CSCI 5523 **HOMEWORK 5** **DUE: April 15, 2019**

Question 1

For each of the following situations, decide what type of clustering should be used (hierarchical or partitional; exclusive, overlapping, or fuzzy; and complete or partial) to obtain the desired grouping. If you believe that data could be clustered using several different types of clustering, please state your assumptions.

In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1. weight sums to 1. Probabilistic.

Example: A nutritionist asks you various questions to assess your risk for diabetes. Based on this data, the nutritionist hopes to group people in three different categories: low, medium and high risk.

Partitional : non-overlapping subsets (clusters) such that each data object is in exactly one subset

Answer: Partitional, exclusive, complete

In non-exclusive clusterings, points may belong to multiple clusters.

(a) A supermarket manager wants to group all goods into several categories, each of which has multiple subcategories.

(b) Now the supermarket manager wants to group all goods according to brand.

(c) Grouping of students in a university based on the organization (department, college, institute etc.) they belong to. A student may belong to multiple organizations. Also, some students don't have declared majors and hence may not belong to any organization.

partial : only want to cluster some of the data

(d) You want to group all the videos on YouTube into several genre/topics, each of which can have several subtopics.

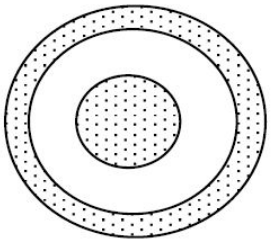
(e) You want to group all locations on Earth as to whether they belong to a tropical rainforest, a deciduous forest, or an evergreen forest. Here, each location corresponds to a grid cell of surface area 1km by 1km, and a location can have more than one variety of forests.

Question 2

For the following sets of two-dimensional points, (1) provide a sketch of how they would be split into clusters by K-means for the given number of clusters and (2) indicate approximately where the resulting centroids would be. Assume that we are using the squared error objective function and random initialization of centroids. If you think that there is more than one possible solution, then please indicate whether each solution is a global or local minimum. Darker areas indicate higher density.

(a) k = 3.

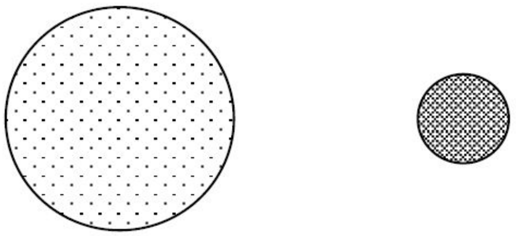
The center of a cluster is often a centroid, the average of all the points in the cluster



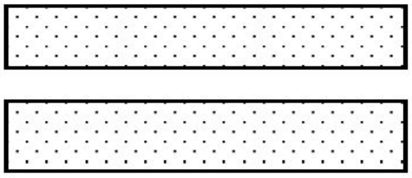
K-means
Each point is assigned to the cluster with the closest centroid

Text

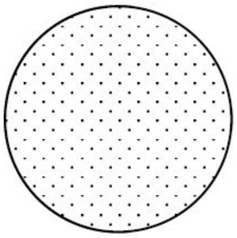
(b) k= 5 (Each circle has equal number of points)



(b) k = 2



(c) k = 2



Question 3

(a) Explain one advantage and one disadvantage of DBSCAN over the K-means clustering algorithm.

(b) List one similarity and one difference between EM and fuzzy k-means.

(c) Discuss MAX and MIN agglomerative clustering for the following cases:
- presence of noise
- different densities of clusters

(d) Discuss k-means and the Group Average agglomerative clustering for the following cases:
- presence of outliers
- categorical data

Agglomerative : Start with the points as individual clusters
At each step, merge the closest pair of clusters until only one cluster
(or k clusters) left

Question 4

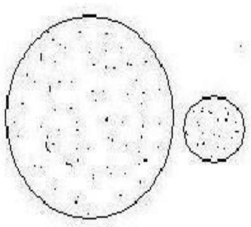
Use the similarity matrix in Table 1 to perform single and complete link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

