Name: Debopriya Lahiri

Roll No.: 13030822008

**PART A**

1. Two most common supervised learning tasks are - Classification & Regression.

2. The response of validation set is to reduce the problem of overfitting and under fitting and by improving generalization it generalization error.

3. There are only one model parameter (Θ) and bias term (Θ0) with a single feature Variable.

4. AUC value of perfect classification is 1.

5. Recall is more important for Spam detection system.

**PART B**

9. Confusion Matrix :- Confusion Matrix is a comprehensive model evaluation on performance measure metrix which is visualised as a tabular form and counts the number of actual outputs versus perdicted output.

N

P

Actual

|  |  |
| --- | --- |
| TP | FN |
| FP  P | TN  N |

Predicted

TR - True Positive is model is predicted positive classes correctly.

FP- False Positive is model predicted positive but actual is negetive (Type 1 error)

FN – False negetive is model predicted negetive but actual is positive. (Type 2 error)

TN- True negetive is model predicted negetive classes correctly

TP+FP = Total +ve predictions.

FN+ TN =Total -ve predictions.

TP+ FP+FN+TN = Total predictions.

TP+FN = Total actual positives.

FP+TN = Total actual negetives.

Importance :-

1. Confusion matrix provides how the model is performing in unseen data, helps to evaluate it’s generalization Capability.
2. It helps to calculate other performance measure metrics.

TP + TN

Accuracy =

TP + FP + FN + TN

TP

Precision =

TP + FP

TP

Recall =

TP + FN

ROC Curve = Plot of TPR and FPR at different threshold.

PR Curve:- Curve of Precision & Recall.

In given classification problem, TN = 82,

FP = 3,

FN = 5,

TP = 10.

10

10

TP

Precision = = =

13

10 + 3

TP + FP

10

10

TP

Recall = = =

15

TP + FN

10 + 5

1

5

5

FN

FN

False Negetive rate = = = = =

3

15

5 + 10

FN + TP

FP + TN

FP

3

3

85

3 + 82

FP + TN

False Positive Rate = = =

6 .) Train – Test Split :- Train – Test Split is defined by dividing the whole dataset into two portions .

1. Train Set :- For train the data to the model . (higher in ratio)
2. Test Set :- For testing the model predictions and performance in unseen data. (smaller in ratio ) . (70:30)

It is generally used in Supervised learning algorithms where we train the model by labelled data and test the model by ~~the~~ providing the unlabelled data and compare the actual and predicted output.

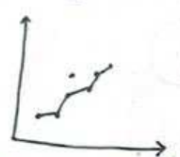
**.** from sklearn . model selection import train test\_split .

Overfitting :- Overfitting occurs when model performs good in training Set and poor in testing Set.

**.** In overfitting model tries to memorize the data rather than generalizing.

**.** In overfitting model has high variance.

**.** Model ~~tries to~~ is too Complex that it tries to predict all points in training but done to high variance it fails to predict unseen data.



Overfitting

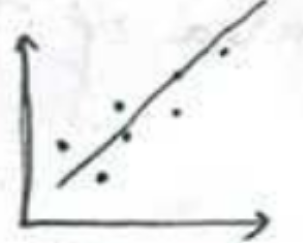
Underfitting :- **.** Underfitting occurs when model performs poor on training and testing data.

**.** In underfitting model doesnot able to extract the true patterns of data and have its own path (far from actual output).

**.** There, model has high bias ~~and low variance.~~

**~~.~~** ~~Class imbalance leads Same thing high bias and~~

**.** Model is to simple or linear and cann’t capture non-linear relations.



Underfitting

Ways to Prevent.

* Use ensemble learning techniques (Bagging & Boosting).
* Use Regularization methods (Ridge, Lasso, Elastic Net)
* Standardization and Normalization.
* For Overfitting, do feature Selection and for under fitting use kernel trick (SVM) and model with non – linear kernels.

7.) Bias :- Bais refers to the difference between actual and predicted output. It defines how close the generalizing hypothesis is of true hypothesis. (Low - Bias:- Near to actual value). (High - Bias: Far from actual, happens due to class imbalance sparse dataset).

Variance :- Variance refers how model predictions change by changing the feature vectors. It refers how scatter the predicted values are (Low Variance 🡪 datas (predicted are not scatter (close to each other)). (High Variance 🡪 predicted points are scattered).

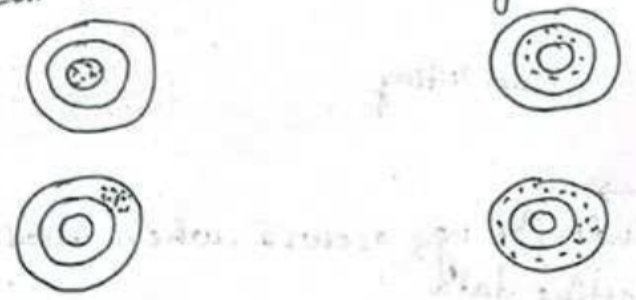
Way to Reduce Them :-

* To reduce bias, we need to balance the dataset by doing oversampling of minor class and under sampling of major class. Do ensemble methods like random forest, bagging.
* To reduce variance we need to do boosting techniques. We do feature Selection (Regularization techniques).

Bias Variance Tradeoff

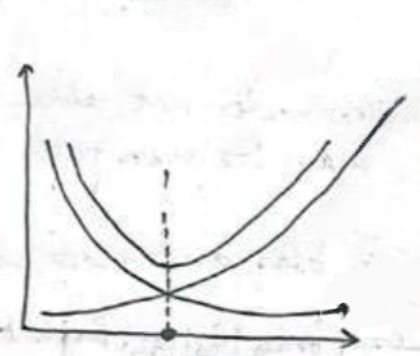
High Variance

Low Variance



High Bias

Low Bias



Generalization graph

Variance

Optional point is where bias and Variance is minimum (Low bias and Low variance)

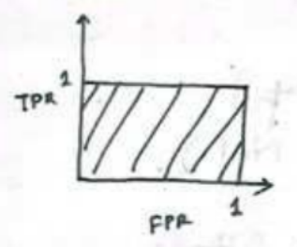
Optional Point

Bias

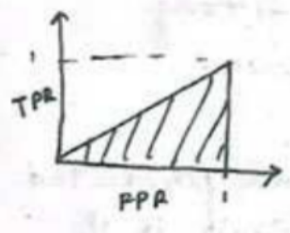
ROC Curve :- ROC (~~Rectif~~ Reciever Operating Characteristics)

refers to the graph of TPR vs FPR at different threasolds.

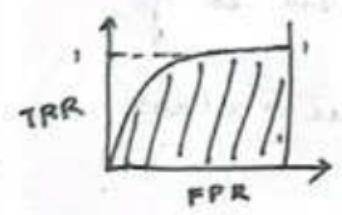
AUC : AUC (Area under Curve) refers to the area of ROC Curve .



Perfect Classifier (AOC = 1)



Random Classifier (AOC = 0.5)

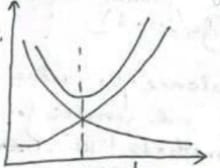


Random Classifier (0.5 <= AOC <= 1)

**Precision Recall Tradeoff**  It defines by the Curve of Precision and Recall. It plots precision and recalls.

Precision (Specificity) & Recall (Sentexity) are inversely proportions.

TP



It defines the change of TP (Precision) and FN (Recall)

Optional Points defines where precision and Recall is minimum.

TP + FN

TP

Recall =

Precision =

TP + FP

Precision

Sensitivity

Optional

Results

Specificity