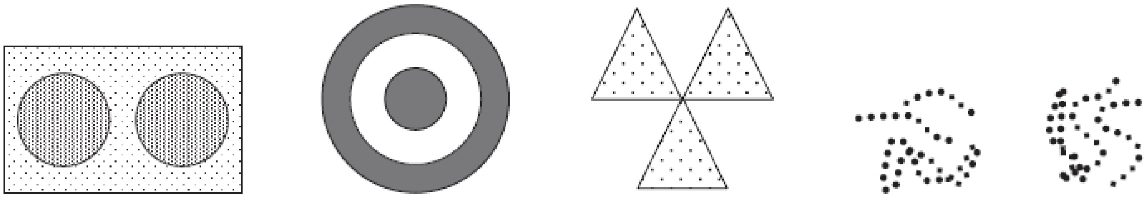
Birla Institute of Technology & Science - Pilani, Hyderabad Campus
Second Semester 2015-2016
CS F415 / CS C415 : Data Mining
Test 2

Type: Closed **Time: 60 mins** **Max Marks: 40** **Date: 09.04.2016**

All parts of the same question should be answered together.

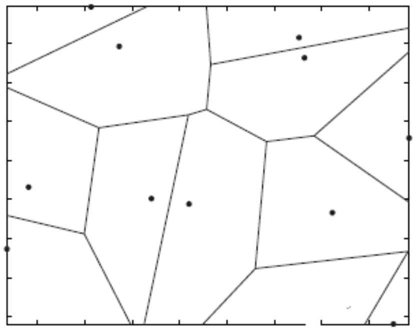
1. Describe the types of modifications necessary to adapt the frequent subgraph mining algorithm to handle: [6 Marks]
(a) Directed graphs
(b) Unlabeled graphs
(c) Acyclic graphs
(d) Disconnected graphs
For each type of graph given above, describe which step of the algorithm will be affected (candidate generation, candidate pruning, and support counting), and any further optimization that can help improve the efficiency of the algorithm.

2. Identify the clusters in the below mentioned 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. [6 Marks]



(a) (b) (c) (d)

3. 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 below). What is the relationship between Voronoi diagrams and K-means clusters? What do Voronoi diagrams tell us about the possible shapes of K-means clusters? [6 Marks]



Voronoi diagram

4. Suppose we find K clusters using Ward's method, bisecting K-means, and ordinary K-means. Which of these solutions represents a local or global minimum? Explain. [4 Marks]

5. Hierarchical clustering is sometimes used to generate K clusters, K > 1 by taking the clusters at the Kth level of the dendrogram. (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: {6, 12, 18, 24, 30, 42, 48}. [12 Marks]
(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. {18, 45}
ii. {15, 40}

(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 single link?

(d) Which technique, K-means or single link, seems to produce the "most natural" clustering in this situation? (For K-means, take the clustering with the lowest squared error.)

(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?

7. You are given two sets of 100 points that fall within the unit square. One set of points is arranged so that the points are uniformly spaced. The other set of points is generated from a uniform distribution over the unit square. [6 Marks]
(a) Is there a difference between the two sets of points?
(b) If so, which set of points will typically have a smaller SSE for K=10 clusters?
(c) What will be the behavior of DBSCAN on the uniform data set? The random data set?