Machine Learning in Finance Assignment 1

1. We want to perform K-means clustering manually with K = 2 on a small example of six observations with two features.

Obs	1	2	3	4	5	6
X1	1	1	6	5	0	4
X2	4	3	2	1	4	0

We use the Euclidean distance. Suppose that we initially assign the observations #1, #2, #3 as cluster 1 and the observations #4, #5, #6 as cluster 2.

a) What are cluster centroids and cluster assignments after the first iteration of K-means clustering?
b) Continue the algorithm of K-means clustering until it converges. Report the cluster centroids and cluster assignments after each iteration.

First question (K-mean clustering question)

a) First iteration centroid of cluster1 (C1) and cluster2 (C2):

$$C_1^{1\,iter} = \left(\left(\frac{1+1+6}{3}\right), \left(\frac{4+3+2}{3}\right)\right) = (2.67,3)$$

$$C_2^{1\,iter} = \left(\left(\frac{5+0+4}{3}\right), \left(\frac{1+4+0}{3}\right)\right) = (3,1.67)$$

Then we compute the distance between each observation and the current centroids:

$$E.D. = \sqrt{(x_1^c - x_1^o)^2 + (x_2^c - x_2^o)^2}$$

Obs	1	2	3	4	5	6
X1	1	1	6	5	0	4
X2	4	3	2	1	4	0
E.d. to C1	1.94	1.67	3.48	3.07	2.85	3.28
E.d. to C2	3.07	2.40	3.02	2.11	3.80	1.94
New Cluster	1	1	2	2	1	2

Second iteration centroid of cluster1 (C1) and cluster2 (C2):

$$C_1^{2\,iter} = \left(\left(\frac{1+1+0}{3}\right), \left(\frac{4+3+4}{3}\right)\right) = (0.67,3.67)$$

$$C_2^{2\,iter} = \left(\left(\frac{6+5+4}{3}\right), \left(\frac{2+1+0}{3}\right)\right) = (5,1)$$

b) Continue the algorithm of K-means clustering until it converges

Then we compute the distance between each observation and the 2nd iteration centroids:

Obs	1	2	3	4	5	6
X1	1	1	6	5	0	4
X2	4	3	2	1	4	0
E.d. to C1	0.47	0.75	5.59	5.09	0.75	4.96
E.d. to C2	5	4.47	1.41	0	5.83	1.41
New Cluster	1	1	2	2	1	2

$$C_1^{3\ iter} = \left(\left(\frac{1+1+0}{3}\right), \left(\frac{4+3+4}{3}\right)\right) = (0.67, 3.67)$$

$$C_2^{3\ iter} = \left(\left(\frac{6+5+4}{3}\right), \left(\frac{2+1+0}{3}\right)\right) = (5, 1)$$

The cluster assignments are unchanged. Thus, after the second iteration, the algorithm converges while the cluster centroids are (0.67,3.67) and (5,1), and the cluster assignments are still the same as before: the 1st, 2nd, 5th observations in cluster 1; the 3rd, 4th, 6th observations in cluster 2.

2. Suppose that we have five observations, for which we compute a similarity (distance) matrix as follows:

$$\begin{bmatrix} 0 & 8 & 3 & 6 & 11 \\ 9 & 0 & 7 & 5 & 10 \\ 3 & 7 & 0 & 9 & 2 \\ 6 & 5 & 9 & 0 & 8 \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix}$$

   a) On the basis of the similarity matrix, sketch the dendogram that results from hierarchically clustering these 5 observations using complete linkage. Be sure to indicate on the plot the height at which each fusion occurs, as well as the observations corresponding to each leaf in the dendrogram.
   b) Repeat (a), this time using single linkage clustering.

Question 2 hierarchically clustering

a) Complete linkage (take the highest value between the 2 merge points

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

2 is the lowest value, we have to select it and merge point 3 and 5.

To update the matrix, we have to update the distance for the new point 35

$d(1,3) = 3$	$D(1,35) = 11$
$d(1,5) = 11$	
$d(2,3) = 7$	$D(2,35) = 10$
$d(2,5) = 10$	
$d(4,3) = 9$	$D(4,35) = 9$
$d(4,5) = 8$	

	35	1	2	4
35	0			
1	11	0		
2	10	9	0	
4	9	6	5	0

5 is the lowest value, we have to select it and merge point 2 and 4.

