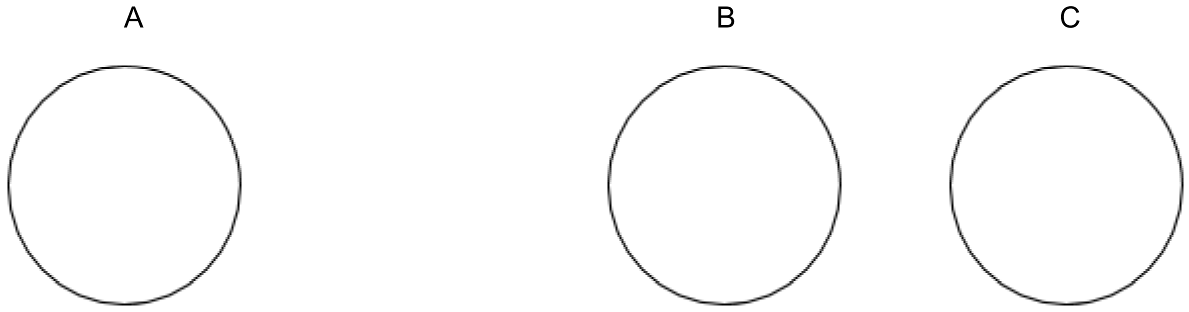
CSCI 5523 **HOMEWORK 6** **DUE: May 6, 2019**

Question 1:

In the figure below, assume that the leftmost circle (A) has 50,000 points and the other two circles (B and C) have 50 points each.



a. Describe the resulting cluster structure if k-means is used to cluster these points into 3 clusters.

b. Describe the resulting cluster structure if MAX is used to cluster these points into 3 clusters.

c. Given your answers for (i) and (ii), what can you say about the relative behavior of MAX and k-means?

Question 2.

Traditional K-means has a number of limitations, such as sensitivity to outliers and difficulty in handling clusters of different sizes and densities, or with non-globular shapes.
Comment on the ability of fuzzy c-means to handle these situations.

Question 3

For the figures below, please match the distance matrices (Fig. (a), (b) and (c)) with their corresponding datasets (Fig. (d), (e) and (f)).
Each dataset contains four clusters, and each cluster has 80 points. In the distance matrices, points are sorted according to the four cluster labels. Different darkness indicates differences in distance: black indicates the lowest distances and white indicates the highest distances.

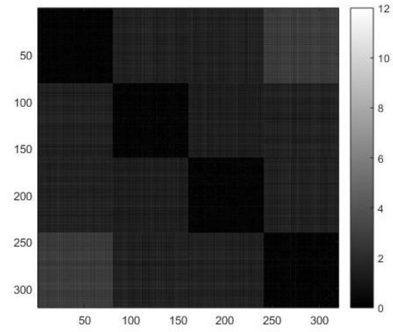


Fig. (a)

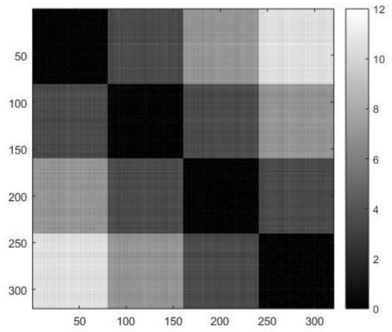


Fig. (b)

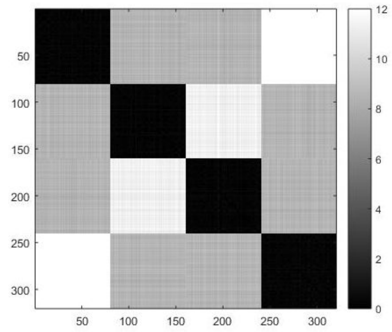


Fig. (c)

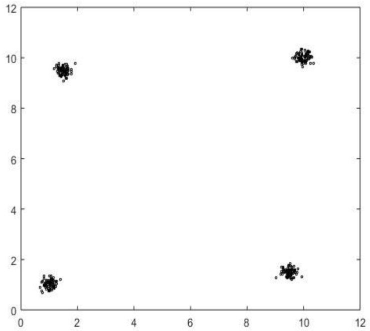


Fig. (d)

