TECHNO MAIN SALT LAKE

(FORMERLY TECHNO INDIA, SALT LAKE)

Name ADITRI CHAUDHURI

Roll No. 13030822039 Stream CSE(AIML) - A

Subject Machine Learning Applications Semester 6

Invigilator's Signature [signature] Date 24/3/25

Part A

1.) Regression and Classification are the two most common supervised learning tasks.

2) The purpose of validation set is to ~~verify whether the model makes accurate predictions, and by~~ to reduce overfitting and ~~then~~ hence improve model generalisation.

3.) There is only one model parameter in a linear regression problem with single feature variable ~~$z = wx + b$~~ 33

4) The AUC value i.e the area under the ROC (Receiver operating characteristics) curve of a perfect classifier is 1 square unit

~~5) Precision~~

5) Out of precision and recall, Recall is the more important evaluation metric for a spam e-mail detection system. Since ~~precision~~ Recall refers to the fraction of total no. of positively classified instances that are actually positive while ~~Recall~~ Precision quantifies the ~~total no.~~ fraction of positive instances that are classified correctly.

6) Train-test-split refers to the process of 'splitting' or partitioning available data into ~~three~~ two distinct sets - ~~the training set and~~ one for training the model and the other for testing it. Typically, the training set is larger then the test set - a eg. ~~80-20~~ 80% of the set is generally used for training while the remaining 20% is used for testing.

'Overfitting' refers to the phenomenon which occurs when the model fits its training data too well - i.e. the variance is very high and bias is low leading to poor generalisation (accuracy) overfitting results, in exceptionally good results when the model is tested with known input parameters and poor results with unseen data.

'Underfitting' refers to the exact opposite phenomenon - when the model is not able to fit the training data too well ~~is~~ and is unable to capture its intricacies. Underfitting is a result of high bias and low variance. ~~It yields decent results~~ Unlike overfitting where the model becomes very complex (eg. a polynomial function or curve) in underfitting the model is simple (eg. a straight line). Both underfitting and overfitting are disadvantageous and we must try to strike a balance so as to reduce both and achieve best prediction results

9) $\text{Precision} = \frac{TP}{TP + FP}$ $FNR = \frac{FN}{FN + TP}$

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

The confusion matrix is important because it enables us to obtain a visual representation of the misclassifications and the no. of correctly classified instances which helps us to evaluate model performance. It also aids in the calculation of evaluation metrics such as precision, Recall, F1 score etc.

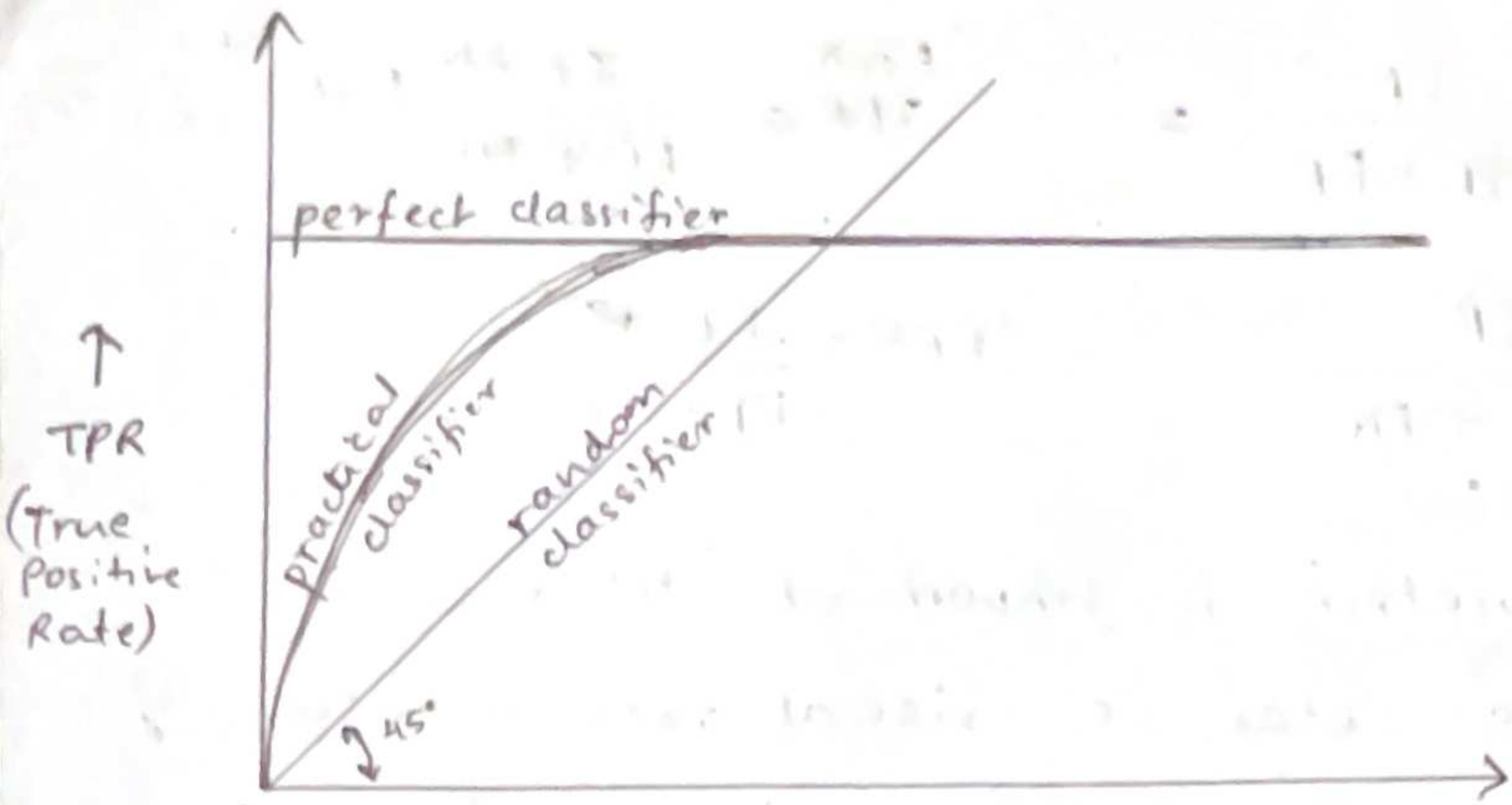
$$\text{Recall} = \frac{TP}{TP + FN} = \frac{10}{10 + 5} = \frac{10}{15} = \frac{2}{3} = 0.667$$

