UECS3213 / UECS3453 DATA MINING

SESSION: January 2019

TUTORIAL 5

Chapter 5 - Cluster Analysis and Anomaly Detection

Part 1: Cluster Analysis

1. Briefly describe the following approaches to clustering: *partitioning methods, hierarchical methods* and *density-based methods*. Give examples in each case.

Answer:
Clustering is the process of grouping data into classes, or clusters, so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters.
There are several approaches to clustering.

- **Partitioning methods**: Given a databases of n objects or data tuples, a partitioning methods constructs k partitions of data, where each partition represents a cluster and k ≤ n. Given k, the number of partitions to construct, it creates an initial partitioning. It then uses an iterative relocation technique that attempts to improve the partitioning by moving objects from one group to another. The general criterion of a good partitioning is that objects in the same cluster are "close" or related to each other, whereas objects of different clusters are "far apart". The k-means algorithm is a commonly used partitioning method.
- **Hierarchical methods**: A hierarchical method creates a hierarchical decomposition of the given set of data objects. It can be either agglomerative or divisive. The agglomerative (bottom-up) approach starts with each object forming a separate group. It successively merges the objects that are close to one another, until all of the groups are merged into one, or until a termination condition holds. The divisive (top-down) approach starts with all of the objects in the same cluster. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object forms its own cluster or until a termination condition holds. AGNES and DIANA are examples of hierarchical clustering. BIRCH integrates hierarchical clustering with iterative (distance-based) relocation.
- **Density-based methods**: These methods are based on the notion of density. The main idea is to continue growing a given cluster as long as the density in its "neighborhood" exceeds some threshold. That is, for each data point within a given cluster, the neighborhood of a given radius has to contain at least a minimum number of points. This method can be used to filter out noise and discover clusters of arbitrary shape. DBSCAN and OPTICS are typical examples of density-based clustering.

2. Identify the clusters in Figure 8.3 using the *center-*, *contiguity-*, and *density-based* definitions. Also indicate the number of clusters for each case and give a brief indication of your reasoning. Note that darkness or the number of dots indicates density. If it helps, assume center-based means *K-means*, contiguity-based means *single link*, and density-based means *DBSCAN*.

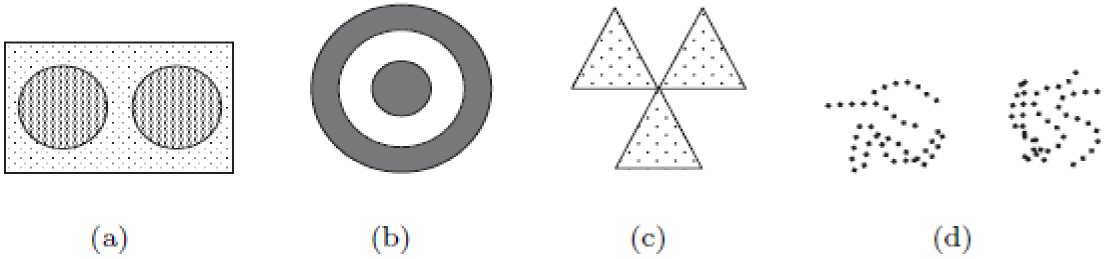


Figure 8.3. Clusters for Exercise 5.

a) center-based 2 clusters. The rectangular region will be split in half. Note that the noise is included in the two clusters.
contiguity-based 1 cluster because the two circular regions will be joined by noise.
density-based 2 clusters, one for each circular region. Noise will be eliminated.

b) center-based 1 cluster that includes both rings.
contiguity-based 2 clusters, one for each rings.
density-based 2 clusters, one for each ring.

c) center-based 3 clusters, one for each triangular region. One cluster is also an acceptable answer.

contiguity-based 1 cluster. The three triangular regions will be joined together because they touch.

density-based 3 clusters, one for each triangular region. Even though the three triangles touch, the density in the region where they touch is lower than throughout the interior of the triangles.

d) center-based 2 clusters. The two groups of lines will be split in two.
contiguity-based 5 clusters. Each set of lines that intertwines becomes a cluster.
density-based 2 clusters. The two groups of lines define two regions of high density separated by a region of low density.

3. Suppose that for a data set

- there are m points and K clusters,
- half the points and clusters are in "more dense" regions,
- half the points and clusters are in "less dense" regions, and
- the two regions are well-separated from each other.

