**Techno Main Salt Lake (Formerly Techno India, Salt Lake)**

Name: Anurag Sen Roll No.: 13002022021

Stream: CSE (AIML) Sec-A

Subject: Application of ML in Industry

Semester: SEM-6 Invigilator's Signature: [Signature] Date: [Blank]

Part-A

1. The two most common Supervised task are Classification and Regression.
2. 'Validation Dataset' is a part of training dataset which is used in cross validation to increase the performance of model. This validation dataset goes under continuous testing with the seen or known data. So that the model gets trained properly & performs performs well on unseen data.
3. For a single feature variable two model parameters are required [y = mx + c]
4. The AUC value of a project classifier is 1.
5. Recall is important for a spam email detection system.

Part-B

1. For a given dataset, train-test split refers to the phenomenon in which a certain part or percentage of dataset is marked as Seen data or known data for the model and a remaining part- of dataset is becomes the testing part- for a particular machine learning model. So training the model is done using train dataset- here in the model is provided input and it's corresponding output, so that the model learns the analogy and thus get trained so it can map the same when there is an unseen sample.

So after training of the model, the time is now to check how our model perform on test set that is the unknown /unseen dataset so here input is not mapped to output, the machine predicts it and we check it if the predicted output is similar to the actual data of the dataset - This gives our accuracy or model performance. generally it's better to split 40% as training dataset and 30% as testing dataset . (Can be 80 - 20 too).

Overfitting means when the model performs well on the training dataset but not on the testing dataset the cause is usually the model is-to memorizes the training dataset- as a result performs poor on unseen data . To prevent this we have to reduce noise in training data thus eliminate Unwanted feature vectors and also co must use ensemble techniques like Random forest to Reduce overfitting .

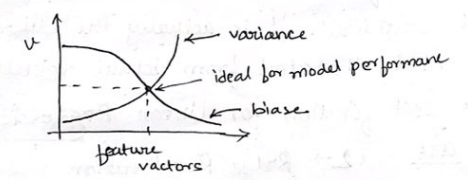
Underfitting means when the model performs well in both training and testing dataset it's Underfit- condition . The usual cause is the model is not very able and adaptive to learn the data from training dataset so performs ill in both cases . To prevent it we much induce proper feature we vectors and their hyperparameters in the dataset and model (tuning) respectively.

1. As the term 'bias' suggests biased towards a particular entity, in machine learning too the understanding remains the same - thus it means for a particular model. the model is biased to some feature vectors and not towards the others as a result modal being biased too wards a certain features learns leaves them properly but the rest remained unlearn as a result poor training of the model on entier dataset thus ultimately underfitting condition .

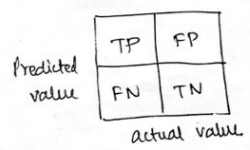
To reduce bias we must choose models and train them with both dataset a number of time , we may use Cross- validation approach for same also ensemble technique like better or results .

Variance In other hand as name suggest 'anomaly' thus Uncertain. due to presence of noisy data in dataset as a result due to presence of extra feature in dataset leads to overfitting condition. It can be reduced by applying feature reduction like PCA or SVD also ensemble techniques helps a lot .

Now Bias- Variance trade off is a phenomenon in which. if for a particular model bias increases then variance should decrease and vias -versa. we can conclude this from our observation of overfit - underfit theory thus if bias is high it's underfit so definitely the feature vectors are less thus 0 variance must be low.



9.Confusion matrix is a metrics in machine learning through which we can get an idea how to our model has classified or predicted in contrast to the actual result-



The main utility of Confusion matrix is to find out the accuracy and trade of misclassifation our model goes through dataset .

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

Precision = TP / (TP+FP) = 10 / (10+3) = 10/13

Recall = TP / (TP+FN) = 10 / (10+5) = 10/15

False Positive rate = FP / (FP+TN) = 3 / (3+82) = 3/85.

False Negative rate = FN / (FN+TP) = 5 / (5+10) = 5/15 = 1/3

8.Cost Function is the term used for loss function for a model where more than one features are considered. It is actually the difference how the model deviated from actual result.

The cost function for Linear Regression are

L2-> Ridge Regularization

L1-> LASSO Regularization & Elastic net- Cost can be for Logistic regression are similar.

Minimize by Least square principle.