Homework # 4

1. Attribute Types

Classify the following attributes as binary, discrete, or continuous. Further classify the attributes as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some of the cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

> **Example:** Age in years. **Answer: Discrete, quantitative, ratio**
>
> **Note:** J. Han's chapter on clustering gives a non-standard definition for interval and ratio scales, so please use the definition in the notes, i.e., *An Introduction to Cluster Analysis for Data Mining*. (In statistics and measurement theory, **an interval scale does not have a true 0, although it does have equal intervals, i.e., is a linear scale**. **A ratio scale has equal intervals and a true 0.** (The part about equal intervals was not mentioned in the notes.))

a) Year that an event happened, e.g., 1917, 1950, 2000.

b) Minnesota student ID number.

c) Temperature in degrees Kelvin.

d) Brightness as measured by a light meter.

e) Brightness as measured by people's judgments.

f) Grades on a scale of 1.0 = F, to 5.0 = A. (No pluses or minuses.)

g) Angles as measured in degrees between 0° and 360°.

h) Bronze, Silver and Gold Medals as awarded at the Olympics.

i) Time, in terms of AM or PM.

j) Number of patients in a hospital.

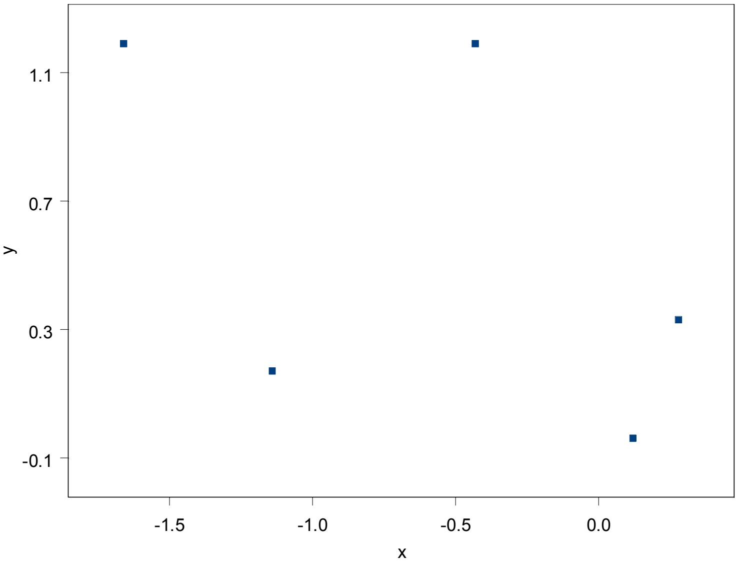
k) ISBN numbers for books. (Format of ISBN numbers is at http://www.isbn.spk-berlin.de/html/userman/usm4.htm)

l) Military rank.

2. Hierarchical Clustering

Five, two-dimensional data points are shown below with their distance matrix for , i.e., the symmetric matrix that gives the pairwise distance between any two points.

x	y
-0.43	1.19
-1.66	1.19
0.12	-0.04
0.28	0.33
-1.14	0.17



	1	2	3	4	5
1	0.00	1.23	1.34	1.12	1.24
2	1.23	0.00	2.16	2.12	1.14
3	1.35	2.16	0.00	0.40	1.28
4	1.12	2.12	0.40	0.00	1.42
5	1.24	1.14	1.28	1.43	0.00

Use the distance matrix to perform the following three types of hierarchical clustering: MIN, MAX, and Group Average. Show your results by drawing a dendogram. Note: the dendogram should clearly show the order in which the points are merged.

3. Understanding the behavior of K-means and MIN

Hierarchical clustering is sometimes used to generate K clusters, $K > 1$ by taking the clusters at the K^{th} level of the dendogram. (Root is at level 1.) By looking at the clusters produced in this way, we can evaluate the behavior of hierarchical clustering on different types of data and clusters, and also compare hierarchical approaches to K-means.