For the given data set, which of the following should occur in order to minimize the squared error when finding K clusters:

a) Centroids should be equally distributed between more dense and less dense regions.
b) More centroids should be allocated to the less dense region.
c) More centroids should be allocated to the denser region.

Answer:

The correct answer is (b). Less dense regions require more centroids if the squared error is to be minimized.

Note: Do not get distracted by special cases or bring in factors other than density.

4. Total SSE is the sum of the SSE for each separate attribute. What does it mean if the SSE for one variable is
   a) low for all clusters?
   b) low for just one cluster?
   c) high for all clusters?
   d) high for just one cluster?
   e) how could you use the per variable SSE information to improve your clustering?

a) If the SSE of one attribute is low for all clusters, then the variable is essentially a constant and of little use in dividing the data into groups.

b) if the SSE of one attribute is relatively low for just one cluster, then this attribute helps define the cluster.

c) If the SSE of an attribute is relatively high for all clusters, then it could well mean that the attribute is **noise**.

d) If the SSE of an attribute is relatively high for one cluster, then it is at odds with the information provided by the attributes with low SSE that define the cluster. It could merely be the case that the clusters defined by this attribute are different from those defined by the other attributes, but in any case, it means that this attribute does not help define the cluster.

e) The idea is to **eliminate attributes that have poor distinguishing power between clusters**, i.e., low or high SSE for all clusters, since they are useless for clustering. Note that attributes with high SSE for all clusters are particularly troublesome if they have a relatively high SSE with respect to other attributes (perhaps because of their scale) since they introduce a lot of noise into the computation of the overall SSE.

5. The *Voronoi diagram* for a set of K points in the plane is a partition of all the points of the plane into K regions, such that every point (of the plane) is assigned to the closest point among the K specified points. (See Figure 8.5.)

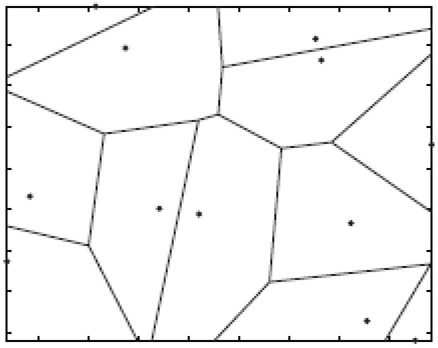


Figure 8.5. Voronoi diagram

a) What is the relationship between Voronoi diagrams and K-means clusters?
b) What do Voronoi diagrams tell us about the possible shapes of K-means clusters?

a) If we have K K-means clusters, then the plane is divided into K Voronoi regions that represent the points closest to each centroid.

b) The boundaries between clusters are piecewise linear. It is possible to see this by drawing a line connecting two centroids and then drawing a perpendicular to the line halfway between the centroids. This perpendicular line splits the plane into two regions, each containing points that are closest to the centroid the region contains.

6. Consider the following four faces shown in Figure 8.7. Again, darkness or number of dots represents density. Lines are used only to distinguish regions and do not represent points.

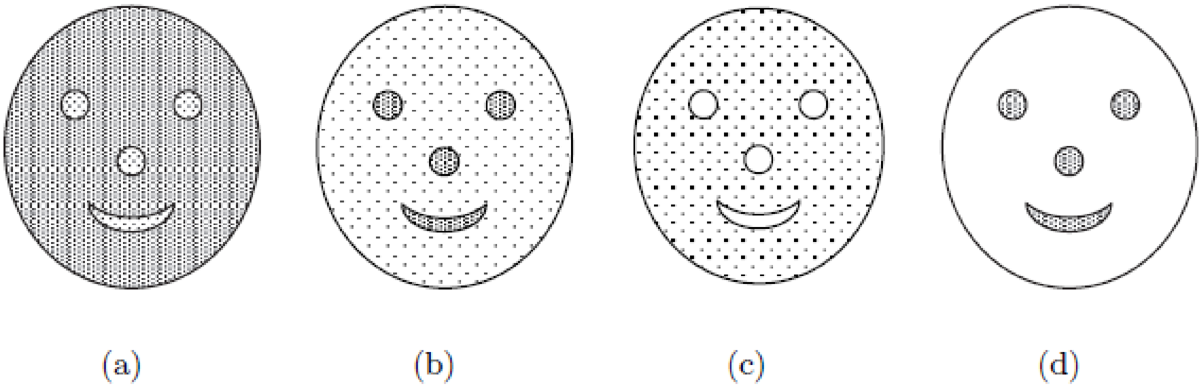


Figure 8.7.

a) For each figure, could you use single link to find the patterns represented by the nose, eyes, and mouth? Explain.