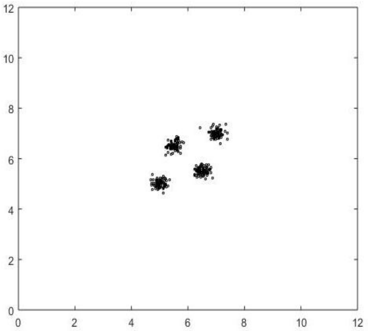


Fig. (e)

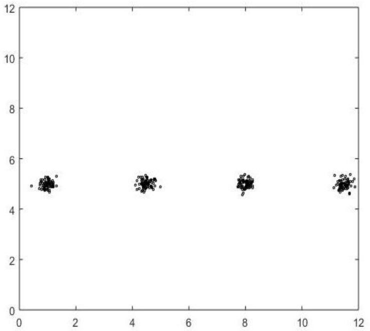
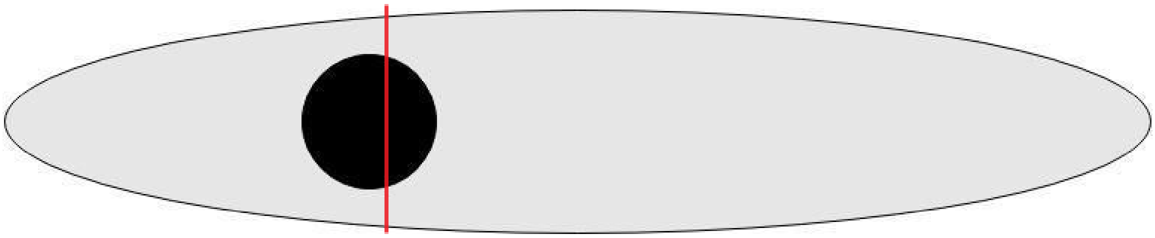


Fig. (f)

The distance matrix	Corresponding dataset ((d), (e) or (f))
Fig. (a)	
Fig. (b)	
Fig. (c)	

Question 4:

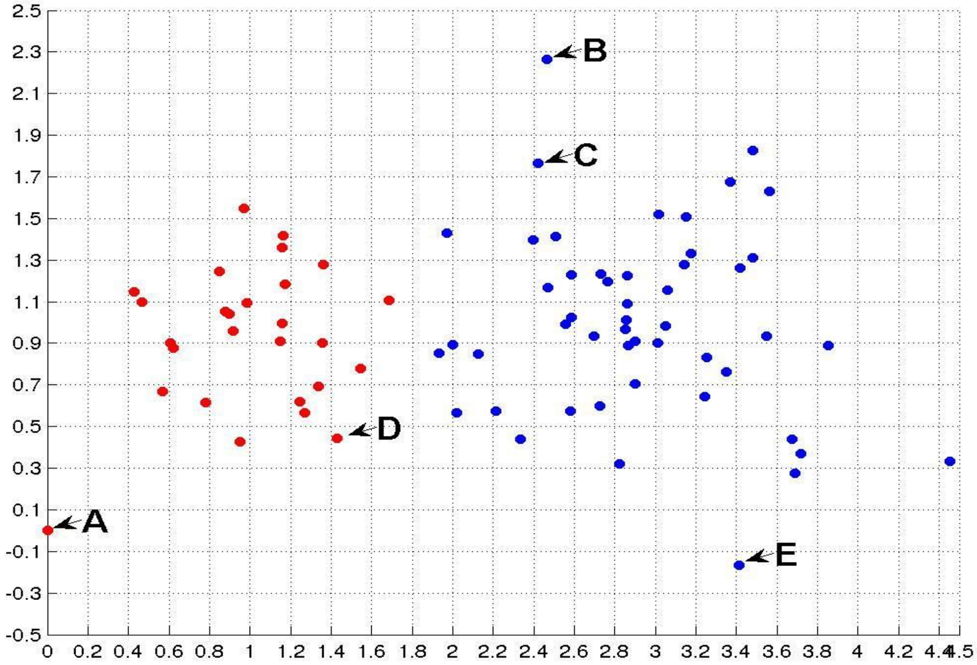
1) For the following set of two-dimensional points, draw a sketch of how they would be split into two clusters by K-means (when global minimum of SSE is achieved) and by Gaussian mixture model clustering. You can assume the density of points in the darker area is much higher than the density of points in the lighter area.



