CSci 5523 Homework 4 Solutions Due Nov 30, 2011

Only a subset of questions will be graded.

Q1. Sequential pattern mining **(15 points)**

Find all the frequent subsequences with support 50% given the sequence database shown in Table 1. Assume that there are no timing constraints imposed on the sequences. List the candidate and frequent subsequences at each level clearly.

Sensor	**Timestamp**	**Events**
S1	1 2 3 4	B A B D, E
S2	1 2 3 4	B A B, C A, D
S3	1 2 3 4	A, B C D C
S4	1 2 3	A, B C, D E
S5	1 2 3	C D C

Table 1. Example of event sequences generated by various sensors.

Answer: The events in the given data set are A, B, C, D and E. So, the candidate level-1 subsequences are : < {A} > (0.8), < {B} > (0.8), < {C} > (0.8), < {D} > (0.8), <{E}>(0.4)

The level-1 frequent subsequences are:
< {A} >, < {B} >, < {C} >, <{D} >.

The candidate level-2 subsequences are: (enumerated as discussed in the book on page 433)
< {A, B} > (0.4), < {A, C} > (0), < {A, D} > (0.2), < {B, C} > (0.2), < {B, D} > (0), < {C, D} >

(0.2),
< {A} {A} > (0.2), < {A} {B} > (0.4), < {B} {A} > (0.4), < {A} {C} > (0.6), < {C} {A} > (0.2), < {A} {D} > (0.8), < {D} {A} > (0),
< {B} {B} > (0.4), < {B} {C} > (0.6), < {C} {B} > (0), < {B} {D} > (0.8), < {D} {B} > (0),
< {C} {C} > (0.4), < {C} {D} > (0.6), < {D} {C} > (0.4),
< {D} {D} > (0)

The level-2 frequent subsequences are:
< {A} {C} >, < {A} {D} >, < {B} {C} >, < {B} {D} >, < {C} {D} >

The candidate level-3 subsequences are: < {A} {C} {D} > (0.4), < {A} {B} {C} > (0), < {A} {B} {D} > (0.4), < {B} {C} {D} > (0.4)
Since there are not candidate level-3 subsequences with support greater than 50%, there are no level-3 frequent subsequences.

Q2. (9 points)

For each of the described data sets, decide what type of clustering should be used (hierarchical or partitional, exclusive or overlapping or fuzzy, complete or partial).
If you find that data could be clustered using multiple types of clustering based on different assumptions, please provide explanations.

An example: Clustering library books based on their literary genre. The genre/topic can have several subtopics as well.
Answer: hierarchical, overlapping, complete

a) Proteins perform different biological functions which are organized into a hierarchical taxonomy(GO) defined by biologists. Some proteins can be multi-functional as well. You want to group them based on those functions. Some proteins may also have missing functional anotation.

Answer: Hierarchical, overlapping, partial

b) A nutritionist asks you several questions(e.g., your calorie intake, types of food you eat, your physical activity labels, and so on) to assess your risks for diabetes in three different groups: low, medium and high.

Answer: Partitional, exclusive, complete

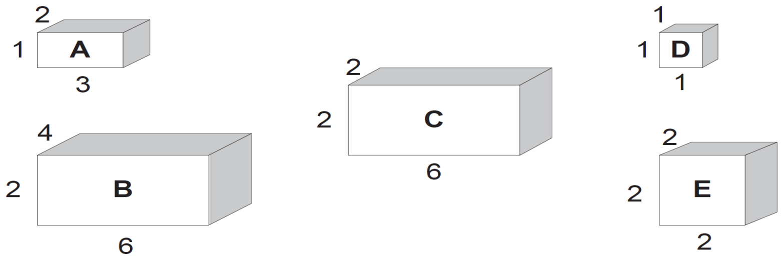
c) An international grad student is allowed to work on campus only at most for 20 hours. You want to assign each student to different job categories (e.g., TA, RA, other on-campus job, jobless). Hint: the sum of these categories should sum upto 20 hours.

