**Birla Institute of Technology & Science, Pilani**

**Work-Integrated Learning Programmes Division**

**Comprehensive Examination**

Course No. :

Course Title : Regression

Nature of Exam : Closed Book

No. of Pages =**2**

No. of Questions = **5**

Weightage : 40%

Duration : 2 Hours

Date of Exam :

Q1. **[ 4 + 4 = 8 Marks]**

1. Discuss the feasibility of applying regression algorithm for finding expected temperature in March ,2022 by considering the following data. If not, suggest suitable ideas to do so.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year & Month | Feb,21 | March,21 | April,21 | May,21 | June,21 | July,21 | Aug,21 | Sept,21 | Oct,21 | Nov,21 | Dec,21 | Jan,22 | Feb,22 |
| Temp(0C) | 12 | 20 | 32 | 40 | 34 | 32 | 28 | 29 | 30 | 18 | 12 | 10 | 14 |

1. Prove or disapprove the statement with examples and algorithms related:

” Regression concept is useful in classification also.”

Q2**. [ 8 + 4 = 12 Marks]**

1. Consider these data points: (x, y); (0,1),(1,3),(3,5), (4,7) ,(5, 9),(60,140)

Consider a linear hypothesis h(x) = θTx

i). Derive the cost (error)function related to this discussion. **(2M)**

ii) Draw the graph of the cost function. **(2M)**

iii)Find the point at which error becomes minimal and also find the error at the point.**(2M)**

iv) Estimate the value of y when x= 8 **(2M)**

**(NOTE: ANSWER THE QUESTIONS AS A DATA SCIENTIST)**

1. Discuss the need for applying Gradient Descent Algorithm.

Consider the data given: (x, y); (0,1),(2,5),(4,8), (5,11) ,(50, 120). Apply Gradient descent Algorithm by choosing initial weights randomly and alpha ( i.e. learning parameter as 0.01). Mention the stopping criteria to stop the iterations. **[4M]**

Q3. **[12 Marks]**

We know that understanding data is very important before training the algorithm. Consider the statistical summary of a data given and answer the questions given below

Table

Description automatically generated

1. Can we consider this as labelled data? Justify your answer**. [2M]**
2. Whether this summary is helpful in the identification of missing values? Justify your answer**. [2M]**
3. Validate the statement: “We need to train Multiple Linear regression only on the corresponding data”. **[2M]**
4. Whether this summary is helpful in identification of outliers? If so, how? **[2M]**
5. Is it possible to take a decision regarding normalization of the corresponding data based on this summary? Discuss in detail. **[3M]**
6. Mention other observations (at least one) which helps in understanding the data from above summary before trying for regression algorithms. **[1M]**

**Q4. [ 4 + 4 Marks]**

1. Discuss the criteria you follow while modelling the data with 20 Features and one target variable (which is a continuous variable) to find the best algorithm for prediction.  **[4M]**
2. (i) Discuss the statistical models/methods/concepts which are helpful to model the data with Regression models. Give suitable example in support. **[2M]**

**(ii).** What is meant by overfitting in regression? How to handle this **[2M]**

**XXXXXXX**