The following is a set of one-dimensional points: { 1, 2, 3, 4, 5, 7, 8 }.

a) For each of the following sets of initial centroids, create two clusters by assigning each point to the nearest centroid, and then calculate the total squared error for each set of two clusters. Show both the clusters and the total squared error for each set of centroids.

i) { 3, 7.5 }

ii) $\{ 2.5, 6\frac{2}{3} \}$.

b) Do both sets of centroids represent stable solutions, i.e., if the K-means algorithm was run on this set of points using the given centroids as the starting centroids, would there be any change in the clusters generated?

c) What are the two clusters produced by MIN?

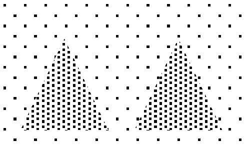
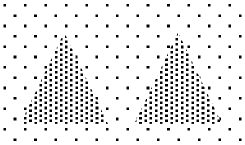
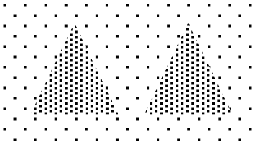
d) Which technique, K-means or MIN, seems to produce the "most natural" clustering in this situation?

e) What definition(s) of clustering does this "natural" clustering correspond to? (Well separated, center-based, contiguous, or density)

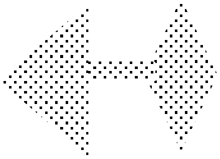
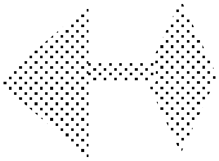
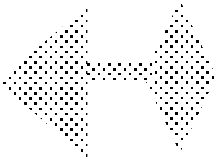
f) What well-known characteristic of the K-means algorithm explains the previous behavior?

4. Identify the clusters in the following diagram using the center-based, contiguous and density definitions. You should draw your answers on this sheet and hand it in. Also indicate the number of clusters for each case and give a brief indication of your reasoning.

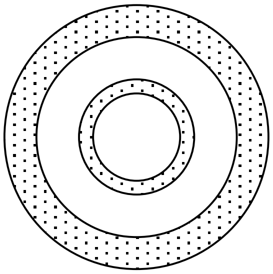
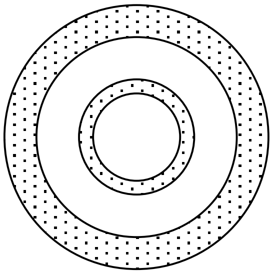
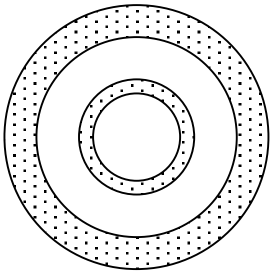
a) contiguous density center-based.

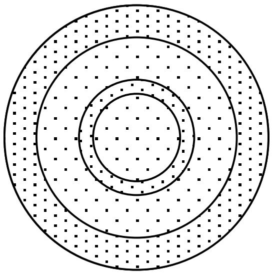
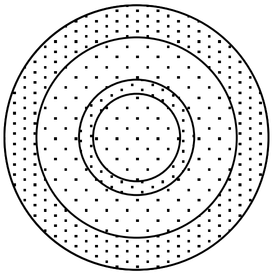
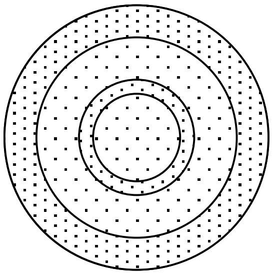
b) contiguous density center-based.

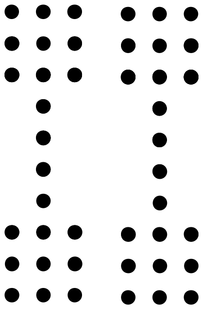
c) contiguous density center-based.

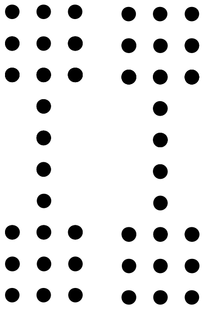




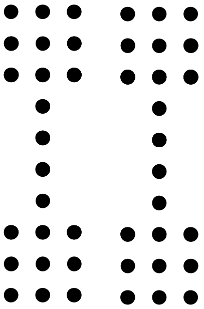
d) contiguous density center-based.

