Data Mining

Homework 6

Important Notes:

1. Submit in electronic form before 11:59pm on Wednesday, May 31, 2017
2. IMPORTANT: If you submit before 11:59pm on Wednesday, May 24, 2017 it is very likely that you will receive feedback through Gradescope.
3. No late homework will be accepted.
4. The homework should be completed and submitted by each individual.
5. The homework should be submitted through Gradescope. Entry Code: **9BW66M**
6. The homework should be written in English.
7. The HW is worth it 10 points.
8. The [Research] questions require from you to do some research on the Web and get to understand things that were not covered during the lecture.
9. For questions, please use Piazza (English only!)

Exercise 1: K-means [2 pts]

Given the following points: 2,4,10,12,3,20,30,11,25. Assume k = 3, and that we randomly pick the initial means $\mu1 = 2$, $\mu2 = 4$ and $\mu3 = 6$. Show the cluster assignments obtained using K-means algorithm after one iteration, and show the new means for the next iteration.

First iteration

1 (2): 2, 3
2 (4): 4
3 (6):10, 12, 20, 30, 11, 25

New means:

1 (2.5)
2 (4)
3 (21.6)

Exercise 2: K-means with different distances [4 pts]

Given the two-dimensional points in the Table below, assume that k=2, and that initially the points are assigned to clusters as follows: C1 = {**x1, x2, x4**} and C2 = {**x3, x5**}. Answer the following questions:

A. Apply the K-means algorithm until convergence, that is, the clusters do not change, assuming the usual Euclidean distance (or otherwise called the L2-norm) as the distance between points, defined as $¿ \backslash x_i - x_j \backslash ¿_2 = \sqrt{\sqsubset}$

   Write down for each iteration (i) the coordinates of the means/centroids and (ii) the distances of all data points to these means/centroids, and (iii) the clusters after each iteration.

B. Apply the K-means algorithm until convergence, that is, the clusters do not change, assuming the Manhattan distance (or otherwise called the L1-norm) as the distance between points, defined as

   $$¿ x_{id} - x_{jd} \backslash ¿$$

   $$¿ \backslash x_i - x_j \backslash ¿_1 = \sum_{d=1}^{m} ¿$$

   Write down for each iteration (i) the coordinates of the means/centroids and (ii) the distances of all data points to these means/centroids, and (iii) the clusters after each iteration.

	X_1	X_2
$\mathbf{x_1}$	0	2
$\mathbf{x_2}$	0	0
$\mathbf{x_3}$	1.5	0
$\mathbf{x_4}$	5	0
$\mathbf{x_5}$	5	2

C1 = (0, 2), (0, 0), (5, 0)

C2 = (1.5, 0), (5, 2)

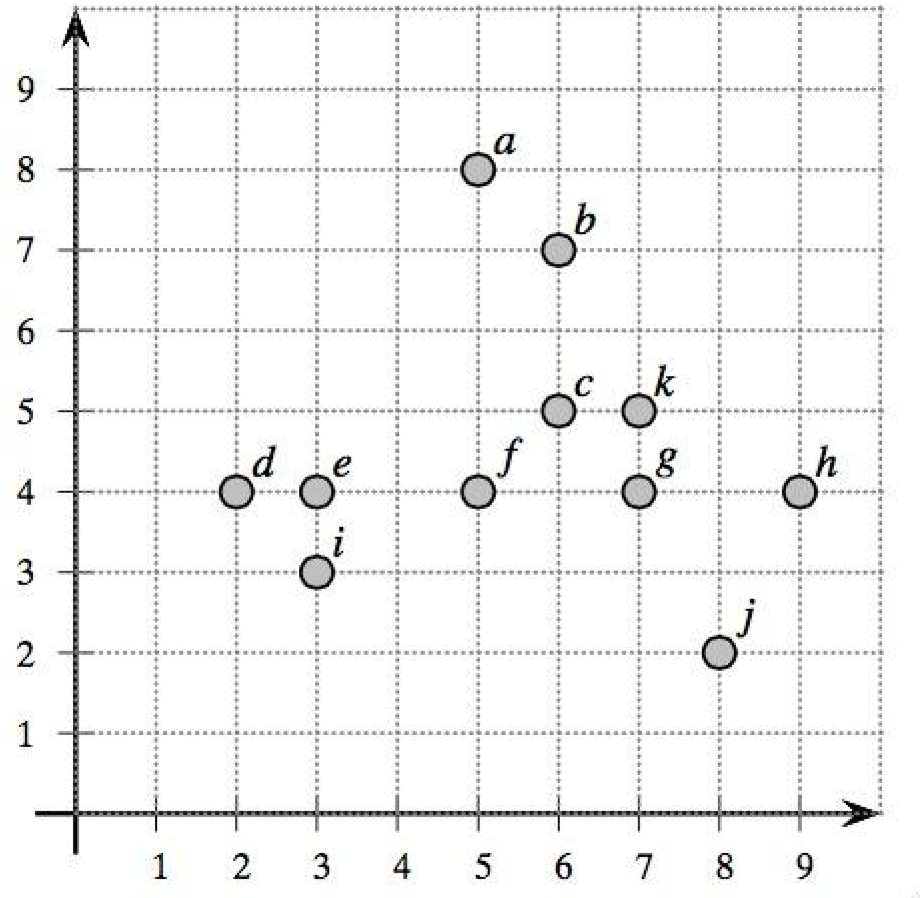
Exercise 3: Hierarchical clustering [4 pts]

Given the dataset in the Figure below, show the dendrogram resulting from the single-link hierarchical agglomerative clustering approach using the L1-norm as the distance between points

$$\text{¿}\, x_{id} - x_{jd} \vee\text{¿}$$

$$\text{¿}\,\backslash x_i - x_j \vee\text{¿}_1 = \sum_{d=1}^{m} \text{¿}$$

Whenever there is a choice, merge the cluster that has the lexicographically smallest labeled point. Show the cluster merge order in the tree, stopping when you have k = 4 clusters.


9
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
9
a
b
c
k
d
e
f
g
h
i
j