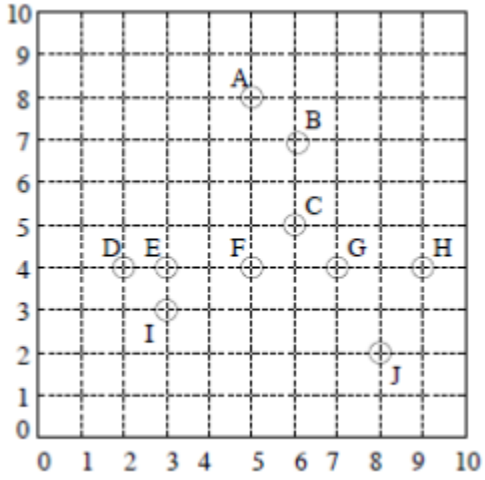
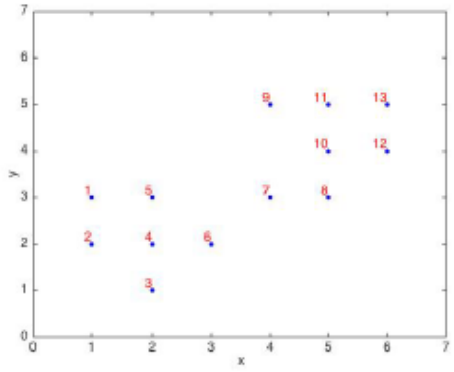
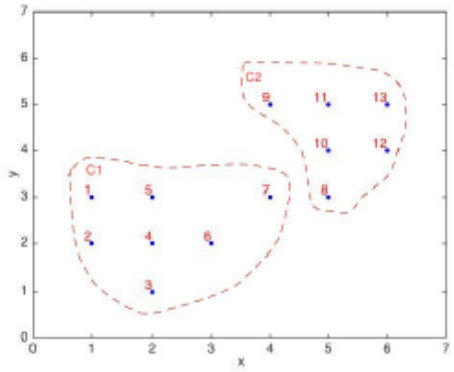


Figure 1

7. Suppose we have 13 data points, which are listed and plotted in the figure below. The ground truth (the correct cluster output) is also provided. Finish filling out the dissimilarity matrix below using Manhattan distance as your metric.

Point	x	y	Cluster
1	1	3	C1
2	1	2	C1
3	2	1	C1
4	2	2	C1
5	2	3	C1
6	3	2	C1
7	4	3	C1
8	5	3	C2
9	4	5	C2
10	5	4	C2
11	5	5	C2
12	6	4	C2
13	6	5	C2





	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
1	0	1	3	2	1	3	3	4	5	5	6	6	7
2		0	2	1	2	2	4	5	6	6	7	7	8
3			0	1	2	2	4	5	6	6	7	7	8
4				0	1	1	3	4	5	5	6	6	7
5					0	2	2	3	4	4	5	5	6
6						0	2	3	4	4	5	5	6
7							0	1	2	2	3	3	4
8								0					
9									0				
10										0			
11											0		
12												0	
13													0

a. Perform AGNES, a hierarchical clustering algorithm on the points above. Please use single link method and adopt Manhattan distance as the dissimilarity measure. You need to list all clusters in each layer of the hierarchy.

b. If we don't know the ground truth, and want to cluster the data set into 2 groups, based on the result above, what are the members of the 2 groups?

c. Based on the given ground truth, what are the Precision and Recall of the output (call class 1 positive)?