Birla Institute of Technology & Science - Pilani, Hyderabad Campus
Second Semester 2014-2015
CS F415 / IS F415: Data Mining
Comprehensive Examination

Type: Closed **Time: 180 mins** **Max Marks: 100** **Date: 10.05.2015**

All parts of the same question should be answered together.

1.a. Robust data loading poses a challenge in database systems because the input data are often dirty. In many cases, an input record may have several missing values and some records could be *contaminated* (i.e., with some data values out of range or of a different data type than expected). Work out an automated *data cleaning and loading* algorithm so that the erroneous data will be marked and contaminated data will not be mistakenly inserted into the database during data loading. [6 Marks]

1.b. Suppose a group of 12 sales price records has been sorted as follows:
5; 10; 11; 13; 15; 35; 50; 55; 72; 92; 204; 215: Partition them into three bins by each of the following methods.
(a) equal-frequency partitioning
(b) equal-width partitioning
(c) clustering [6 Marks]

1.c. Association rule mining often generates a large number of rules. Discuss effective methods that can be used to reduce the number of rules generated while still preserving most of the interesting rules. [6 Marks]
Note: You are not supposed to increase / decrease the support and confidence values to reduce rules.

2.a. Suppose that the data mining task is to cluster the following eight points (with (*x; y*) representing location) into three clusters. *A*1(2*;* 10)*;A*2(2*;* 5)*;A*3(8*;* 4)*;B*1(5*;* 8)*;B*2(7*;* 5)*;B*3(6*;* 4)*;C*1(1*;* 2)*;C*2(4*;* 9)*:* [4 Marks]
The distance function is Euclidean distance. Suppose initially we assign *A*1, *B*1, and *C*1 as the center of each cluster, respectively. Use the *k-means* algorithm to show *only*
(a) The three cluster centers after the first round of execution and
(b) The final three clusters

2.b. Prove that in DBSACN, for a fixed MinPts value and two neighborhood thresholds €1 < €2, a cluster C with respect to €1 and MinPts must be a subset of a cluster C' with respect to €2 and MinPts. [8 Marks]

Note: A formal proof is only considered for evaluation.

2.c. Illustrate the strength and weakness of k-means algorithm in comparison with a hierarchical clustering schemes. [4 Marks]

2.d. 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. [6 Marks]

3.a. Explain the difference between likelihood and probability. Show mathematically that the maximum likelihood estimate of μ and σ *for a normal distribution* are the sample mean and the sample standard deviation, respectively [2 + 2 Marks]

3.b. We take a sample of adults and measure their heights. If we record the gender of each person, we can calculate the average height and the variance of the height, separately, for men and women. Suppose, however, that this information was not recorded. Would it be possible to still obtain this information? Explain. [4 Marks]

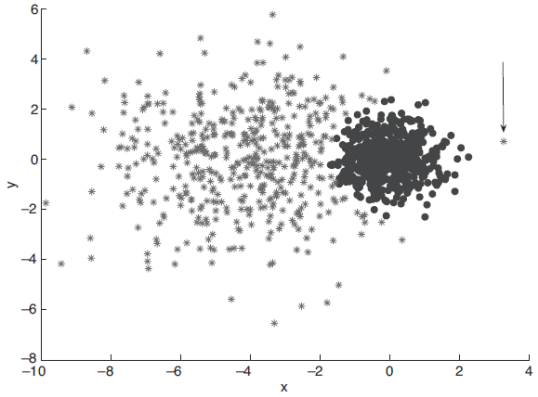
4.a. Show that the MST clustering technique produces the same clusters as single link. To avoid complications and special cases, assume that all the pairwise similarities are distinct. [8 Marks]

4.b. One way to sparsify a proximity matrix is the following: For each object (row in the matrix), set all entries to 0 except for those corresponding to the objects k-nearest neighbors. However, the sparsified proximity matrix is typically not symmetric. [3 + 5 Marks]

(a) If object a is among the k-nearest neighbors of object b, why is b not guaranteed to be among the k-nearest neighbors of a?

(b) Suggest at least two approaches that could be used to make the sparsified proximity matrix symmetric.

5.a.The following figure shows a clustering of a two-dimensional point data set with two clusters: The leftmost cluster, whose points are marked by asterisks, is somewhat diffuse, while the rightmost cluster, whose points are marked by circles, is compact. To the right of the compact cluster, there is a single point (marked by an arrow) that belongs to the diffuse cluster, whose center is farther away than that of the compact cluster. Explain why this is possible with EM clustering, but not K-means clustering. [8 Marks]



5.b.Consider a Gaussian mixture model in which the marginal distribution $p(\mathbf{z})$ for the latent variable is given by (1), and the conditional distribution $p(\mathbf{x}|\mathbf{z})$ for the observed variable is given by (2). Show that the marginal distribution $p(\mathbf{x})$, obtained by summing $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ over all possible values of $\mathbf{z}$, is a Gaussian mixture of the form (3).[8 Marks]

Note: The notation used here is same as that we discussed in class.

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}. \quad \text{----------(1)}$$

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \quad \text{……(2)}$$

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \quad \text{……..(3)}$$