Semester - 6th
Subject Name - Machine Learning
Subject Code - CS 3035

SPRING MID SEMESTER EXAMINATION-2024

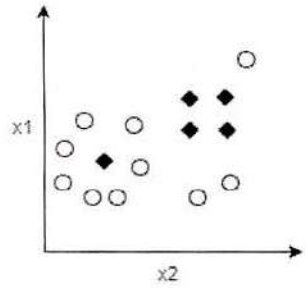
School of Computer Engineering
Kalinga Institute of Industrial Technology, Deemed to be University
Machine Learning
CS 3035

Time: 1 1/2 Hours **Full Mark: 20**

Answer Any four questions including question No.1 which is compulsory.
The figures in the margin indicate full marks. Candidates are required to give their answers in their own words as far as practicable and all parts of a question should be answered at one place only.

1. Answer all the questions. [1 Mark X 5]

a) Consider the following dataset. The data points are classified into 2 classes denoted by different colors. x1 and x2 are the predictor variables.



Draw the approximate decision boundary for the given dataset based on the Nearest Neighbour algorithm.

b) Show that the derivative of the logistic function g(x) is (1 - g(x))g(x).

c) A data scientist is evaluating different binary classification models. A false positive result is 5 times more expensive (from a business perspective) than a false negative result. What criteria will you use to evaluate your models and why?

d) When is Lasso regression preferred over ridge regression?

e) What are the advantages and disadvantages of stochastic gradient descent?

2.

a. Define the Learning Rate in the Gradient Descent Algorithm. Explain what happens if the learning rate is too high or too low. Using a cost function v/s iteration plot, show how the cost function changes per iteration when the learning rate is too high, too low, and moderate. [3 Marks]

b. Suppose you have a dataset of animals and you want to use KNN (with K=5 using Euclidean distance) to predict whether a new animal is a cat or a dog based on its weight and height. You have the following dataset Predict the species of a new animal that weighs 4 kg and is 30 Cm tall.

Animal	Weight (Kg)	Height (Cm)	Species
1	4	35	Cat
2	6	40	Dog
3	3	25	Cat
4	7	45	Dog
5	5	30	Cat
6	8	50	Dog
7	2	20	Cat

3.

a. The following table gives a data set about species. Using Naive Bayes classifier, identify the species of an entity with the following attributes. X={Color=Green, Legs=2, Height=Tall, Smelly=No}

Color	Legs	Height	Smelly	Species
White	3	Short	Yes	M
Green	2	Tall	No	M
Green	3	Short	Yes	M
White	3	Short	Yes	M
Green	2	Short	No	H
White	2	Tall	No	H
White	2	Tall	No	H
White	2	Short	Yes	H

[3 Marks]

b. Prove binary logistic regression model is also a linear classifier. [A linear classifier is an algorithm that separates two types of objects by a line or a hyperplane] [2 Marks]

4.

a. Briefly explain normalization and standardization with their properties. Explain their importance with an example. [3 Marks]

b. Argue why linear regression is scale-invariant but Ridge regression is not. (scale-invariant means multiplying any feature by a constant value does not affect the output value after training).[2 Marks]

5.

a. Consider the following dataset that contains information on whether someone will go out or not based on temperature, wind direction, rain, and humidity. Calculate the information gain of each of the attributes. Based on the information gain, state which attribute would be used to split the root node in the ID3 algorithm.

Temperature	Wind Direction	Rainy	Humidity	Going Out
High	South	N	High	Y
High	West	Y	High	Y
Medium	West	N	High	Y
Medium	South	N	High	N
High	West	N	Medium	N
Medium	South	Y	High	N
High	West	N	High	Y
Medium	South	N	High	Y
Medium	South	Y	High	N
High	South	N	Medium	Y

[5 Marks]