Homework 2. Released: 2/3/16; Due: 2/13/16.
Note: This h/w contains two parts.

Part A is for grading and you have to submit your best answer. If the question is ambiguous for you, then make suitable assumptions (and justify your assumptions).

Part B is for practice; you don't need to submit solution.

Part A (50 points). All answers should be your own.

Table 1: Data for K-Means

ID	A1	A2
1	1	1
2	2	1
3	1	2
4	2	2
5	3	3
6	3	4
7	4	5
8	2	4

Q1. K-Means Clustering (15 points)

(a) Why k-means clustering is sub-optimal? (1 points)

(b) List 3 advantages and 3 disadvantages of k-means clustering? (3 points)

(c) As K-means is sub-optimal, describe 3 ways to select initial centroids that may help finding good solution. (3 points)

(d) Using the data in Table 1, show first 3 iterations of K-means with following initial centroids. (Show both calculations and scatter plot for each iteration) (7 points)

(e) Using the k-means cluster solution found in (d), predict label for points (3,1), and (2,3). (1 point)

Initial centroids for K-Means: m1 = (1,3), m2 = (2,3)

Table 2: Data for Hierarchical clustering

ID	A1	A2
1	1	1
2	2	1
3	2	3
4	3	4
5	5	3

Q2. Hierarchical Clustering (15 points)

- **(a)** Using the data in table 2, perform hierarchical clustering using min, and max; show resulting distance matrix at each step (10 points)
- **(b)** Draw nested cluster and dendrogram for each of final clustering.

Q3. DBSCAN algorithm (10 points)

- (a) Define the core, border, and noise points (3 points)
- (b)

(b) Mark core, border, and noise points for DBSCAN data given in Figure 1. Assume unit squares. MinPoints = 2, and radius (EpsilonDist) = 1.5 units. Please note all data points (*) are at the intersection of grid lines. (5 points)

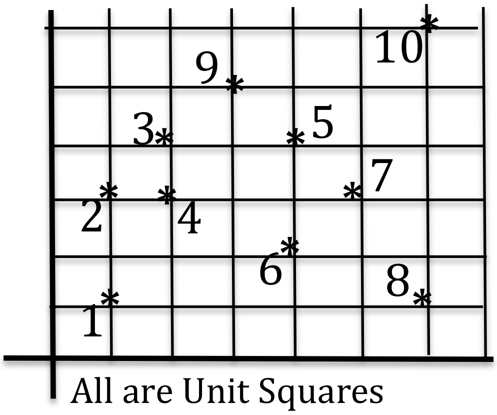

Figure 1. Data for DBSCAN

(c) Explain the conditions (situations) where DBSCAN does not work well (**1 point**)

(d) Is DBSCAN deterministic? If not, explain how can you make it deterministic. (**1 point**)

Q4. "R" Mini Project (10 points)

Using the supplied data (hw2-data.csv)

- (a) Perform K-Means clustering (for K=2, 3, 4, 5, 6) and for best clustering submit scatter plot diagram (each cluster should be colored differently) (**5 points**)
- (b) Perform Hierarchical clustering using single, complete, average, and centroid measures. Cut the tree such that you will produce 4 clusters. Show the resulting dendrograms (after cutting; not full trees) for each distance measure. (**5 points**)

Part B (Don’t submit solution; for your practice; from end of book chapter exercise). Collaboration is encouraged to solve Part B.

(From Chapter 8.7 Exercises)
3, 5, 7, 9, 14, 20