- Only for (b) and (d). For (b), the points in the nose, eyes, and mouth are much closer together than the points between these areas. For (d) there is only space between these regions.

b) For each figure, could you use K-means to find the patterns represented by the nose, eyes, and mouth? Explain.

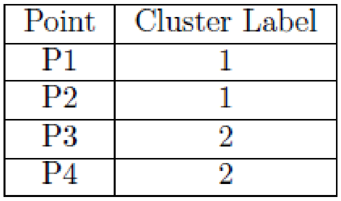
- Only for (b) and (d). For (b), K-means would find the nose, eyes, and mouth, but the lower density points would also be included. For (d), K-means would find the nose, eyes, and mouth straightforwardly as long as the number of clusters was set to 4.

c) What limitation does clustering have in detecting all the patterns formed by the points in Figure 8.7(c)?

- Clustering techniques can only find patterns of points, not of empty spaces.

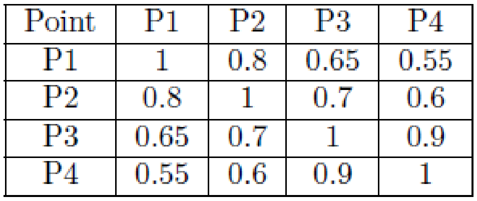
7. Given the set of cluster labels and similarity matrix shown in Tables 8.4 and 8.5, respectively,
   a) compute the correlation between the similarity matrix and the ideal similarity matrix, i.e., the matrix whose ij-th entry is 1 if two objects belong to the same cluster, and 0 otherwise.

Table 8.4. Table of cluster labels



Point	Cluster Label
P1	1
P2	1
P3	2
P4	2

Table 8.5. Similarity matrix



Point	P1	P2	P3	P4
P1	1	0.8	0.65	0.55
P2	0.8	1	0.7	0.6
P3	0.65	0.7	1	0.9
P4	0.55	0.6	0.9	1

Answer:

Point	P1	P2	P3	P4
P1	1	1	0	0
P2	1	1	0	0
P3	0	0	1	1
P4	0	0	1	1

We need to compute the correlation between the vector x =< 1, 0, 0, 0, 0, 1 > and the vector y =< 0.8, 0.65, 0.55, 0.7, 0.6, 0.9 >, which is the correlation between the off-diagonal elements of the distance matrix and the ideal similarity matrix.

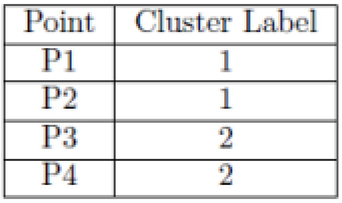
We get:
Standard deviation of the vector **x** : $\sigma_x = 0.5164$
Standard deviation of the vector **y** : $\sigma_y = 0.1304$

Covariance of **x** and **y**: cov(**x**, **y**) = 0.05

Therefore, corr(**x**, **y**) = cov(**x**, **y**)/$\sigma_x\sigma_y$ = 0.05/(0.5164 × 0.1304) = 0.7425

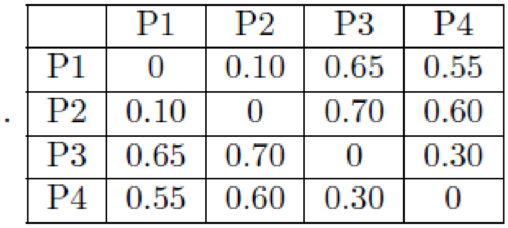
8. Compute the *silhouette coefficient* for each point, each of the two clusters, and the overall clustering, given the distances between points given in Table 8.3.

