DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Assignment 2

Data preprocessing and clustering

ENGG 5103: Techniques for data mining Deadline: 5:00 pm, Nov 22th, 2016

Submission: On CUHK E-learning

Name: _____

SID: _____

1. (**Data preprocessing, 10%**) Following tale records the information of some students.

Student id	**Name**	**Gender**	**Current GPA**	**CS Student**	**Course Enrolled**
1034	Jerry	Male	3.2	Yes	Data Mining
1034	Jerry	Male	3.2	Yes	Physics
2578	Tom	Male	3.8	No	Data Mining
2578	Tom	Male	3.8	No	Linear Algebra
	Lion		2.9		
3729	Spike	Male	2.0	No	Linear Algebra
4280	Lily	Female	4.0	No	Physics
5530	Lucy	Female	3.7	Yes	Data Mining

a. Please clean the records with missing values in at least one column.

b. The teacher wants to see the students' records without names. Please do the coding process using feature selection.

c. Please perform data transformation so that the variable in the column "CS student" is 0-1 binary one and the variable in "Course enrolled" takes value from {1,2,3}.

d. Draw a histogram to show the distribution of students' GPA after processing steps (a) (b) (c).

2. (**K-means clustering, 18%**) Given the following points, finish the questions.

x	y
1.5	0.8
3.3	2.5
4.8	3.8
4	1.2
3.2	5
2.5	4.6
0.5	3
0.6	2.1
2.8	2.3
4.2	4.4
3.8	3.9
4	2.4
0.8	0.4
1.7	5
2.1	2.9
3.5	1.7
2.5	0.8
0.6	2.8
4.2	0.3
3.1	0.1

a. Suppose the k-means clustering (c = 2) is initialized by (1.8, 2.3) and (2.3, 1.4). Plot the objective function value in each iteration (until convergence),

$$\sum_{i=1}^{2} \sum_{(x,y)\in cluster\ i} dist((x,y), center\ of\ cluster\ i)$$

and show the clustering result, with Euclidean distance as distance metric.

b. Suppose the k-means clustering (c = 2) is initialized by (1.8, 2.3) and (2.3, 1.4). Plot the objective function value in each iteration and show the clustering, with Cityblock distance as distance metric

c. Suppose the k-means clustering (c = 2) is initialized by (1.8, 2.3) and (2.3, 1.4). Plot the objective function value in each iteration and show the clustering result using Minkowski distance with lambda = 3.

d. Answer the following questions based on your results:

[1]. Is the clustering result the same given the same initialization center and different distance metric? If not, describe the difference.

[2]. List 3 real world clustering examples, for which Euclidean distance, Cityblock distance, and Minkowski distance (lambda = 3) are most suitable distance metric respectively. Give your reasons.

3. (**Hierarchical Clustering, 18%**) Given the points in question 2, answer the following questions. The distance is set to be Euclidean distance.

a. Calculate initial proximity matrix between points, and make an initial separation of them into 5 clusters.

b. Based on the 5 clusters, perform hierarchical clustering using complete linkage. Show the proximity matrices for each step.

c. Based on the 5 clusters, perform hierarchical clustering using single linkage. Show the proximity matrices for each step.

d. Based on the 5 clusters, perform hierarchical clustering using group average. Show the proximity matrices for each step.

e. Answer the following questions based on your results:

[1]. How could the difference proximity matrices in the process lead to their own results?

[2]. List 3 real world examples, for which complete linkage, group averages, and single linkage are most suitable respectively. Give your reasons.

4. (**SOM clustering, 18%**) You are required to perform SOM on points in question 2. The sampling sequence is from top point to the bottom one.

a. Given initial guess centroid (1,3.1), (2,2.2), (1.5,2.1), (3.1,1.1). Write the topological order of them in memory space.