	P1	P2	P3	P4	P5
P1	1.00	0.10	0.41	0.55	0.35
P2	0.10	1.00	0.64	0.47	0.98
P3	0.41	0.64	1.00	0.44	0.85
P4	0.55	0.47	0.44	1.00	0.76
P5	0.35	0.98	0.85	0.76	1.00

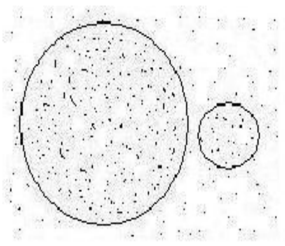
Table 1: Similarity matrix.

Question 5

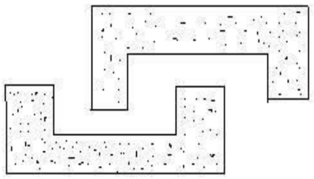
How will single-link, complete-link, and DBSCAN will perform for following cases? The points are evenly distributed for first three cases (a-c), while the last case (d) has one dense cluster with 50000 points and one relatively sparse cluster with 50 points with noise data points in between. Assume that the points inside a boundary are denser than the points outside the boundary, which represent the noise points (case b,d).



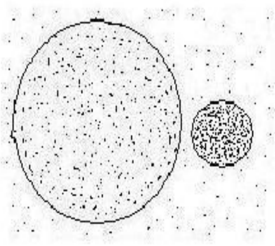
(a)



(b)



(c)



(d)

PRACTICE QUESTIONS

Question 1: 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

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

I(A,B)=.01

I(C,D)=2

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

Question 2: A pharmaceutical company has manufactured two drugs for the cure of a disease. A survey was conducted to compare the performances of the two drugs. The details of the survey are as follows –

- Drug A was given to 1000 patients, out of which 100 were in critical condition. The drug successfully cured 810 patients, out of which 10 were in critical condition.
- Drug B was given to 1500 patients, out of which 500 were in critical condition. The drug successfully cured 1000 patients, out of which 100 were in critical condition.

(i) Which drug has an overall higher success rate?
(ii) Which drug would you recommend to a patient to go for? Does your recommendation differ for a critical and a non-critical patient? Justify.

Question 3: The interestingness measure Odds Ratio (OR) for a pair of items (A,B) is defined as mentioned in Table 5.9of textbook on page 407 . Consider the following two pairs of items (A,B) and (C,D) characterized by their contingency tables:

	A	$\overline{A}$
B	20	400
$\overline{B}$	50	10

	C	$\overline{C}$
D	100	50
$\overline{D}$	40	200

Calculate the value of the *OR* measure for the pairs *(A,B)* and *(C,D).* Based on these values, comment on which of these pairs is more interesting? Why? **Hint:** The more extreme the value of OR(.) relative to 1 in either direction, the stronger the association.

Question 4.

Each of the following parts describes a collection of groups. Describe each of these groups in terms of the characteristics that we applied to sets of clusters. More specifically, classify the groups as to whether they are

- hierarchical or partitional
- overlapping or non-overlapping
- complete or incomplete

Each part should be labeled with four characteristics, e.g., partitional, overlapping and incomplete. Also, if you feel there may be some ambiguity about what characteristics a grouping has, provide a short justification of your answer.

a) Grouping of movie actors based on the genres of movies (comedy, drama, sci-fi etc.) they have acted in.

b) Grouping of students in a university based on the organization (department, college, institute etc.) they belong to. A student may belong to multiple organizations. Also, some students don't have declared majors and hence may not belong to any organization.

c) Grouping of all the students in the Computer Science department based on the letter grade they get in the data mining (CSci 5523) class.

Question 5

For each sequence shown in parts (i), (ii), and (iii) indicate whether it is a subsequence of the sequence <{b c}{a}{e d}>.

If a sequence is not a subsequence, give a one sentence explanation why this is so.

- <{a}{c}>
- <{a b c}>
- <{e}{d}>

Question 6

What are two key differences and two key similarities between clustering and association analysis?