Dr. Sakshi Babbar

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

END SEMESTER EXAMINATION-2015

B.Tech VIth Semester

COURSE CODE: 10B22CI622 MAX. MARKS: 45

COURSE NAME: Data Mining

COURSE CREDITS: 4 MAX. TIME: 3 HRS

Note: *All questions are compulsory. Carrying of mobile phone during examinations will be treated as case of unfair means.*

Section A **(9 x 1= 9)**

1. Suppose a classification models gives TP=70 and TN=4. Let there are 300 data points presents in training set. What is the accuracy of rule?
2. What are the methods to overcome model underfitting problem in classification?
3. How will you measure impurity of a discrete node?
4. Let A and B be two nominal attributes. Suppose A= $\{a_1, a_2 \ldots, a_c\}$ and B=$\{b_1, b_2, \ldots, b_k\}$. How will you compute correlation between attributes A and B?
5. What is the complexity of K-means clustering algorithm?
6. How will you measure goodness of split?
7. Consider a rule A,B →C. How will you measure coverage and Accuracy of rule?
8. What does joint probability distribution in Bayesian network tell? How can it be computed?
9. Explain difference between mutually exclusive and exhaustive rule?

Section B **(5 x 3= 15)**

1. (a) Prove that if a rule X → Y-X does not stratify the confidence threshold then any rule X'→Y- X', where X' is a subset of X, must not satisfy the confidence threshold as well

   (b) Consider a training set contains 160 positive examples and 230 negative examples. Suppose we are given the following two candidate rules:

   R1: covers 50 positive examples and 11 negative examples

   R2: covers 140 positive examples and 2 negative examples

Evaluate R1 and R2 using likelihood ratio statistics to identify which rule is better for said training set.

2. Write an algorithm for direct method for rule extraction in rule based classifier. Derive complexity of algorithm.
3. Construct FP tree for set of transactions in Table 1. Consider minsupp=3. Discover frequent 2-itemsets from constructed FP tree.

Section C (7 x 3= 21)

1. Suppose we have 4 objects in training set and each object have 2 dimensions as shown in Table 2. Apply K means clustering algorithm to cluster data points in two groups. Consider A and B as initial seeds.
2. Consider data set in Table 3 for binary class problem. Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?
3. For training set in Table 4, predict class for test set ={Color= yellow, Type = SUV, Origin=Domestic) using Naive Bayes classifier.

Id	Items
1	b,c,e,g
2	a,c,d,e
3	b,c,d
4	b,,d
5	b,c
6	a,b,d
7	b,d,e
8	a,f,d
9	a,b,d
10	c,d,f

Table 1

Weight	pH
1	2
2	2
4	3
5	4

Table 2

A	B	Class
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

Table 3

Color	Type	Origin	Stolen
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Domestic	Yes
Yellow	Sports	Domestic	No
Yellow	Sports	Imported	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Red	SUV	Imported	No
Red	Sports	Imported	Yes

Table 4