2) Name one other clustering method that might be able to accurately capture the two clusters.

Question 5

Mark whether the points A,B,C,D,E are border, core and noise point for a) eps=0.4 and b) eps=0.6 with mincout=4 for both cases. You don't need to compute the actual distance between the points, but should roughly estimate the neighborhood of each point to find the border, core and noise points.



Question 6

a. Give two advantages of statistical approaches to anomaly detection over cluster-based approaches.

b. Give two advantages of cluster-based approaches to anomaly detection over statistical approaches.

c. Why does high-dimensionality make density-based outlier detection particularly challenging?

d. The quality of cluster-based anomaly detection is heavily dependent on the quality of the clustering, which in turn depends on the inherent cluster structure present in the data. Does density-based anomaly detection also have this heavy dependence on the data having inherent cluster structure?

e. When the data has regions of differing density, the LOF anomaly detection algorithm is more effective than the k-nearest neighbor-based outlier detection algorithm.

Question 7

1. Give one advantage of statistical approaches over clustering clustering-based approaches and vice-versa.

2. Given a data set with a significant amount of noise and outliers, which technique would you prefer given a choice of either a clustering or density-based anomaly detection approach? Why? Assume the clustering approach is k-means and the density-based approach is LOF.

Practice Questions

We strongly encourage you to do questions:

- *Chapter 9 - questions:* ***3, 10, 11, 12, 17, 18, 20*** *(pages: 696-702)*
- *Chapter 10 - questions:* ***2, 3, 4, 7, 11*** *(pages: 809-814)*

from your book "Introduction to Data Mining, Second Edition, P.N. Tan, M. Steinbach, A. Karpatne, V. Kumar."

As well as:

Question 1.

Answer the following questions.

1) State two similarities and two differences between SNN density-based clustering and DENCLUE.

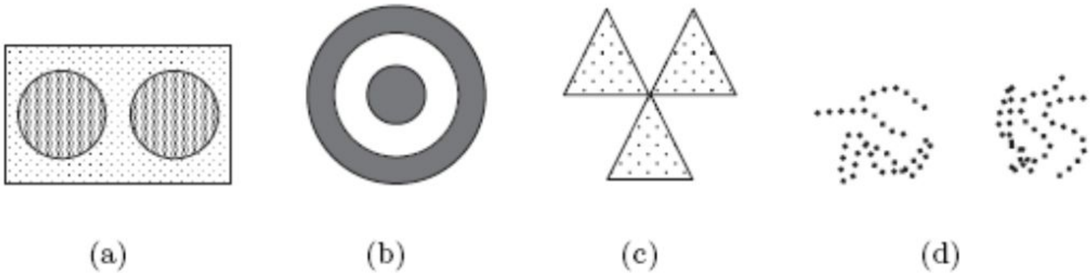
2) Explain two major advantages of Chameleon over hierarchical clustering using group average.

3) Whenever two clusters are merged together, the resulting SSE after merging is always greater than or equal to the SSE before merging. True or False? Explain briefly.

4) When applying DBscan to data sets that contain a large number of noise and outliers, it is better to use a large value for MinPts than a large value for EPS. True or False? Explain briefly.

Question 2:

Identify the clusters in the following Figure 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.



Question 3:

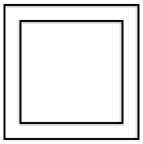
Following is the list of **all** frequent itemsets of size 3 in a given dataset:

{x, y, z}, { y, z, w}, {x, z, w}, {x, z, t}, {x,y,w}

a. Which of the following itemsets should be merged to generate a candidate itemset of size 4 according to the apriori $F_{k-1} \times F_{k-1}$ method given in the book? Also answer whether the candidate itemset would be pruned in the pruning step or not. (If the itemsets could be merged, list the candidate itemset and write down whether it will be pruned or not (Yes/No). If the itemsets could not be merged, write a **None** in the Candidate Itemset column and an **NA** (Not Applicable) in the Pruned Column.

