Semester: 6th
Subject Name:- Machine Learning.&
Code:- CS 3035

SPRING MID SEMESTER EXAMINATION-2023

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 Q.No.1 which is Compulsory.
The figures in the margin indicate full marks. Candidates are required to give theiranswers in their own words as far as practicable and all parts of a question should beanswered at one place only.

1. Answer all the questions. [1 x 5]

a) Which of the following tasks is NOT a suitable machine learning task?
A. Finding the shortest path between a pair of nodes in a graph
B. Predicting if a stock price will rise or fall
C. Predicting the price of petroleum
D. Grouping mails as spams or non-spams
Correct Answer : A. Finding the shortest path between a pair of nodes in a graph

b) In Linear Regression the output is:
A. Discrete
B. Continuous and always lies in a finite range
C. Continuous
D. May be discrete or continuous
Correct Answer : C. Continuous

c) The KNN algorithm should be used for large datasets? If no, give the reason.
The KNN algorithm should not be used for the large dataset.It is not feasible to store a large amount of data and also it is computationally costly to keep calculating and sorting all the values.

d) What is true for Stochastic Gradient Descent?
A. In every iteration, model parameters are updated for multiple training samples
B. In every iteration, model parameters are updated for one training sample
C. In every iteration, model parameters are updated for all training samples
D. None of the above
Correct Answer : B. In every iteration model parameters are updated for one training sample.

e) Under what conditions Minkowski distance is same as Euclidean distance.
When p=2

2. Long Answer Type Question

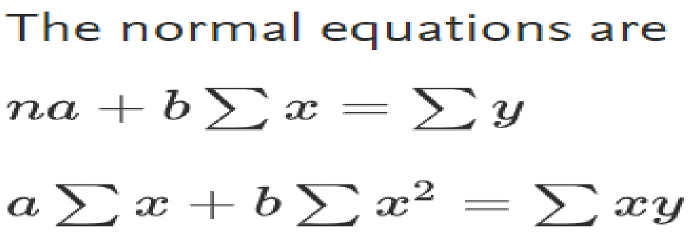
a) Fit a straight line Y=a+bX to the data by the method of least squares. [3 Marks]

X	1	3	4	2	5
Y	3	4	5	2	1

Answer:

X	Y	X^2	XY
1	3	1	3
3	4	9	12
4	5	16	20
2	2	4	4
5	1	25	5

$\sum X=15$ $\sum Y=15$ $\sum X^2=55$ $\sum XY=44$



The normal equations are

$$na + b\sum x = \sum y$$

$$a\sum x + b\sum x^2 = \sum xy$$

By putting the values in the above equation,we get b=-0.1 and a=3.3

Equation of best fit,Y=3.3-0.1 X

b) Write the effect of learning rate on the performance of a Gradient Descent algorithm?

Explanation: 2 mark [2 Marks]

3. a) Perform KNN Classification on the following training instances(see table),each having two attributes(X_1 and X_2).Compute the class label for the test instance t_1=(3,7) with K=3 using Euclidean distance. [3 Marks]

Training instances	X_1	X_2	output
I_1	7	7	0
I_2	7	4	0
I_3	3	4	1
I_4	1	4	1

Answer:

Training instances	X_1	X_2	output	Distance	Neighbor rank
I_1	7	7	0	4	3
I_2	7	4	0	5	4
I_3	3	4	1	3	1
I_4	1	4	1	3.6	2

For K=3,the test instance t_1=(3,7) belong to output 1

b) What do you mean by Generalization?Write the reasons for poor Generalization.

Generalization--- 1 mark

Reasons for poor Generalization --- 1 mark

4. a) Explain the merits and demerits of Cosine distance measure. Find the cosine distance between (1, 6, 1, 0), and (0, 1, 2, 2). [3 Marks]

merits and demerits --- 1 mark

$$\text{Cosine distance }(d1,d2) = \frac{d1.d2}{||d1||.||d2||}$$

$$=0.431$$

b) Explain bias-variance trade-off in context of model fitting using visual representations.

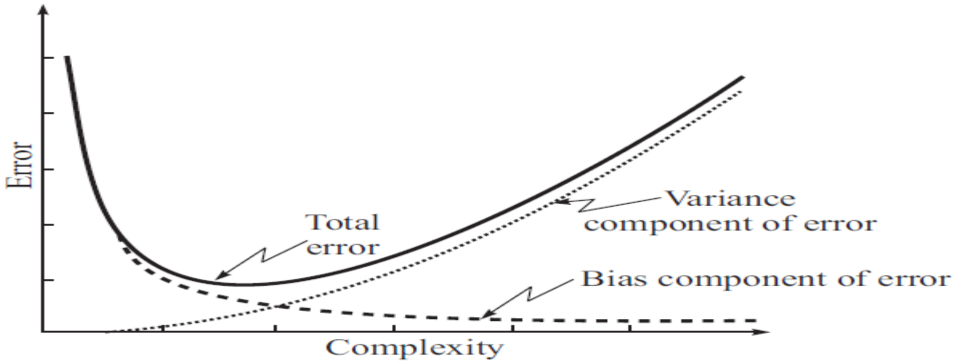


Figure 2.4 Bias-variance trade-off

[2 Marks]

5. a) Apply K-means clustering algorithm on given data for K=2. Use C1(4), C2(12) as initial cluster centers. Data:{2, 3, 4, 10, 11, 12, 20, 25, 30} [3 Marks]

Answer:C1: {2,3,4,10,11,12}
C2: {20, 25, 30}

b) Write the exact expression for ridge regression estimates of the coefficients. How is it different from the exact solution given by ordinary least squares (OLS)? [2 Marks]

Expression for ridge regression-----------------------------------1 mark

How it is different from OLS-----------------------------------1 mark