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{10}{10 + 3} = \frac{10}{13} = 0.76$$

$$\text{False Negative Rate } (FNR) = \frac{FN}{FN + TP} = \frac{5}{5 + 10} = \frac{1}{3} = 0.33$$

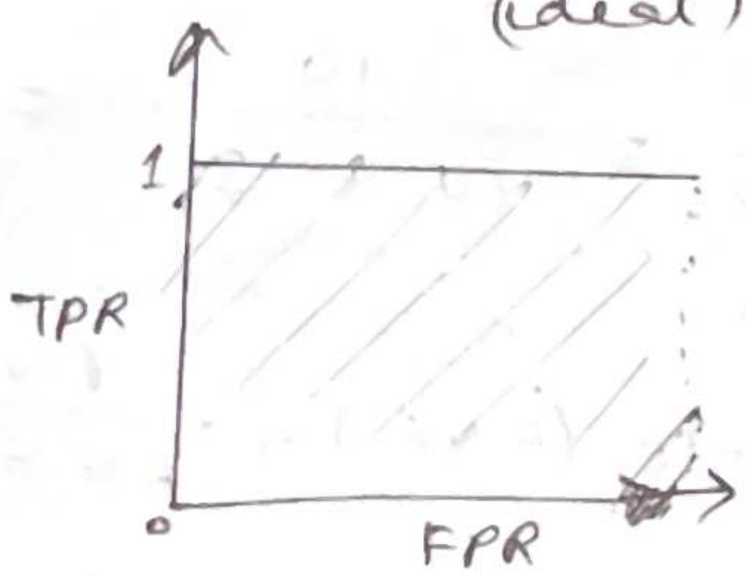
$$\text{False Positive Rate } (FPR) = \frac{FP}{FP + TN} = \frac{3}{3 + 82} = \frac{3}{85}$$

10) ROC (Receiver operating characteristics) curve refers to the graph or curve obtained by plotting the True Positive Rate against the False positive Rate. AUC refers to the area enclosed under the ROC curve.



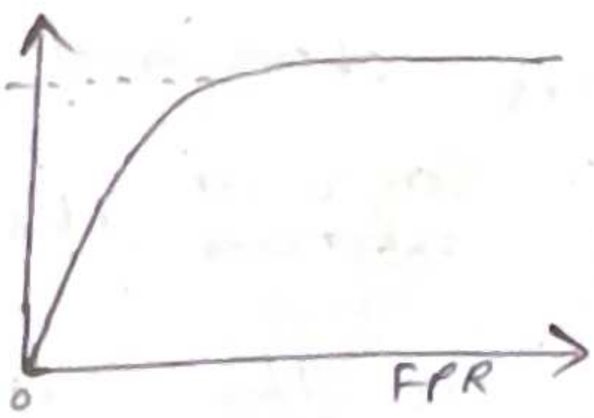
The adjacent graph depicts the ROC curve for perfect, pratical and random classifiers

11) case (a) represents a case where there is ~~no~~ absolute no overlap between the probability distributions of negative and positive prediction. It ~~represents~~ implies that there are absolutely zero misclassification (FPR=0) ie both FP and FN are equal to zero. Thus ~~(ideal)~~ case (a) represents a perfect classifier (ideal) with the ROC curve :—



AUC = 1

case (b) ~~represents~~ depicts a partial overlap which occurs in the case of most ~~cla~~ practical classifiers indicating that some positive and negative instances have been incorrectly classified yielding the following ROC curve —



0 < AUC < 1

case (c) depicts a case where ~~the~~ both probability distributions ~~pred~~ completely superimpose which occurs in case of a Random classifier yielding the following ROC curve :