TECHNO MAIN SALT LAKE

(FORMERLY TECHNO INDIA, SALT LAKE)

Name ..... Kewal Raj .....

Roll No. 13030822018 ..... Stream ..... CSE (AIML) .....

Subject ..... Machine Learning (PCC-AIML601) ..... Semester ..... 6th .....

Invigilator's Signature ..... Date .....

(Part-A)

1. The two most Common tasks are:-

In age recognition and Customer-churn Analysis.

2. The purpose of Validation ~~test~~ Set are:-

After analysing the data through trainset and testing ~~Set~~ through test Set for multiple times to find the final accuracy Validation Set is needed.

3. (i) Coefficient of the Variable.

(ii) Loss function i.e mean Square error (MSE)

(iii) Gradient decent.

4.) AUC value of perfect classifier $\tan\theta \geq 1$.

5.) Out of precision, recall, recall is more important for a Spam email detection System.

(Part-B)

6.) Train-test-split:-

when we ~~fit~~ get a new dataset, we ~~how~~ divide the dataset initially into train-test-split. i.e. 70-80% of data are kept initially for training and after traing the ~~data~~ model, we use 30-20% of dataset for testing On this basis, we train the model as well as test the model Simultaneously, this division of dataset into 70-30 is spliting of data for the purpose of training and testing of data.

Overfitting of data: And Underfitting:

In Overfitting of data, we basically try to fit data more precisly and accurately, we try to cover each and every point possible, which is very strict in nature And, Underfitting of data says that, we are less Concerned of Considering ~~dog~~ data, we traverse in dataset, and whatever number of data we get is the possible data, But it does not gives more accuracte models, because Sample Size of data may decrease.

8) General algorithms that are available to minimize Cost function:-

→ Naives Bias algorithm
→ Random forest algorithm
→ Decision trees

Cost functions depends on the regression model, we select, Cost function may reduce for Some problem and also may increase for Some problem,
Basically, Cost function is low for linear regression model and at the Same time it is Increased for logistic regression model.

In linear regression, basically we Similify the problem with given datasets, which takes less time than Logistic regression problem, So, it will lower the Cost function and Simultaneously it will increase the Cost function for Logistic regression problems.

(9) Confusion matrix is the essential aspects of training model, we will get Confusion matrix for every model, we train. It basically tells about the model in more details, without Confusion matrix we Can't be able to tell about the precision, recall, false negative rate, false positive rate of model.

Given: TN = 82
FP = 3
FN = 5
TP = 10

false negative rate = $\frac{FN}{TN + FN}$

$= \frac{5}{82 + 5}$

$= \frac{5}{87}$

FNR $\simeq$ 0.05

false positive rate = $\frac{FP}{TP + FP}$

$= \frac{3}{10 + 3}$

$= \frac{3}{13}$

FPR $\simeq$ 0.29

PRECISION = $\frac{10}{10 + 3 + 3 + 82} = \frac{10}{100} = 0.1$

Recall = $\frac{TP}{TN + TP} = \frac{10}{82 + 10} = \frac{10}{92} = 0.108$

(7) Bias and Variance of a Machine Learning model.

→ Bias are dataset (how much deviated from the actual central tendency)

→ Variance is the (how dataset is or far from the central nature of a given model.

To reduce Bias and Variance, we should avoid:-

→ overfitting and underfitting

→ increase Sample space of datasets

→ more frequent train-test-split.

Bias-variance trade off is basically, first we know how much deviation is there in the model, then we fix the model by knowing how far we are from the central tendency, this process is called Bias-Variance trade off.