Answer: Partitional, fuzzy, complete

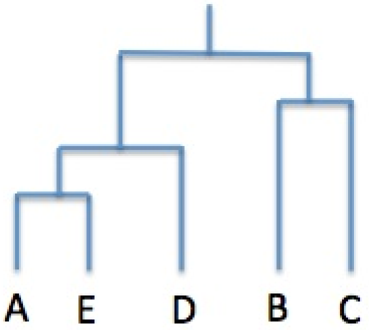
Q3. (15 points)

Consider the five objects (A,B,C,D, and E) shown in Figure 1. Each object has three features: length, width and height. For example, the features of object A are (3,2,1).

a) Suppose we apply the single link (MIN) algorithm to cluster the objects. Draw the dendrogram for the clusters assuming the similarity measure is Euclidean distance.



Answer:

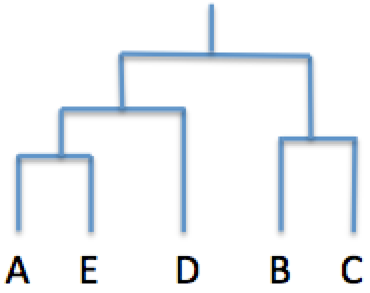


b) Repeat the question in part (a) assuming that the similarity measure is correlation. Are you able to compute similarity for every pair? Which objects are you having problems with and why?

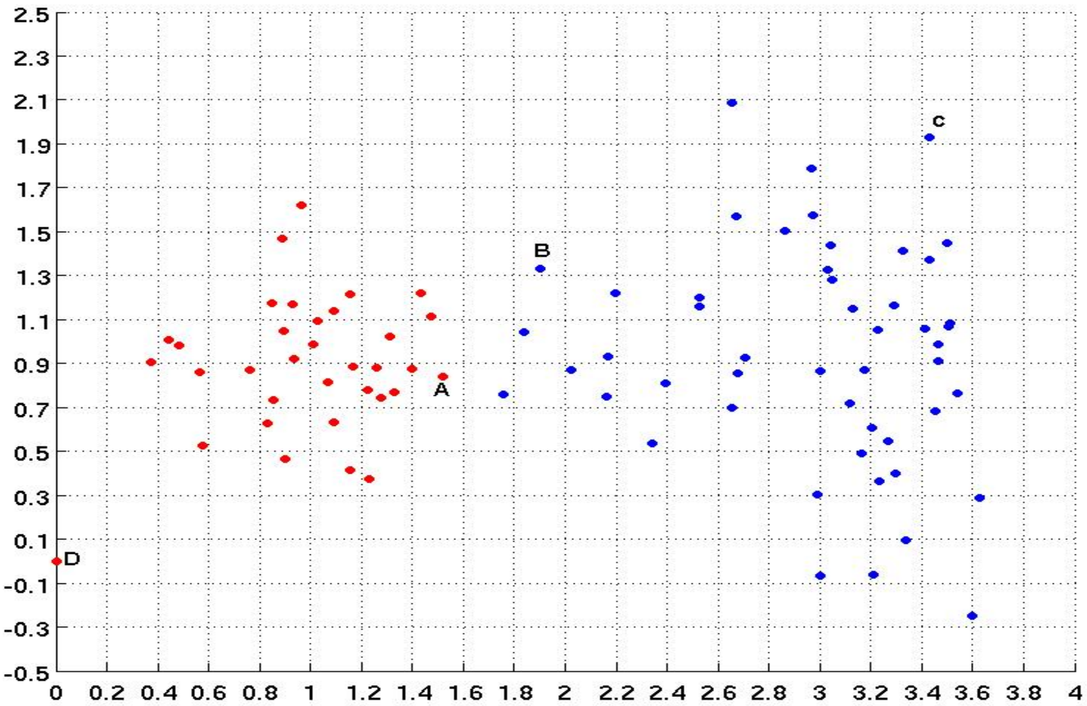
Answer: We cannot compute the correlation between objects D and E and any other object, since they are constant vectors, and the correlation is undefined. We therefore cannot perform the clustering.

c) Suppose we apply the complete link (MAX) algorithm to cluster the objects. Draw the dendrogram for the clusters assuming the similarity measure is Euclidean distance.

