Birla Institute of Technology & Science-Pilani, Hyderabad Campus
1st Semester 2011-2012
Data Mining (CS / IS C415)
Test I (Regular)
Type: Closed Book Time: 50 mins Wt.age: 20% Date: 15/9/2011

PART - I

Answer all the following questions using true/false or in one line. (5 Marks)

1. Can you call the following tasks as data mining answer using [yes/no]?

a. Sorting the student according to ID numbers.
b. Monitoring the heart rate of the customers to find abnormalities.
c. Dividing the customers according to their profitability.
d. Predict the profitability of a new customer.
e. The bronze, silver or gold medal awarded at the Olympics is what kind of attribute?

2. Answer in one line (5 Marks)

a. If the mean is larger than the median then this might be an indication that the data is what?
b. If I have 100 values in my data and I add 5.0 to all of the values, then how will this change the median?
c. If I have 100 values in my data and I add 5.0 to all of the values, then how will this change the standard deviation?
d. What is the complexity of the KNN algorithm as a function of the number of elements in the training set(q) and the number of elements (n) to be classified ?
e. Consider a supervised learning problem in which the training examples are points in 2-dimensional space. The positive examples are (1*;* 1) and (-1;-1). The negative examples are (1;-1) and (-1*;* 1). Are the positive examples linearly separable from the negative examples in the original space?

PART - II

1. Suppose a student collected the price and weight of 20 products in a shop with the following result **(6*2=12 Marks)**

Price	5.89	49.59	59.98	159	17.99	56.99	62.75	102.19	31	125.5
Weight	1.4	1.5	2.2	2.7	3.2	3.9	4.1	4.1	4.6	4.8
Price	4.5	84	52.9	61	33.5	328	128	142.19	229	189.4
Weight	4.9	5.3	5.5	5.8	6.2	8.9	11.6	18.0	22.9	38.2

a. Calculate the mean, Q1, median, Q3 of **price**
b. Draw the boxplots for **weight**;
c. Normalize the **two variables** based on the min-max normalization (min = 1, max = 10);
d. Normalize the **two variables** based on the z-score normalization;
e. Take the **price** of the above 20 products, partition them into four bins by equal-width partitioning
f. Take the **price** of the above 20 products, partition them into four bins by equal-depth (equal-frequency) partitioning

2. For this question, you're going to answer a couple questions regarding the dataset shown below. You'll be trying to determine whether Andrew finds a particular type of food appealing based on the food's temperature, taste, and size. **(3+3+4)**

Appealing	Temperature	Taste	Size
No	Hot	Salty	Small
No	Cold	Sweet	Large
No	Cold	Sweet	Large
Yes	Cold	Sour	Small
Yes	H	Sour	Small
No	H	Salty	Large
Yes	H	Sour	Large
Yes	Cold	Sweet	Small
Yes	Cold	Sweet	Small
No	H	Salty	Large

a) What is the initial entropy of Appealing?
b) Assume that Taste is chosen for the root of the decision tree. What is the information gain associated with this attribute?
c) Draw the full decision tree learned for this data.

3. A trainee manager wondered whether the length of time his trainees revised for an examination had any effect on the marks they scored in the examination. Before the exam, he asked a random sample of them to honestly estimate how long, to the nearest hour, they had spent revising. After the examination he investigated the relationship between the two variables. **(3 Marks)**

Trainee	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**
Revision time	4	9	10	14	4	7	12	22	1	17
Exam mark	31	58	65	73	37	44	60	91	21	84

If the regression equation derived is : **y = 21.7 + 3.47x**
a) Predict the examination mark for a trainee who revises for 15 hours.
b) Predict the examination mark for a trainee who revises for 35 hours.
c) Do you have any reservations about your answer to (b)?

4. Suppose we are given the following dataset, where A,B,C are input binary random variables, and y is a binary output whose value we want to predict. **(3+2 Marks)**

A	**B**	**C**	**y**
0	0	1	0
0	1	0	0
1	1	0	0
0	0	1	1
1	1	1	1
1	0	0	1
1	1	0	1

a) How would a naive Bayes classifier predict y given this input: A = 0,B = 0,C = 1. Assume that in case of a tie the classifier always prefers to predict 0 for y.
b) Suppose you know for fact that A,B,C are independent random variables. In this case is it possible for any other classifier (e.g., a decision tree or a neural net) to do better than a naive Bayes classifier? If so why?