Lab 8: 21 May 2012

Exercises on Clustering

1. Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). Suppose that the initial seeds (centers of each cluster) are A1, A4 and A7. Run the k-means algorithm for 1 epoch. At the end of this epoch show:
    a. The new clusters (i.e. the examples belonging to each cluster);
    b. The centers of the new clusters;
    c. Draw a 10 by 10 space with all the 8 points and show the clusters after the first epoch and the new centroids.
    d. How many more iterations are needed to converge? Draw the result for each epoch.

Solution

The Euclidean distances between the given points are in the following matrix:

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

a.

seed1=A1=(2,10), seed2=A4=(5,8), seed3=A7=(1,2)

epoch1 – start:

A1:
d(A1, seed1)=0 as A1 is seed1
d(A1, seed2)= $\sqrt{13}$ >0
d(A1, seed3)= $\sqrt{65}$ >0
➔A1 ∈ cluster1

A2:
d(A2,seed1)= $\sqrt{25}$ = 5
d(A2, seed2)= $\sqrt{18}$ = 4.24
d(A2, seed3)= $\sqrt{10}$ = 3.16 ← smaller
➔ A2 ∈ cluster3

A3:
d(A3, seed1)= $\sqrt{36}$ = 6
d(A3, seed2)= $\sqrt{25}$ = 5 ← smaller
d(A3, seed3)= $\sqrt{53}$ = 7.28
➔ A3 ∈ cluster2

A4:
d(A4, seed1)= $\sqrt{13}$
d(A4, seed2)=0 as A4 is seed2
d(A4, seed3)= $\sqrt{52}$ >0
➔ A4 ∈ cluster2

A5:

d(A5, seed1)= $\sqrt{50}$ = 7.07

d(A5, seed2)= $\sqrt{13}$ = 3.60 ← smaller

d(A5, seed3)= $\sqrt{45}$ = 6.70

➔ A5 $\in$ cluster2

A6:

d(A6, seed1)= $\sqrt{52}$ = 7.21

d(A6, seed2)= $\sqrt{17}$ = 4.12 ← smaller

d(A6, seed3)= $\sqrt{29}$ = 5.38

➔ A6 $\in$ cluster2

A7:

d(A7, seed1)= $\sqrt{65}>0$

d(A7, seed2)= $\sqrt{52}>0$

d(A7, seed3)=0 as A7 is seed3

➔ A7 $\in$ cluster3

A8:

