BUSINESS DATA MINING (IDS 572)

HOMEWORK 9
DUE DATE: WEDNESDAY, NOVEMBER 23 AT 3:00 PM

- Please provide succinct answers to the questions below.
- Please include all the R codes you use to get the results.
- You should submit an electronic pdf or word file in blackboard.
- Please include the names of all team-members in your write up and in the name of the file.

Problem 1. Take the following points in two dimensional space:

$$(8,4), (3,3), (4,5), (0,1), (10,2), (3,7), (0,9), (8,1), (4,3), (9,4).$$

For this exercise, use the Manhattan distance metric: for instance, the distance from (3,3) to (8,1) is

$$|3-8| + |3-1| = 7.$$

(a) Beginning with centroids at (1,1) and (8,8), do two iterations of the 2-means clustering algorithm, that is:

- allocate the points to centroids, then find the new centroids.
- again allocate the points to the centroids, and then get the new centroids.

If a point is equidistant between the centroids, assign it to the centroid that starts at (1,1). What are the resulting centroids and resulting clusters?

(b) Suppose we are interested in a binary (yes, no) output. Suppose outputs for the points above are yes, yes, no, no, yes, yes, no, no, yes, yes respectively. Consider the point (5,3).

i. What are the three closest points in our data set?

ii. Using the K-nearest neighbors approach, what would be the predicted output for (5,3) using K = 3 neighbors? (Use equal weights for each of three closest neighbors.)

Problem 2. Use single and complete link agglomerative clustering to group the data described by the following distance matrix. Show the dendrograms.

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

Problem 3. Download the file prospects.csv and load it into SPSS Modeler. The meaning of the fields is as follows (interval means continuous):

Name	Model Role	Measurement Level	Description
AGE	Input	Interval	Age in years
INCOME	Input	Interval	Annual income in thousands
MARRIED	Input	Binary	1=married, 0=not married
GENDER	Input	Binary	F=female, M=male
OWNHOME	Input	Binary	1=homeowner, 0=not a homeowner
LOCATION	Rejected	Nominal	Location of residence (A-H)
CLIMATE	Input	Nominal	Climate code for residence (10,20, & 30)
FICO	Input	Interval	Credit score
ID	ID	Nominal	Unique customer identification number

In your clustering model exclude the fields LOCATION and ID. Use R to answer the questions below.

(a) Use the K-means method to cluster the prospects dataset. Set the number of clusters to four. How many points are in each cluster? What are cluster means and variances?

(b) For each of the four clusters, briefly describe the characteristics of members of that cluster.

(c) What is the best value of k for this data set?