B.Sc. SWE 5th Semester **27 February 2020 (Morning)**

ISLAMIC UNIVERSITY OF TECHNOLOGY (IUT)

ORGANISATION OF ISLAMIC COOPERATION (OIC)

Department of Computer Science and Engineering (CSE)

MID SEMESTER EXAMINATION **WINTER SEMESTER, 2019-2020**

DURATION: 1 Hour 30 Minutes **FULL MARKS: 75**

CSE 4553: Machine Learning

Programmable calculators are not allowed. Do not write anything on the question paper.

There are **4 (four)** questions. Answer any **3 (Three)** of them.

Figures in the right margin indicate marks.

1. a) Write down the Bayes theorem with explanation. Discuss the use of Naive Bayes classifier in real world applications. 4+3

b) Consider the dataset given in Table 1 of a credit card promotion database. The credit card company has authorized a new life insurance promotion similar to the existing one. We are interested in building a classification data mining model for deciding whether to send the customer promotional material. 12

Table 1: Dataset of credit promotion

Customer ID	Magazine Promotion	Watch Promotion	Credit Card Insurance	Sex	Life Insurance Promotion
1	Y	N	N	M	N
2	Y	Y	Y	F	Y
3	N	N	N	M	N
4	Y	Y	Y	M	Y
5	Y	N	N	F	Y
6	N	N	N	F	N
7	Y	Y	Y	M	Y
8	N	N	N	M	N
9	Y	Y	Y	M	N
10	N	Y	N	F	Y

Use the Naive Bayes classifier to determine the value of Life Insurance Promotion for the following instance:

Magazine Promotion = Y, Watch Promotion = Y, Credit Card Insurance = N, Sex = F, Life Insurance Promotion = ?

c) You are given a data set of 10,000 students with their sex, height, and hair color. You are trying to build a classifier to predict the sex of a student, so you randomly split the data into a training set and a testing set. Here are the specifications of the data set: 3+3

- sex ϵ { male, female}
- height ϵ [0,300] centimeters
- hair ϵ {brown, black, blond, red, green}

Under the assumptions necessary for Naive Bayes answer each question with T (True) or F (False) and provide a justification of your answer:

i. As height is a continuous valued variable, Naive Bayes is not appropriate since it cannot handle continuous valued variables.

ii. P(height, hair|sex) = P(height|sex) P(hair|sex).

2. a) Suppose you want to build a decision tree for a problem. In the dataset, there are two classes, with 150 examples in the '+' class and 50 examples in the '−' class.

   i. What is the entropy of the class variable? 2

   ii. For this data, suppose the Color attribute takes on one of 3 values (red, green, and blue), and the split into the two classes across red/green/blue is '+' : (120/10/20) and '−': (0/10/40). Write down an expression for the class entropy in the subset containing all green examples. Is this entropy greater or less than the entropy in the previous question? 4

   iii. Is Color a good attribute to add to the tree? Explain your answer 2

   iv. What is the information gain for a particular attribute if every value of the attribute has the same ratio between the number of + examples and the total number of examples? 4

b) Imagine you grow a very large, complex decision tree from a training set which contains many features (attributes). 2×4

   i. You find that the training set accuracy for your tree is very high, but the test set accuracy (as measured on held out data) is very low. Explain why the accuracy on the training data could be so much higher than on the test data.

   ii. You decide to use pruning to create a series of successively smaller subtrees from your full tree (e.g. χ-squared pruning, or truncating the tree at smaller depths). As you prune more and more of the tree, the test accuracy goes up at first, but then starts to drop again after extensive pruning. Explain the rise and then fall of test accuracy

c) Discuss the advantages and disadvantages of using Decision Tree classifier. 5

3. a) Explain the principle of the gradient descent. Consider a linear regression problem $y = w_1x + w_0$, with a training set having m examples $(x_1, y_1), \ldots, (x_m, y_m)$. Suppose that we wish to minimize squared error (loss function) given by: 5+5

$$Loss = \frac{1}{2m}\sum_{i=1}^{m}(y_i - w_1x - w_0)^2$$

Derive a batch gradient descent algorithm that minimizes the loss function.

b) i. What is the difference between Linear and Logistic regression? 4×2

   ii. Suppose you are given a dataset of cellular images from patients with and without cancer. If you are required to train a classifier that predicts the probability that the patient has cancer, would you prefer to use Decision trees over logistic regression? If not why?

   iii. Consider a logistic regression model with weights, $\beta = (-ln(4), ln(2), -ln(3))$. A given document has feature vector $X = (1, 1, 1)$. What is the probability that the document is about sports?

   iv. Suppose you train a logistic classifier where the hypothesis is

$$h_\theta(x) = g(\theta_0 + \theta_1x_1 + \theta_2x_2)$$

If $\theta_0 = 6$, $\theta_1 = -1, \theta_2 = 0$, draw the figure of the decision boundary found by this classifier?

c) Consider a (binary) linear classifier model: y= g(w^{T}x), where input x is a d-vector and w is the parameters to learn. Suppose we want to use such binary classifier models to classify k classes. Explain in detail how this can be done. 3

d) Is it possible to get multiple local optimum solutions if we solve a linear regression problem by minimizing the sum of squared errors using gradient descent? Give proper explanation. 4

4. a) Determine which is the best approach for each problem (i-v) and why? 5

- Supervised learning – classification
- Supervised learning – regression
- Unsupervised learning - clustering

i. A robot is learning to sort garbage using visual identification. It sits all day picking out recyclable items from garbage as it passes on a conveyor belt. It places items such as glass, plastic and metal into 12 bins. Each item is labeled with an identification number on a sticker

ii. A friend invites you to his party where you meet totally strangers. Now you will classify them in the basis of gender, age group, dressing, educational qualification or whatever way you would like.

iii. Suppose you have never seen a Cricket match before and by chance watch a video on internet, now you can classify players on the basis of different criterion: Players wearing same sort of kits are in one class, Players of one style are in one class (batsmen, bowler, fielders), or on the basis of playing hand (RH vs LH)

iv. Consider the problem of predicting the marks of a student based on the number of hours he/she put for the preparation

v. Determine whether a credit card transaction is valid or fraudulent.

b) Consider a 2-class classification problem where the number of data points (we also call them examples) in class 0 is 990 and number of data points in class 1 is 10. Suppose the classification model (or algorithm) predicts everything to be class 0. Which of the following metrics correctly measures the performance of the model: Accuracy, Precision, Recall, False positive rate, False negative rate? Justify your answer. 5

c) What are the benefit of feature scaling? Suppose students have taken some class, and the class had a midterm exam and a final exam. You have collected a dataset of their scores on the two exams, which is as follows: 3+6

Mid I (X1)	Mid II (X2)	Final
89	7921	96
72	5184	74
94	8836	87
69	4761	78

You would like to use polynomial regression to predict a student's final exam score from their midterm exam score. Before that you plan to use feature scaling so that each feature in the data have zero-mean and unit-variance. Which scaling technique you should apply and how? What will be the normalized value of feature X_2 ?

d) Write down the differences among hold out, k-fold cross-validation and leave-one-out cross-validation? 6