d(A8, seed1)= $\sqrt{5}$

d(A8, seed2)= $\sqrt{2}$ ← smaller

d(A8, seed3)= $\sqrt{58}$

➔ A8 $\in$ cluster2

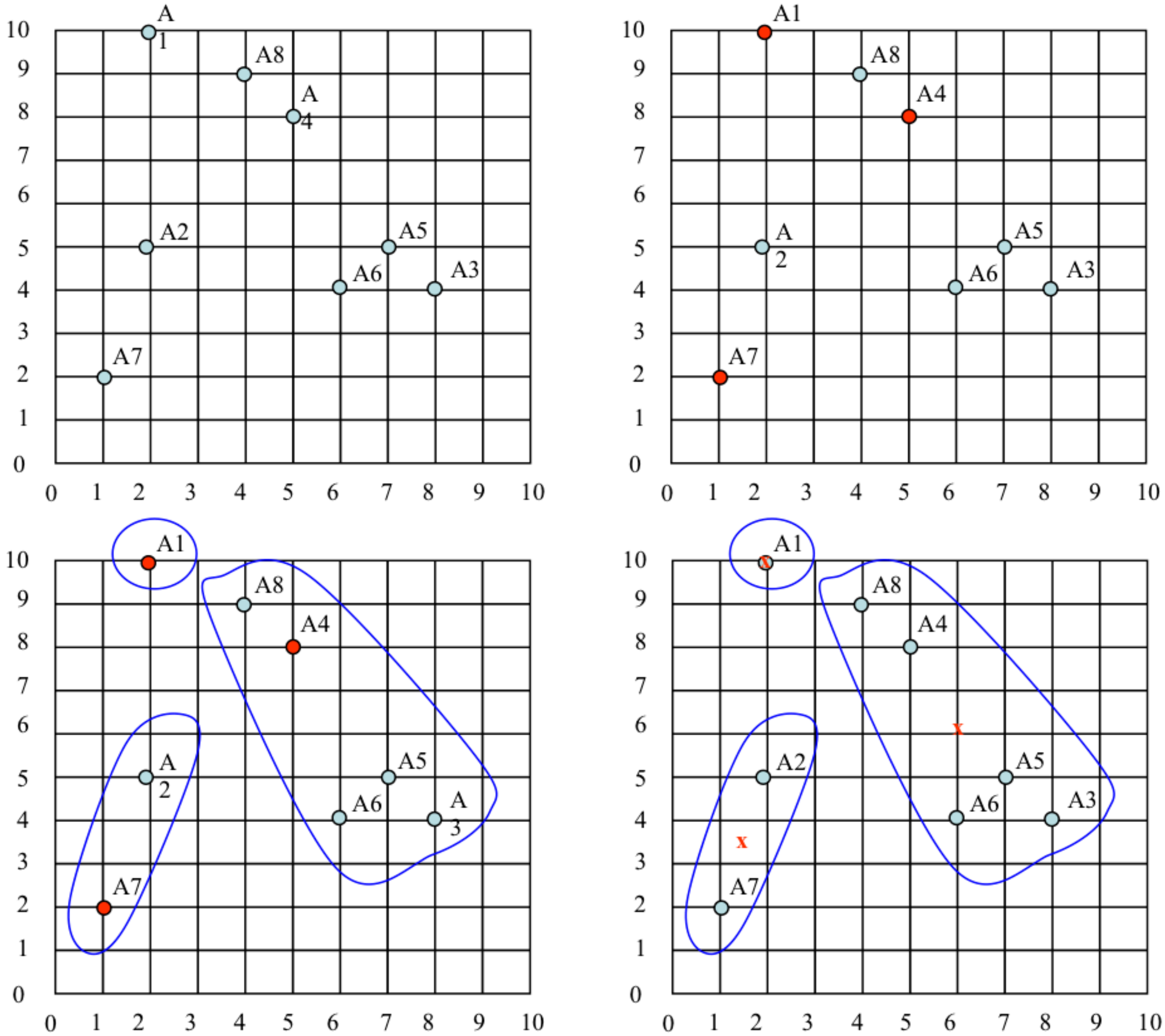
end of epoch1

new clusters: 1: {A1}, 2: {A3, A4, A5, A6, A8}, 3: {A2, A7}

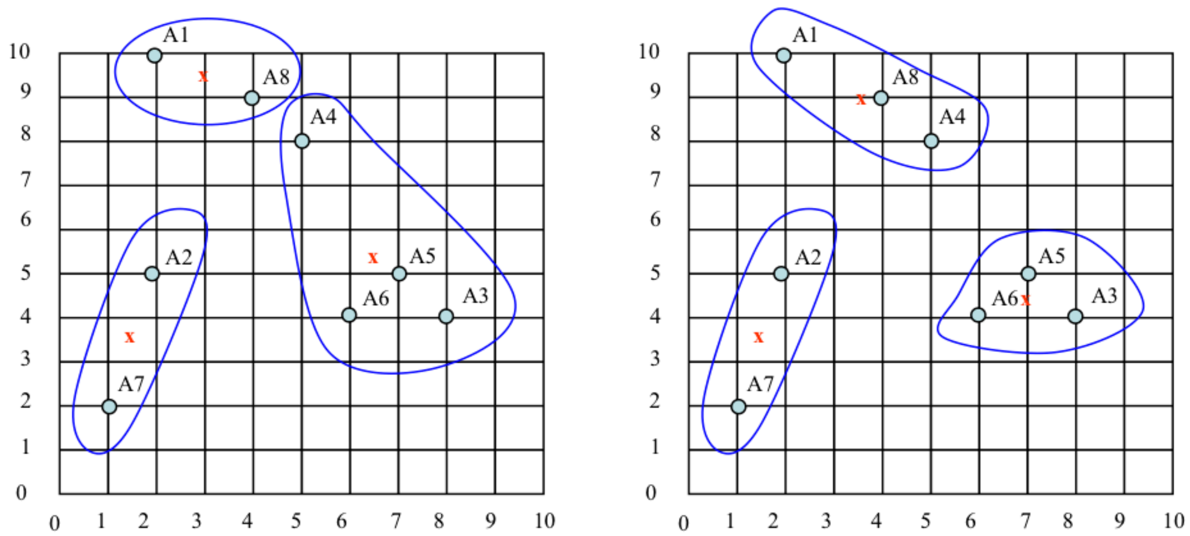
b) centers of the new clusters:

C1= (2, 10), C2= ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6), C3= ((2+1)/2, (5+2)/2) = (1.5, 3.5)

c)



d)
We would need two more epochs. After the 2nd epoch the results would be:
1: {A1, A8}, 2: {A3, A4, A5, A6}, 3: {A2, A7}
with centers C1=(3, 9.5), C2=(6.5, 5.25) and C3=(1.5, 3.5).
After the 3rd epoch, the results would be:
1: {A1, A4, A8}, 2: {A3, A5, A6}, 3: {A2, A7}
with centers C1=(3.66, 9), C2=(7, 4.33) and C3=(1.5, 3.5).