To update the matrix, we have to update the distance for the new point 24

$d(35,2) = 10$	$D(35,24) = 10$
$d(35,4) = 9$	
$d(1,2) = 9$	$D(1,24) = 9$
$d(1,4) = 6$	

	35	24	1
35	0		
24	10	0	
1	11	9	0

9 is the lowest value, we have to select it and merge point 24 and 1.

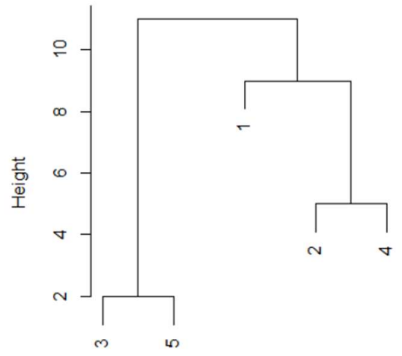
To update the matrix, we have to update the distance for the new point 241

$d(35,24) = 10$	$D(35,241) = 11$
$d(35,1) = 11$	

	35	241
35	0	
241	11	0

Finally!

Plot for complete linkage



b) Single linkage (take the lowest value between the 2 merge points)

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

2 is the lowest value, we have to select it and merge point 3 and 5.

To update the matrix, we have to update the distance for the new point 35

$d(1,3) = 3$	$D(1,35) = 3$
$d(1,5) = 11$	
$d(2,3) = 7$	$D(2,35) = 7$
$d(2,5) = 10$	
$d(4,3) = 9$	$D(4,35) = 8$
$d(4,5) = 8$	

	35	1	2	4
35	0			
1	3	0		
2	7	9	0	
4	8	6	5	0

3 is the lowest value, we have to select it and merge point 35 and 1.

To update the matrix, we have to update the distance for the new point 351

$d(2,35) = 7$	$D(2,351) = 7$
$d(2,1) = 9$	
$d(4,35) = 8$	$D(4,351) = 6$
$d(4,1) = 6$	

	351	2	4
351	0		
2	7	0	
4	6	5	0

5 is the lowest value, we have to select it and merge point 2 and 4.

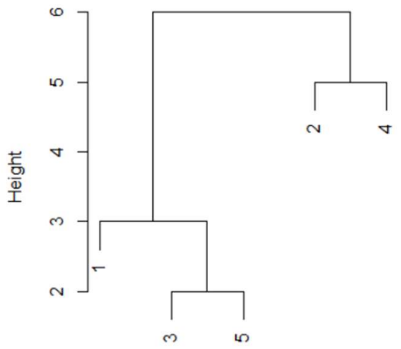
To update the matrix, we have to update the distance for the new point 24

$d(351,2) = 7$	$D(351,24) = 6$
$d(351,4) = 6$	

	351	24
351	0	
24	6	0

Finally!

Plot for single linkage



3. The table below provides a training data set containing eight observations, two predictors (having the same scale), and one quantitative response.

Obs	1	2	3	4	5	6	7	8
X1	5	5	2	2	0	3	3	2
X2	1	4	4	3	4	0	3	5
Y	4.75	5.56	2.16	1.97	0.40	2.38	2.95	2.37

We wish to use this data set to make a prediction for Y when (X1, X2) = (1, 3) using K-nearest neighbors (KNN) methods. We want our prediction has low mean square error.

a) Compute the Euclidean distance between each observation and the test point (X1, X2) = (1, 3).

b) What is your KNN prediction with K = 3?

c) What is your KNN prediction with K = 5?

a)

Obs	1	2	3	4	5	6	7	8
X1	5	5	2	2	0	3	3	2
X2	1	4	4	3	4	0	3	5
E.d. with (1,3)	4.47	4.12	1.41	1	1.41	3.61	2	2.236
Rank	8	7	2	1	2	6	4	5
Y	4.75	5.56	2.16	1.97	0.40	2.38	2.95	2.37

b)

K=3, $y = \frac{1.97+0.40+2.16}{3} = 1.51$

c)

K=5, $y = \frac{1.97+0.40+2.16+2.95+2.37}{3} = 1.97$