Table 8.4. Table of cluster labels



Point	Cluster Label
P1	1
P2	1
P3	2
P4	2

Table 8.3. Table of distances



	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

Cluster 1 contains {P1, P2}, Cluster 2 contains {P3, P4}. The dissimilarity matrix that we obtain from the similarity matrix is the following:

Let a indicate the average distance of a point to other points in its cluster.
Let b indicate the minimum of the average distance of a point to points in another cluster.
Point P1: SC = 1- a/b = 1 - 0.1/((0.65+0.55)/2)= 5/6 = 0.833
Point P2: SC = 1- a/b = 1 - 0.1/((0.7+0.6)/2) = 0.846
Point P2: SC = 1- a/b = 1 - 0.3/((0.65+0.7)/2) = 0.556
Point P2: SC = 1- a/b = 1 - 0.3/((0.55+0.6)/2) = 0.478
Cluster 1 Average SC = (0.833+0.846)/2 = 0.840
Cluster 2 Average SC = (0.556+0.478)/2 = 0.517
Overall Average SC = (0.840+0.517)/2 = 0.68

9. In a *silhouette* analysis, the average distance of a point $p_{i,}$ which is in cluster A to each of the 3 existing clusters were found to be:

Cluster	Average distance from p_i to each cluster member
A	24
B	48
C	72

Calculate the *silhouette coefficient* of point p_i to clusters A, B and C.

Answer:

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & \text{if } a(i) < b(i) \\ 0 & \text{if } a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 & \text{if } a(i) > b(i) \end{cases}$$

- a(i) : the average distance between 'i' and all other data within the same cluster (source)
- b(i) : the lowest average distance of 'i' to all points in any other cluster, of which 'i' is not a member (source)
- So, from the question, a(i) will be 24 as point p_i belongs to cluster A and b(i) will be 48 as it is the least average distance that p_i has from any other cluster than A (to which it belongs).
- So, as a(i) < b(i), silhouette coefficient s(i) = 1 - 24/48 = **0.5**

10. 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).

Calculate the distance matrix (Euclidean distance) of points A1 to A8. Fill in the table below.

	A1	A2	A3	A4	A5	A6	A7	A8
A1		$\sqrt{25}$						
A2	$\sqrt{25}$							
A3								
A4								
A5								
A6								
A7								
A8								

Suppose that the initial seeds (centers of each cluster) are A1, A4 and A7. Run the k-means algorithm for 1 epoch only. 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.

Answer:

The distance matrix based on the Euclidean distance is given below:

	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)

d(a,b) denotes the Eucledian distance between a and b. It is obtained directly from the distance matrix or calculated as follows: d(a,b)=sqrt((x_b-x_a)^2+(y_b-y_a)^2))
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 $\in$ cluster1