Answer:

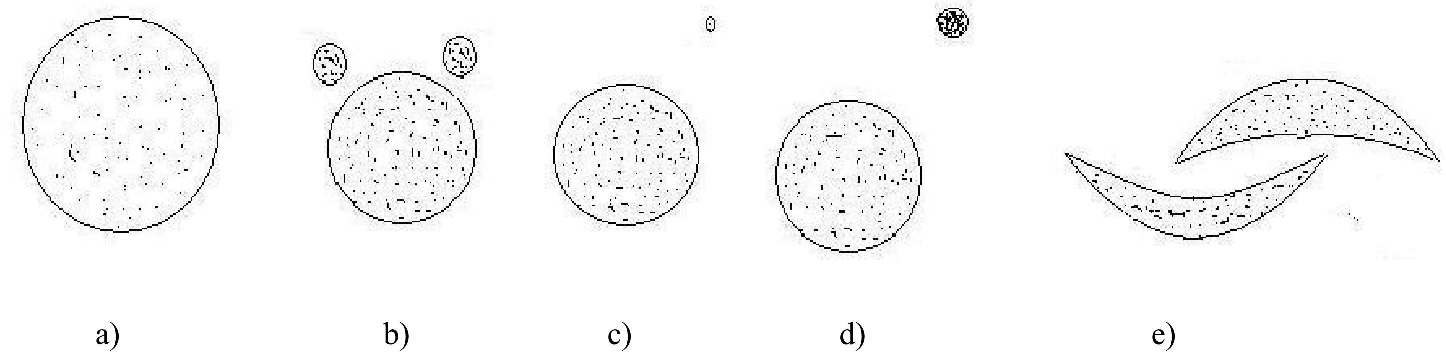


Q4. (10 points) Mark whether the points A,B,C,D are border, core and noise point for a) eps=0.4 and b) eps=0.6 with mincout<=4 for both cases. You don't need to compute the actual distance between the points, but should roughly guess the neighborhood of each point to find the border, core and noise points. How many clusters do you think DBSCAN will find for both cases?



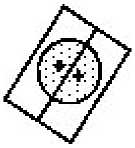
Answer:
a) eps = 0.4: core- A, border-B, noise-C,D
b) eps=0.6: core-A,B,C, border-D.
There will be 1 and 1 cluster found for eps=0.4 and 0.6 respectively.

Q5. (15 points) For the following sets of two-dimensional points, (1) draw 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. If you think that there is more than one possible solution, then please indicate whether each solution is a global or local minimum. Note that the label of each diagram in Figure 1 matches the corresponding part of this question, e.g., Figure 1(a) goes with part (a).Assume that the points in the cluster are of equal density unless mentioned explicitly.



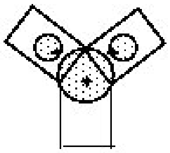
a) K=2. How many possible ways are there (in theory) to partition the points into two clusters? What can you say about the positions of the two centroids? (Again, you don't need to provide exact centroid locations, just a qualitative description.)

Answer:
In theory, there are an infinite number of ways to split the circle into two clusters - just take any line that bisects the circle. This line can make any angle between [0◦, 180◦] with the x axis. The centroids will lie on the perpendicular bisector of the line that splits the circle into two clusters and will be symmetrically positioned. All these solutions will have the same, globally minimal, error.

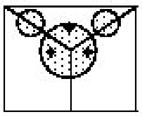


b) K= 3. Hint: Use the symmetry of the situation

Answer:
For the solution shown in the top figure, the two top clusters are enclosed in two boxes, while the third cluster is enclosed by the regions defined by a triangle and a rectangle. (The two smaller clusters in the drawing are supposed to be symmetrical.) I believe that the second solution—suggested by a student—is also possible, although it is a local minimum and might rarely be seen in practice for this configuration of points. Note that while the two pie shaped cuts out of the larger circle are shown as meeting at a point, this is not necessarily the case—it depends on the exact positions and sizes of the circles. There could be a gap between the two pie shaped cuts which is filled by the third (larger) cluster. (Imagine the small circles on opposite sides.) Or the boundary between the two pie shaped cuts could actually be a line segment.