Merging itemsets	Candidate Itemset	Pruned (Yes / No/NA)
{x, y, z}, { y, z, w}		
{x, z, w}, {x, z, t}		
{x,y,z}, {x, y, w}		

Question 4: Consider the *bond* measure for an itemset, *X* = {*A, B*}, containing two items, *A* and *B* given by



$P(A,B)/P(A^{\vee} B)$

where $P(A,B)$ is the support of the itemset and $P(A^{\vee}B)$ is the fraction of transactions that contain either A or B (or both).

a) What is the minimum value for *bond* and under what condition does *Bond(X)* attain its minimum value?

b) What is the maximum value for *bond* and under what condition does *Bond(X)* attain its maximum value?

c) How does Bond(X) behave when P(A) is increased while P(A,B) and P(B) remain unchanged?

Question 5: The interestingness measure *I* for a pair of items *(A,B)* is defined as:

$$I(A,B)=\frac{P(AB)}{P(A)P(B)}$$

Consider the following pairs of items: *(A,B)* and *(C,D)*, characterized by the following contingency tables:

	A	*Not A*
B	2	398
Not B	498	102

I(A,B)=.01

	C	*Not C*
D	240	60
Not D	160	540

I(C,D)=2

Which of these pairs of items: *(A,B)* or *(C,D)*, is more interesting (strongly associated)? Why?

Question 6.

(a) State two similarities and two differences between SNN density-based clustering and DENCLUE.

(b) Algorithms such as SNN produce incomplete clusterings (i.e., they do not cluster all points). For what kind of problems is this feature of SNN particularly useful?

(c) The objects of a clustering problem are stocks listed on the New York Stock Exchange. Each stock is represented by a time series of its daily closing price. Stocks with similar market behavior (i.e. similar time series) are put in the same group. Which type of clustering is more natural for this problem: fuzzy or crisp? Briefly justify your answer.

Question 7

To answer the three following true / false questions about how k-means operates, refer to figures (a), (b), and (c), below. Note that we are referring to the very basic k-means algorithm presented in class and not to any of it more sophisticated variants, such as bisecting k-means or k-means++.

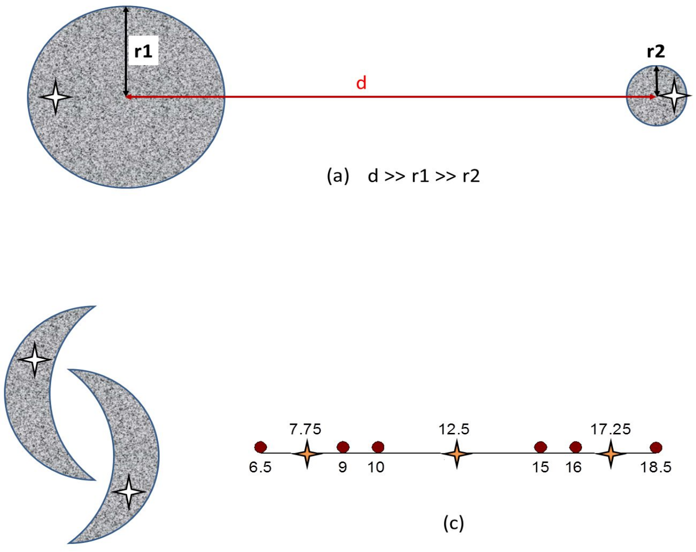
Note that for all three figures, the initial centroids are given by the symbol: ✦Initial point

For figures (a) and (b) assume the shaded areas represent points with the same uniform density. For Figure (c), the data points are given as red dots and their values are indicated under the dots. No explanation for your answer is necessary unless you feel there is some ambiguity in the figure or the question.

a) **True or False:** For Figure (a) and the given initial centroid: When the k-means algorithm completes, each shaded circle will have one cluster centroid at its center.

b) **True or False:** For Figure (b) and the given initial centroids: When the k-means algorithms completes, there will be one cluster centroid in the center of each of the two shaded regions and each of the two final clusters will consist only of points from one of the shaded regions. In other words, none of the two final clusters will have points from both shaded regions.

c) **True or False:** For Figure (c) and the given initial centroids, the final clustering for k-means contains an empty cluster.



(a) d >> r1 >> r2

(b)

(c)