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

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

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

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

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

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

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

→ A5 $\in$ cluster2

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)= $\sqrt{58}$

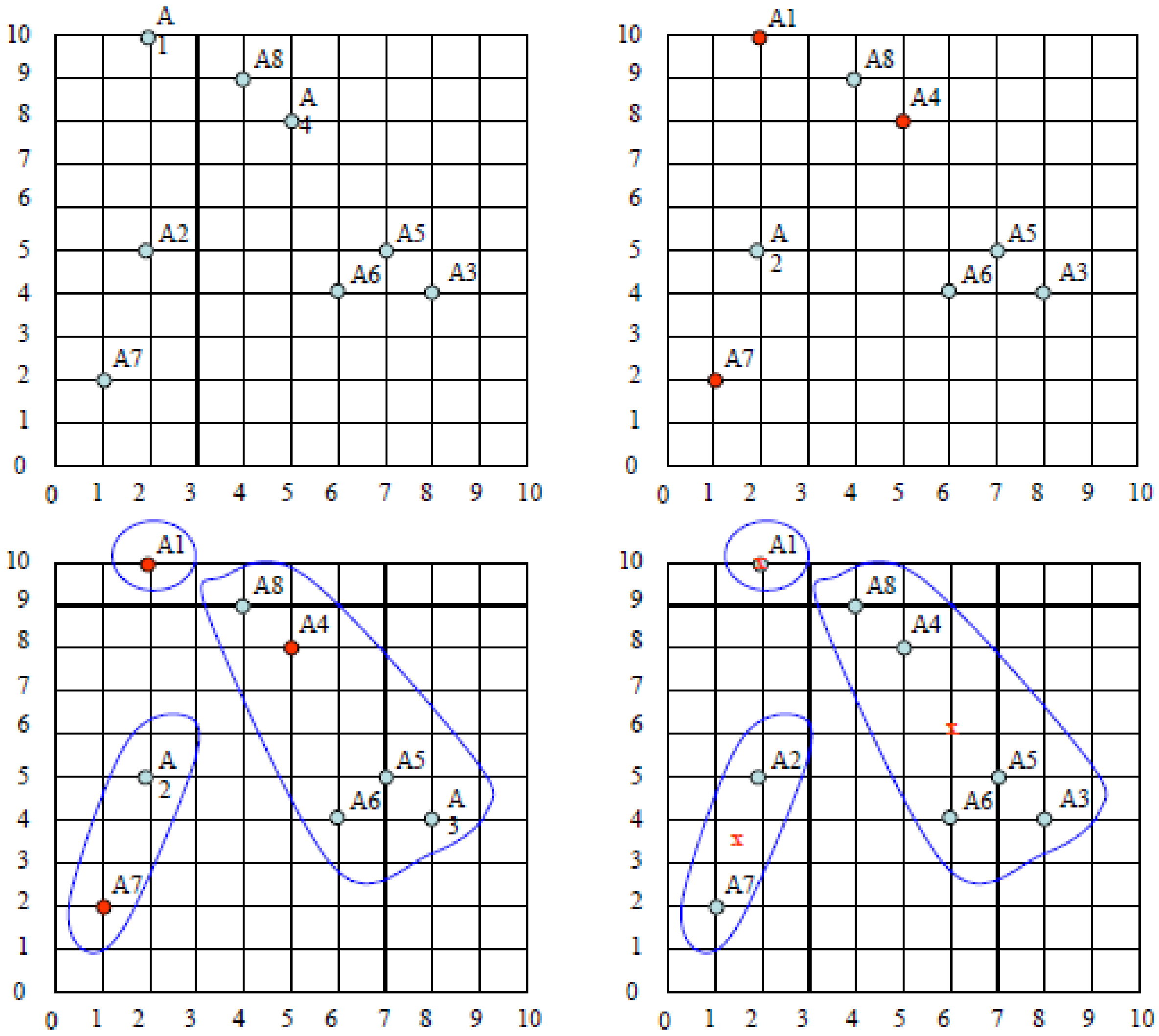
→ A8 $\in$ cluster2

end of epoch1

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

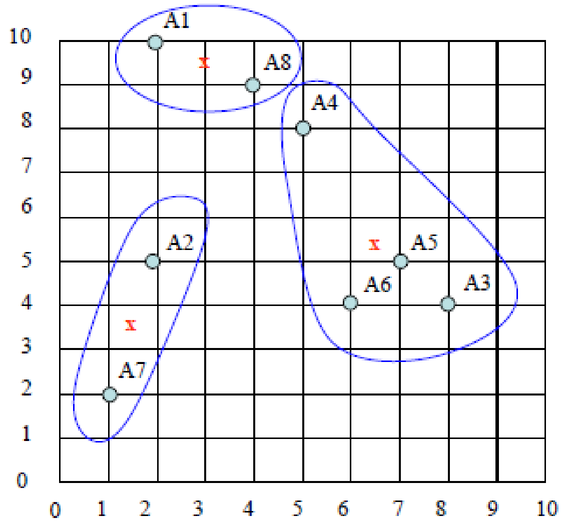
b) centers of the new clusters:
C1= (2, 10), C2= ((8+5+7+6+4)/5, (4+8+5+4+9)/5) = (6, 6), C3= ((2+1)/2, (5+2)/2) = (1.5, 3.5)

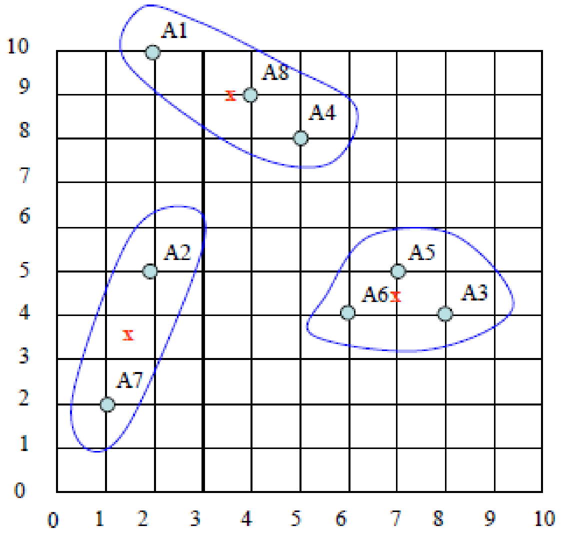
c)



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).





