Institute of Information Technology
Jahangirnagar University
Professional Masters in IT

1st Semester Final Examination — **Semester: Fall & Summer 2018**

Duration: 3 hours — **Full Marks: 60**

Course Code: IT-6307 — **Course Title: Data Mining & Knowledge Discovery**

Do not write anything on the question paper.
There are **7 (Seven)** questions. Answer any **5 (Five)** of them.
Figures in the right margin indicate marks.

1. a) What is Data Mining? Why is data mining important in our daily life? 4
   b) What do you mean by supervised and unsupervised classification? 4
   c) Why do we divide data in two parts before data mining starts? 4

2. a) Why do we apply aggregation on data? 4
   b) If you have missing data and noise exist in your data then what are the steps you should take? 4
   c) What are the advantages of tree based classification? 4

3. a) When do we need to use discretization and binarization? Explain with an example. 4
   b) How do you validate a classification model? What are the functions of ROC curve in validation? 4
   c) In a validation process, among 120 test data 40 data were classified as "Cheat=Yes" where among these 22 are actually "Cheat=Yes". Actual cases of "Cheat=No" is 58. Calculate different parameters related with validation and comment on the results. 4

4. a) Why and when do we use Gini coefficient or entropy? 4
   b) Suppose you have two variables (Gender and Refund) to select any one of them in constructing a tree. Under variable Gender, distribution of class variable are Cheat =5 and No Cheat =10 in one hand and in other hand it is 3, 12 whereas under variable Refund it is 3, 7 in one hand and 6 , 8 in other hand respectively. Which variable you should select? (Use Gini coefficient or entropy to give your answer.) 4
   c) Explain, how do you discretize a numeric attribute? i.e. Income 4

5. a) If you have a data set with class attribute and a new data without class attribute. You want to predict the value of class attribute of new data. Write the process of classifying this new data using decision tree classification (Hunt's Algorithm). 4
   b) How do you perform KNN? Write the limitations of KNN. 4
   c) How ANN classifier works? 4

6. a) What is ensemble method of classification? Explain with pictorial example. 4
   b) What are the different methods of calculating similarity and dissimilarity? 4
   c) Why and when do researchers like to use SVM classifier? 4

7. a) What do you mean by centeroid in k-means clustering 4
   b) How do you calculate distance between two clusters? 4
   c) Write two limitations of k-means clustering. How can we minimize these limitations? 4