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

	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

Solution

1. Single link: distance between two clusters is the shortest distance between a pair of elements from the two clusters.

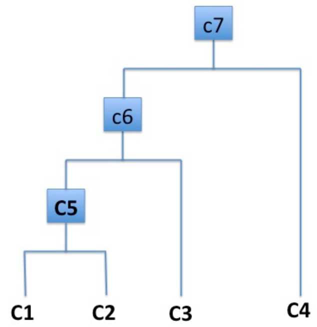
   We apply the algorithm presented in lecture 10 (ml_2012_lecture_10.pdf), page 4.

   At the beginning, each point A,B,C, and D is a cluster → c1 = {A}, c2={B}, c3={C}, c4={D}

   Iteration 1
   The shortest distance is d(c1,c2)=1 → c1 and c2 are merged → the clusters are c3={C}, c4={D}, c5={A,B}
   The distances from the new cluster to the others are d(c5,c3) = 2, d(c5,c4)=5

   Iteration 2
   The shortest distance is d(c5,c3)=2 → c5 and c3 are merged → the clusters are c6={A,B,C}, c4={D}
   The distances from the new cluster to the others are: d(c6,c4)=3

   Iteration 3
   c6 and c4 are merged → the final cluster is c7={A,B,C,D}



The dendrogram is

2. Complete link: The distance between two clusters is the distance of two furthest data points in the two clusters
   We apply the algorithm presented in lecture 10 (ml_2012_lecture_10.pdf) page 4.

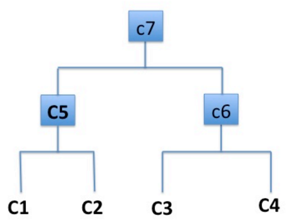
   At the beginning, each point A,B,C, and D is a cluster → c1 = {A}, c2={B}, c3={C}, c4={D}

   Iteration 1
   The shortest distance is d(c1,c2)=1 → c1 and c2 are merged → the clusters are c3={C}, c4={D}, c5={A,B}
   The distances from the new cluster to the others are: d(c5,c3) = 4, d(c5,c4)=6

   Iteration 2
   The shortest distance is d(c3,c4)=3 → c3 and c4 are merged → the clusters are c6={C,D}, c5={A,B}
   The distances from the new cluster to the others are: d(c6,c5)=6

   Iteration 3
   c6 and c5 are merged → the final cluster is c7={A,B,C,D}

   The dendrogram is



3. Use single-link complete-link, average-link, and centroid agglomerative clustering, to cluster the following 8 examples: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9). Show the dendrograms.

Solution
The solutions for single-link and complete-link are analogous to the previous one. The solutions for average-link and centroid are also similar, what is changing is the calculation of the distances between clusters.

- For average link the distance is the average of all the distances between points belonging to the two clusters. For instance if c1={A,B} and c2={C,D},
  dist(c1, c2) = (dist(A,B) + dist(A,D) + dist(B,C) + dist(B,D)) / 4

- For centroid the distance between two cluster is the distance between their centroids.

4. Consider a data set in two dimensions with five data points at: $\{(1, 0), (-1, 0), (0, 1), (3, 0), (3, 1)\}$. Run two iterations of k-means by hand with initial points at $(-1, 0)$ and $(3, 1)$. What are the assignments at each iteration and what are the centroids? Has the algorithm converged?

Solution
The solution is analogous to the solution of Exercise 1.

5. How can we make k-means robust to outliers? Explain the two methods we have seen.

Solution
Refer to lecture 9 (ml_2012_lecture_09.pdf), pages 15-16.

6. Explain the main similarities and differences between k-means and hierarchical clustering.

Solution
Refer to lecture 9 (ml_2012_lecture_09.pdf) and lecture 10 (ml_2012_lecture_10.pdf).

7. Give two examples of real-world applications of clustering.

Solution
Refer to lecture 9 (ml_2012_lecture_09.pdf), page 9.

8. Which are the stopping criteria for the k-means algorithm?

Solution
Refer to lecture 9 (ml_2012_lecture_09.pdf), page 12.

9. Is the result of k-means clustering sensitive to the choice of the initial seeds? How? Make an example.

Solution
Refer to lecture 9 (ml_2012_lecture_09.pdf), page 17.

10. Which is a good algorithm for finding clusters of arbitrary shape? Is finding these clusters always a good idea? When it is not?

Solution
Refer to lecture 9 (ml_2012_lecture_09.pdf), page 21 and to lecture 10 (ml_2012_lecture_10.pdf), page 5.

11. Explain the general algorithm for agglomerative hierarchical clustering.

Solution
Refer to lecture 10 (ml_2012_lecture_10.pdf), pages 3-4.

12. Explain the single-link and the complete-link methods for hierarchical clustering.

Solution
Refer to lecture 10 (ml_2012_lecture_10.pdf), pages 5-6.

13. Make 2 examples of distance functions that can be used for numeric attributes.

Solution
Refer to lecture 10 (ml_2012_lecture_10.pdf), pages 8-9.