TECHNO MAIN SALT LAKE

(FORMERLY TECHNO INDIA, SALT LAKE)

Name.....Debapriya Lahiri.....

Roll No. .....13030822008.....Stream..CSE-AIML.....

Subject .Machine learning Applications (PCCAIML601).....Semester....6.....

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

Part A

1. Two most common supervised learning tasks are - Classification & Regression.
2. The purpose of Validation Set is to reduce the problem of overfitting and underfitting and by improving generalization it reduce generalization error.
3. There are only one model parameter (θ) and bias term (θ_0) with a single feature variable.
4. AUC value of perfect classifier is 1.
5. Recall is more important for Spam detection System.

Part B

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

Actual \ Predicted	P	N
P	TP	FN
N	FP	TN

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

FP - ~~True~~ False Positive is model predicted positive but actual is negative (Type I error).

FN - False negative is model predicted negative but actual is positive. (Type II error).

TN - True negative is model predicted negative classes correctly.

TP + FP = Total +ve predictions

FN + TN = Total -ve predictions

TP + FN = Total actual positives

FP + TN = Total actual negatives.

TP + FP + FN + TN = Total predictions

□ Importance :-

① Confusion matrix provides how the model is performing in unseen data, helps to evaluate its generalization capability.

② It helps to calculate other performance measure metrics

$$\text{accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

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

PR Curve :- Curve of Precision & Recall.

In given classification problem, TN = 82

FP = 3,

FN = 5.

TP = 10.

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{10}{10+3} = \frac{10}{13}$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{10}{10+5} = \frac{10}{15} = \frac{2}{3}$$

$$\text{False negative rate} = \frac{FN}{FN+TP} = \frac{5}{5+10} = \frac{5}{15} = \frac{1}{3}$$

$$\text{False Positive rate} = \frac{FP}{FP+TN} = \frac{3}{3+82} = \frac{3}{85}.$$

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)

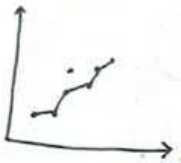
It is generally used in Supervised learning algorithms where we train the model by labeled data and test the model by the providing the unlabelled data and compare the actual and predicted output.

- 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 ~~and low bias~~.
• 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

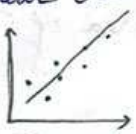
Underfitting :- • Underfitting occurs when model performs poor in training and testing data.

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

• Here, model has high bias ~~and low variance~~.

• ~~Class imbalance leads some 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).
□ Standaddization and Normalization.
□ For Overfitting, do feature selection and for underfitting use kernel fnith (SVM) and model with non-linear ~~re~~ 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).

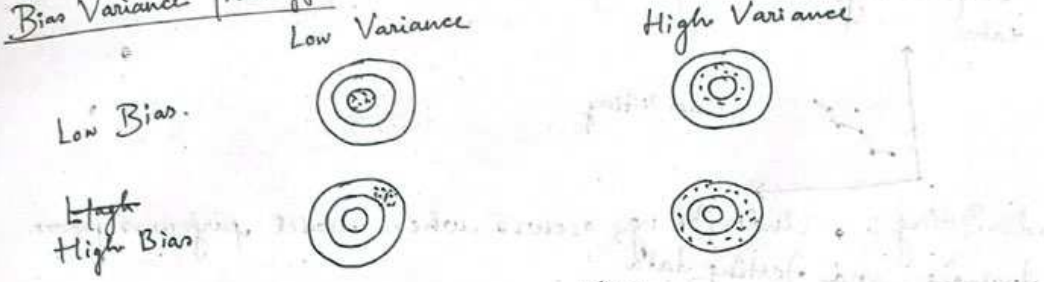
Varsince :- Variance refers how model predictions change by changing the ~~features~~ feature vectors. It refers how scatter the predicted values are (low Variance → datas (predicted are not scatter (close to each other)
High Variance → predicted points are scattered).

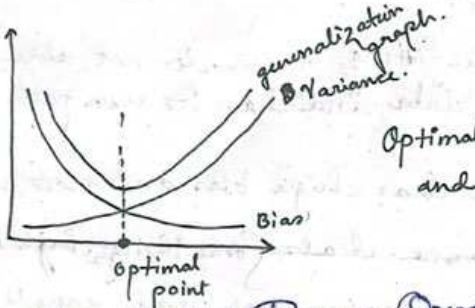
□ ~~Reducing~~ Way to Reduce Them. :-

• 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 ~~boosting~~. random forest, bagging

• To reduce Variance we need to do boosting techniques. We do feature selection (Regularization techniques)

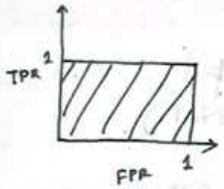
Bias Variance Tradeoff



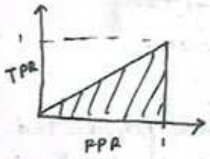


Optimal point is where bias and Variance is minimum. (Low bias and low Variance)

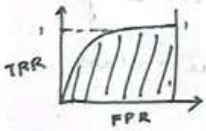
10. ROC Curve:- ROC (Recieuer Operating Characteristics) refers to the graph of TPR vs FPR at different thresholds.
AOC: AUC (Area under Curve) refers to the area of ROC Curve.



Perfect Classifier (AOC = 1).



Random Classifier (AOC = 0.5)

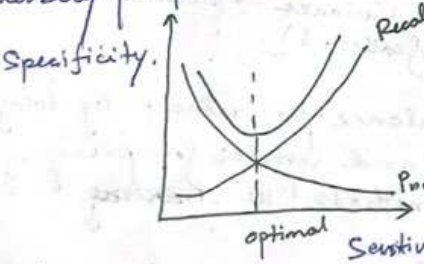


Random Classifier ($0.5 \leq AOC \leq 1$).

Precision Recall Tradeoff. It defines by the Curve of Precision and Recalls. It plots precision and recalls.
Precision (Specificity) & Recall (Sensitivity) are inversely proportions.

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$



It defines the change of FP (Precision) and FN (Recall).

Optimal points defines where precision and Recall is minimum