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

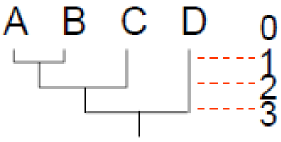
	A	B	C	D
A	0	1	4	5
B		0	2	6
C			0	3
D				0

Answer:

Agglomerative → initially every point is a cluster of its own and we merge cluster until we end-up with one unique cluster containing all points.

a) single link: distance between two clusters is the shortest distance between a pair of elements from the two clusters.

d	k	K	Comments
0	4	{A}, {B}, {C}, {D}	We start with each point = cluster
1	3	{A, B}, {C}, {D}	Merge {A} and {B} since A & B are the closest: d(A, B)=1
2	2	{A, B, C}, {D}	Merge {A, B} and {C} since B & C are the closest: d(B, C)=2
3	1	{A, B, C, D}	Merge D

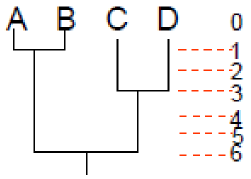


b) complete link: distance between two clusters is the longest distance between a pair of elements from the two clusters.

d	k	K	Comments
0	4	{A}, {B}, {C}, {D}	We start with each point = cluster
1	3	{A, B}, {C}, {D}	d(A,B)=1<=1 → merge {A} and {B}

2	3	{A, B}, {C}, {D}	d(A,C)=4>2 so we can't merge C with {A,B}
			d(A,D)=5>2 and d(B,D)=6>2 so we can't merge D with {A, B}
			d(C,D)=3>2 so we can't merge C and D

3	2	{A, B}, {C, D}	- d(A,C)=4>3 so we can't merge C with {A,B} - d(A,D)=5>3 and d(B,D)=6>3 so we can't merge D with {A, B} - d(C,D)=3 <=3 so merge C and D
4	2	{A, B}, {C, D}	{C,D} cannot be merged with {A, B} as d(A,D)= 5 >4 (and also d(B,D)= 6 >4) although d(A,C)= 4 <= 4, d(B,C)= 2<=4)
5	2	{A, B}, {C, D}	{C,D} cannot be merged with {A, B} as d(B,D)= 6 > 5
6	1	{A, B, C, D}	{C, D} can be merged with {A, B} since d(B,D)= 6 <= 6, d(A,D)= 5 <= 6, d(A,C)= 4 <= 6, d(B,C)= 2 <= 6



12. If Epsilon is 2 and minpoint is 2, what are the clusters that DBScan would discover with 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).

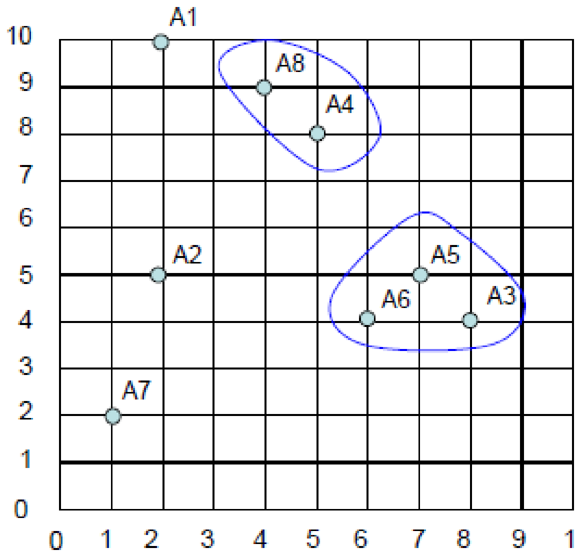
    The distance matrix is the same as the one in Exercise 1. Draw the 10 by 10 space and illustrate the discovered clusters. What if Epsilon is increased to 10?