Global minimum



Local minimum

c) K = 2. Here there are two outlier points with distance d from the center of the circle and d > 2R, where R is the radius of the circle.

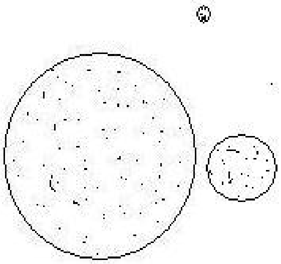
Answer: There will be two clusters on having centroid in the original circle and another will have the centroid in between the outlier and the circle.

d) K = 2. There are two circles. The bigger circle has only 100 points while the smaller circle is much denser with 100000 points. The distance between the two centers is greater than 2R, where R is the radius of the larger circle. Hint: the two centroids are more likely to be initialized with the points of denser circle.

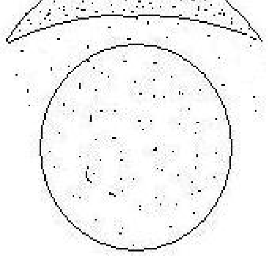
Answer: The denser circle will be divided into two cluster. One cluster will also contain the larger circle and its centroid will be slightly shifted towards the larger circle.

e) K = 2.
Answer: Since K-Mean only can find globular clusters, it will break each elliptical between the two clusters.

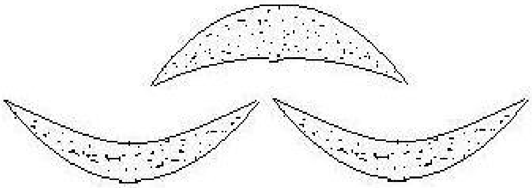
Q6. (12 poitns) How will single-link and complete-link will perform for following cases? The points are evenly distributed for first three cases(a-c), where the last case(d) has one dense cluster with 50000 points and one relatively sparse cluster with 50 points only without any noisy data points in between. Assume that the points inside a boundary are denser than the points outside the boundary, which represent the noise points(case b).



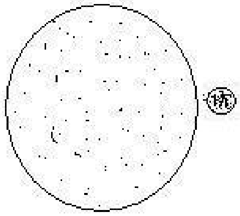
a)



b)



c)

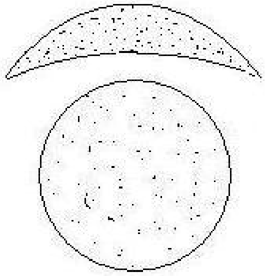


d)

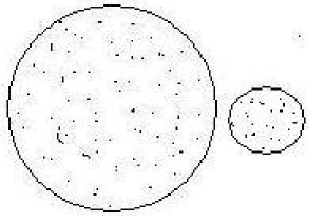
Answer:

a) Both of them will perform well, since they are less susceptible to outliers. MAX can break the big clusters into small ones for the favor of globular shape. b) MIN is sensitive to noise, MAX is good to remove the noise but can break the elliptical cluster into globular shapes. c) MIN will perform well, MAX will break also the clusters into globular shapes. d) None of them can handle the varying densities in clusters. MAX will be slightly better than MIN, since MAX will merge the two cluster in the last phase of dendrogram.

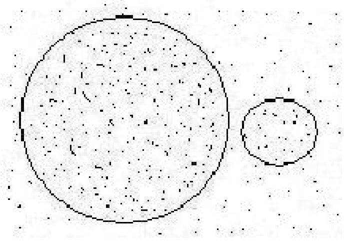
Q7. (12 points) How will DBSCAN perform for the following cases? Describe the effect of eps parameter on clustering them with a fixed mincount. Assume that the points inside a boundary are denser than the points outside the boundary, which represent the noise points(case c and d).



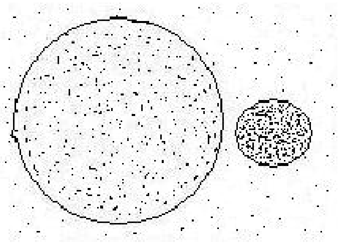
a)



b)



c)



d)

Answer:
For all cases except d), DBSCAN will perform well. If eps is too large, then the noise points will be also considered as the points of the cluster. If eps is too small, then only the denser cluster will be detected, but the other one will not be detected.