Name:Keyal raj

Roll No.: 13030823018

**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 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: tanθ ≥ 1.

5. Out of precision recall, Recall is more important for a spam email detection system.

**PART-B**

**6. Train-test split:**When we get a new dataset, we divide the dataset initially into train-test split. i.e., 70-80% of data are kept initially for training and after training 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 datasets into 70-80 is splitting of data for the purpose of training and testing of data.

**7. Explain Overfitting and Underfitting of data:**  
In overfitting of data, we basically try to fit data more precisely and accurately, we try to cover each and every point possible, which is very strict in nature.  
In underfitting of data says that, we are less concerned of considering data, we traverse in dataset, and whatever number of data we get is the possible data, but it does not give more accurate models, because sample size of data may decrease.

**8. General algorithm that are available to minimize Cost function:**

* Naive Bayes algorithms
* Random forest algorithms
* Decision trees

cost functions depends on the regression model .we select, cost function may reduce for some problems and also may increase for some problems.  
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 simplify 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 cannot be able to tell about the precision, recall, false negative rate, false positive rate of model.

**Given:**

* TN = 82 false negative rate=
* FP = 3 =
* FN = 5 =
* TP = 10 =0.05

False positive rate =

=3/(10+3)

=3/13

=0.29

Precision =10/(10+3+3+82)=10/100=0.1

Recall=TP/(TN+TP)=10/(82+10)=10/92=0.108

10. Bias and Variance of a Machine Learning model.

* **Bias** is dataset (how much devided from the actual central tendency).
* **Variance** is the (how dataset is or far from the central nature of a given dataset).

To reduce Bias and Variance, we should avoid:

* overfitting and underfitting.
* Increase sample space of datasets.
* Use more frequent train-test split.

**Bias-Variance trade-off** is basically first we know how much damaging is there in the model, then use it. It is moved by knowing how far we are from the central tendency of both process or causal Bias-Variance tradeoff.