Department of Computer Science and Engineering
University of Barishal

Final Examination-2020

Course Title: Machine Learning and Data Mining
Course Code: CSE-4213
4th Year 2nd Semester
Admission Session: 2016-17

Time: 3 Hours **Marks:** 60

Answer any five Questions from the followings.

1. a) What is machine learning? Write down the relation between artificial intelligence (AI) and Machine Learning (ML)? [2.5]

b) What is the difference between supervised and unsupervised machine learning? Give example. [3]

c) How can you define deep learning, and how does it contrast with other machine learning algorithms? [2.5]

d) Solve the equations given below using metrics. [4]

$$3x + 4y + 5z = 36$$
$$2x + 5y + 7z = 33$$
$$x + y + z = 6$$

2. a) How can you convert the non-linear decision boundary to linear? Write with an example. [2]

b) Discuss the Naïve Bayes' classification algorithm. Consider a table below, the training set of weather (in the morning) and the corresponding target variable 'Heavy rain' (i.e., the possibilities of heavy rain in the day). [5]

Weather	Heavy rain	Weather	Heavy rain
Sunny	NO	Overcast	Yes
Overcast	NO	Dark cloudy	Yes
Dark cloudy	Yes	Dark cloudy	No
Overcast	NO	Overcast	Yes
Dark cloudy	Yes	Dark cloudy	Yes
Sunny	NO	Sunny	No
Overcast	Yes	Dark cloudy	Yes

There will heavy rain if weather is overcast. Is this statement correct? Justify your answer

c) What is information gain? Consider the example in the following table where variable-1 and variable-2 are used to determine whether to continue with the experiment or to stop the experiment. [5]

Variable-1	Variable-2	Outcome
3	5	Stop
7	6	Continue
3	3	Stop
4	8	Continue
3	9	Continue
6	5	Stop
5	8	Continue
6	4	Continue

Determine the first decision based on the information gain.

3. a) What do you mean by regression? When should multiple regression analysis be used? [3]

b) How do you perform a linear regression? [3]

c) The sales of a company (in million dollars) for each year are shown in the table below. [6]

x (year)	0	1	2	3	4
y (sales)	12	19	29	37	45

i) Find the least square regression line y = a x + b.
ii) Use the least squares regression line as a model to estimate the sales of the company in year 7.

4. Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters:
A1 = (2,10), A2 = (2,5), A3 = (8,4), A4 = (5,8), A5 = (7,5), A6 = (6,4), A7 = (1,2), A8 = (4,9)

The distance matrix based on the Euclidean distance is given below:

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Suppose that the initial seeds (centers of each cluster) are A1, A4 and A7. Run the k-means algorithm for 1 epoch only. At the end of this epoch show:

a) The new clusters (i.e., the examples belonging to each cluster) [4]
b) The centers of the new clusters [3]
c) Draw a 10 by 10 space with all the 8 points and show the clusters after the first epoch and the new centroids. [5]

5. a) What is the basic difference of Recurrent Neural Network (RNN) from Artificial Neural Network? Give some examples of RNN application. [4]
b) What is the vanishing Gradient Problem of RNN? [3]
c) Shortly describe Long-Term Memory (LSTM) networks and explain how it can avoid long term dependency problem. [5]

6. a) What is the basic principle of a Support Vector Machine? [3]
b) Why would you use the Kernel Trick in the Support Vector Machine? Explain with example. [3]
c) What happens when there is no clear Hyperplane in SVM? [3]
d) When would you use SVMs over Random Forest and vice-versa? [3]

7. a) What are the differences between supervised and unsupervised methods? [3]
b) How does supervised learning work? [3]
c) Briefly describe the steps involved in Supervised Learning. [3]
d) Briefly describe the types of unsupervised learning algorithms. [3]

8. a) What do mean by machine learning life cycle? [2]
b) Describe the steps involved in ML life cycle. [2]
c) (True or False?) Justify your answer. [3]
i) If you are given m data points, and use half for training and a half for testing, the difference between training error and test error decreases as m increases.
ii) Overfitting is more likely when the set of training data is small
iii) Overfitting is more likely when the hypothesis space is small
d) Describe the methods used to split the dataset in machine learning. [5]