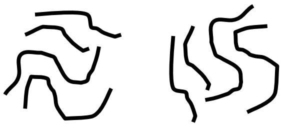
e) contiguous density center-based.

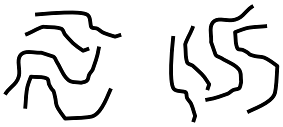


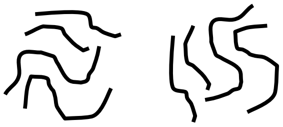




f) contiguous density center-based.

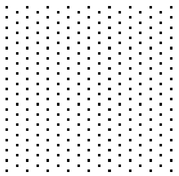




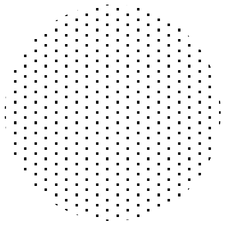


5. **For the following sets of two-dimensional points, provide a) a sketch of how they would be split into clusters by k-means for the given number of clusters and b) indicate approximately where the resulting centroids would be. Assume that we are using the squared error objective function. If you think there is more than one possible solution, then please indicate for each solution whether it is a global or local minimum.**

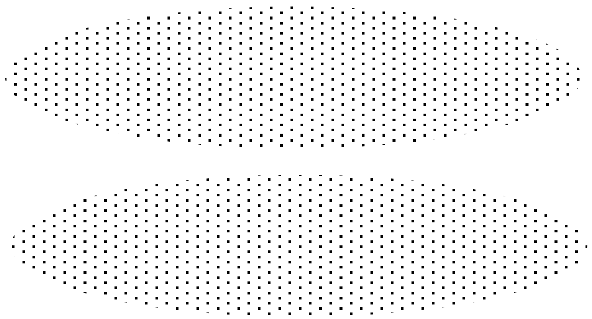
a) K = 2. Assuming that the points are uniformly distributed in the square, how many possible ways of partitioning the points into two clusters are there? What can you say about the positions of the resulting two centroids? (Again you don't need to provide exact centroid locations, just a qualitative description.)



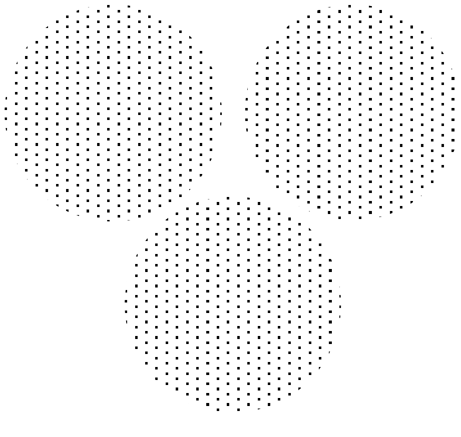
b) K = 3.



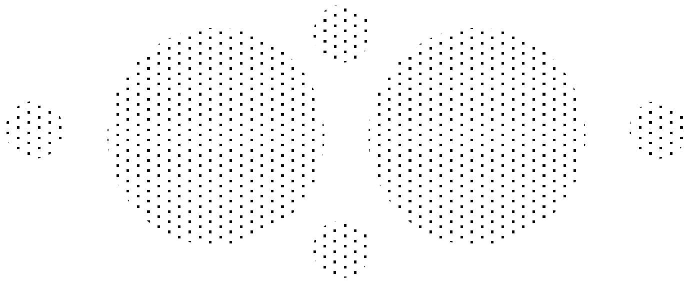
c) K = 3.



d) K = 2.



e) K = 2.



6. Limitations of clustering

Consider the following two diagrams:

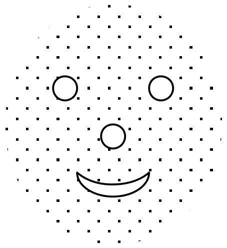


Figure 1

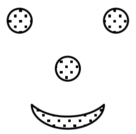


Figure 2

a) In which of the two figures will case will traditional clustring techniques, e.g., MIN, find the patterns represented by the nose, eyes and mouth?

b) What limititation does clustering have in detecting all the patterns formed by patterns of points?

7. Well-separated Clusters

Find all well separated clusters in the following set of points.