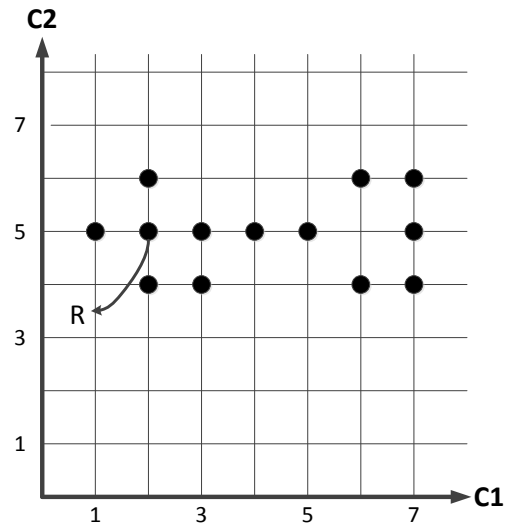
b. Choose a=0.3 and a(neighbor)=0.2, size of neighbor being 3, and perform SOM clustering and show the final results.

c. For the kth iteration of feeding the sample, Choose a=0.3-0.02k and a(neighbor)=0.2-0.02k, size of neighbor being 3, perform SOM clustering and show the final results.

d. Based on your result, discuss the following issues:

[1]. How would the choice of a and a(neighbor) influence the clustering result?
[2]. Including a shrinking rate for a and a(neighbor) is useful or not? Why?
[3]. The relation between SOM clustering and sequential k-means.

5. (**Density based clustering, 18%**) You are required to perform DBSCAN on the following points.



a. Suppose Eps is 1.5 and Minpts is 3, label each points as core, border, or noise points.

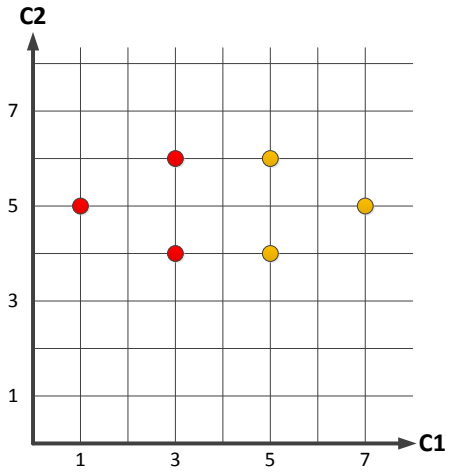
   Notes: the criteria for labeling a core point is that within Eps, the number of points is greater than or equal to the Minpts

b. Please describe the points that are directly and indirectly reachable from point R.

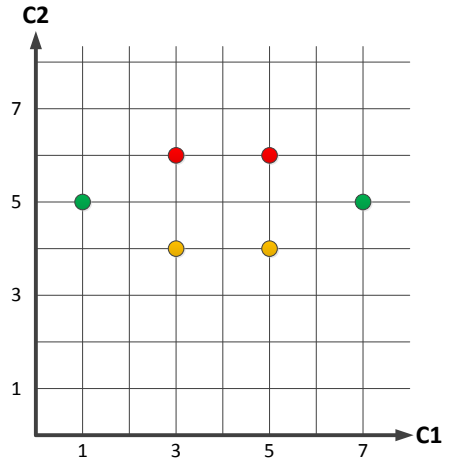
c. Suppose Eps and Minpts are not given and you are asked to set them. Describe how would you set the values and why.

6. (**Cluster cohesion and separation, 18%**) Suppose we have the following points with coordinates showed in figures.

a. They are divided into 2 clusters (red points and yellow points). Calculate the SSE, SSB and TSS.



b. They are divided into 3 clusters (red points, green points and yellow points). Calculate the SSE, SSB and TSS.



c. Consider a more general case. Suppose we have the dataset as a matrix

$$X = [X_1, \dots, X_k]$$

with each X_i contains n_i points (each point is a column vector) that belong to cluster i. Assume that the center of the whole dataset is 0. Introduce a vector e as a column vector of all ones (with same dimension as the point). Write SSE, SSB and TSS using the symbols provided (X, n_i, e).

d. Show the relations between SSE, SSB and TSS. Based on your finding, please write a new objective function for question 2 (a) so that maximizing the objective function equals to minimizing the in-cluster distances.