    Answer:

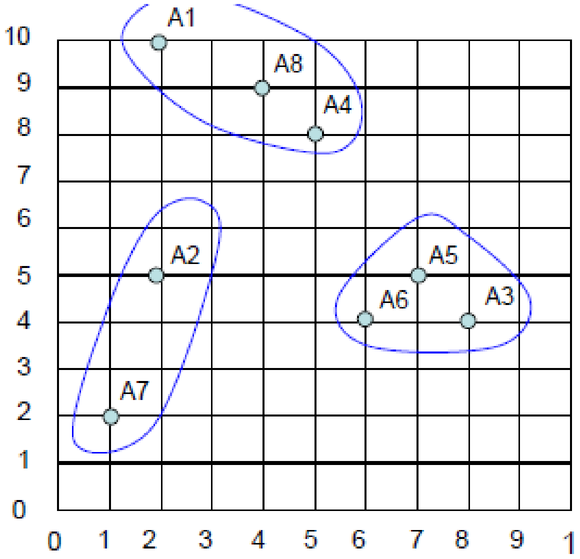
What is the Epsilon neighborhood of each point?
$N_2(A1)=\{\}$; $N_2(A2)=\{\}$; $N_2(A3)=\{A5, A6\}$; $N_2(A4)=\{A8\}$; $N_2(A5)=\{A3, A6\}$;
$N_2(A6)=\{A3, A5\}$; $N_2(A7)=\{\}$; $N_2(A8)=\{A4\}$

So A1, A2, and A7 are outliers, while we have two clusters C1={A4, A8} and C2={A3, A5, A6}

If Epsilon is $\sqrt{10}$ then the neighborhood of some points will increase:
A1 would join the cluster C1 and A2 would joint with A7 to form cluster C3={A2, A7}.



Epsilon = 2



Epsilon = $\sqrt{10}$

Part 2: Anomaly Detection

1. Compare and contrast the different techniques for anomaly detection: *proximity-based, density-based* and *model-based*. In particular, try to identify circumstances in which the definitions of anomalies used in the different techniques might be equivalent or situations in which one might make sense, but another would not. Be sure to consider different types of data.

- First, note that the proximity- and density-based anomaly detection techniques are related. Specifically, high density in the neighbourhood of a point implies that many points are close to it, and vice-versa. In practice, density is often defined in terms of distance, although it can also be defined using a grid-based approach.
- The model-based approach can be used with virtually any underlying technique that fits a model to the data. However, note that a particular model, statistical or otherwise, must be assumed. Consequently, model-based approaches are restricted in terms of the data to which they can be applied. For example, if the model assumes a Gaussian distribution, then it cannot be applied to data with a non-Gaussian distribution.
- On the other hand, the proximity- and density-based approaches do not make any particular assumption about the data, although the definition of an anomaly does vary from one proximity- or density-based technique to another. Proximity-based approaches can be used for virtually any type of data, although the proximity metric used must be chosen appropriately. For example, Euclidean distance is typically used for dense, low-dimensional data, while the cosine similarity measure is used for sparse, high-dimensional data.
- Since density is typically defined in terms of proximity, density-based approaches can also be used for virtually any type of data. However, the meaning of density is less clear in a non-Euclidean data space.
- Proximity- and density-based anomaly detection techniques can often produce similar results, although there are significant differences between techniques that do not account for the variations in density throughout a data set or that use different proximity measures for the same data set.
- Model-based methods will often differ significantly from one another and from proximity and density-based approaches.

2. Consider a set of points that are uniformly distributed on the interval [0,1]. Is the statistical notion of an outlier as an *infrequently observed value* meaningful for this data?

- No. The traditional statistical notion of an outlier relies on the idea that an object with a relatively *low probability* is suspect. With a uniform distribution, no such distinction can be made.

3. Discuss techniques for combining multiple anomaly detection techniques to improve the identification of anomalous objects. Consider both supervised and unsupervised cases.

- In the supervised case, we could use ensemble classification techniques. In these approaches, the classification of an object is determined by combining the classifications of a number of classifiers, e.g., by voting.
- In the unsupervised approach, a voting approach could also be used. Note that this assumes that we have a binary assignment of an object as an anomaly. If we have anomaly scores, then the scores could be combined in some manner, e.g., an average or minimum, to yield an overall score.

4. Compare the following two measures of the extent to which an object belongs to a cluster: (1) distance of an object from the *centroid* of its closest cluster; and (2) the *silhouette coefficient*.

- The first measure is somewhat limited since it disregards that fact that the object may also be close to another cluster.
- The **silhouette coefficient** takes into account both the distance of an object to its cluster and its distance to other clusters. Thus, it can be more informative about how strongly an object belongs to the cluster to which it was assigned.

The End