University of Asia Pacific

Department of Computer Science and Engineering

Mid-Semester Examination Spring-2020

Program: B.Sc. in Computer Science and Engineering

Course Title: Machine Learning **Course No.:** CSE 427 **Credit:** 3.00

Time: 1.00 Hour. **Full Mark:** 60

Instruction(s): Answer any three questions including 1 and 2.

1. a. Suppose we have a dataset as follows— **[12]**

x	*y*
4	10
6	16
8	19

We want to apply linear regression to predict the value of y. Our hypothesis function is:

$$h(\theta) = \theta_0 + \theta_1 x$$

At initial step, initialize the values of θ_0 and θ_1 as the last two digits of your ID. For Example,

If your ID is 113026, then $\theta_0 = 2$ and $\theta_1 = 6$.
If your ID is 113007, then $\theta_0 = 0$ and $\theta_1 = 7$. etc.

Here, learning rate, alpha = 0.1

Now, what will be values of θ_0 and θ_1 after updating them only once using gradient decent? (Your task is to calculate the updated values of θ_0 and θ_1 after one iteration)

b. "Logistic regression is not a regression algorithm" – do you agree with this statement. Explain why or why not. **[8]**

2. a. Suppose we have a dataset as follows— **[12]**

x_1	x_2	x_3
a	b	c
3	10	150
6	50	540

Where, a = your ID mod 5
b= your ID mod 7
c = your ID mod 9

For example, if your ID is 113026,

a = 113026 mod 5 = 1

b = 113026 mod 7 = 4

c = 113026 mod 9 = 4

Now, normalize this dataset.

b. "By k-fold cross validation we can get to a balance point between overfitting and underfitting" – do you agree with this statement? Explain why or why not. **[8]**

3. a. Suppose, you have given the following data where *x* and *y* are the two input variables that corresponds the coordinate points and *Class* is the dependent variable— **[12]**

x	*y*	*Class*
-1	1	-
0	1	+
0	2	-
1	-1	-
1	0	+
1	2	+
2	2	-
2	3	+

i. Suppose, you want to predict the class of new data point (x, y) using Euclidian distance in 3-NN. In which class this data point belongs to?

ii. You now want to use 7-NN instead of 3-KNN which of the following (x, y) will belong to?

The new data point (x, y) is the last two (02) digits of your ID. That is, if your ID is 113026, then the new data point will be (2, 6). Again, if your ID is 113007, then the new data point will be (0, 7) etc.

b. Discuss the terminating condition of k-means clustering, i.e. when we stop the iteration in k-means clustering. **[8]**

Or,

4. a. Consider the following dataset of genetic mutations— **[12]**

Mutation Rate	*New species emergence*
S	N
M	Y
M	N
M	Y
L	Y
M	N
S	Y
S	N
L	Y
L	Y
M	N

Here, first column represents the mutation rate and second column represents new species emergence.

In first attribute, S means small mutation rate, M means medium mutation rate and L means large mutation rate.

In second attribute, N means NO and Y means YES.

Now, determine the probability of new species emergence is 'YES' given that the mutation rate is 'medium' using Naïve Bayes Algorithm.

b. Suppose, there is a dataset which has some positive labels and negative labels. The number of positive labels is your ID mod 4 and the number of negative labels is your ID mod 7. Now determine the entropy for that dataset. **[8]**

For example, if your ID is 113058, then the number of positive labels is 113058 mod 4 = 2 and the number of negative labels is 113058 mod 7 = 1.