TECHNO MAIN SALT LAKE

(FORMERLY TECHNO INDIA, SALT LAKE)

Name ANURAG SEN

Roll No. 13030822021 Stream CSE (AIML) sec-A

Subject Application of ML in industry
PCC-AIML601 Semester SEM-6

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

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 continous testing with the seen or known data, so that the model gets trained properly & ~~performan~~ performs well on unseen data.

3) For a single feature variable two model parameters are required $\boxed{y = mx}$

4) The AUC value of a perfect classifier is 1.

5) Recall is important for a spam email detection system.

PART-B

6) For a given dataset, train-test split refers to the phenomeanan in which a certain part or pecentage of dataset is marked as seendata 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 corrosponding 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 ~~traing~~ 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-pridects it and we check it if the
model.

predicted output is similar to the actual data of the dataset. - This gives our accuracy or model performance. geneally it's better to split 70% 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 perfoms poor on ~~te~~ unseen data. to prevent this we have to ~~d~~ reduce noise in training data thus eliminate unwanted feature vectors and also we must use ensemble techniques like Random Forest to reduce ~~d~~ overfitting.

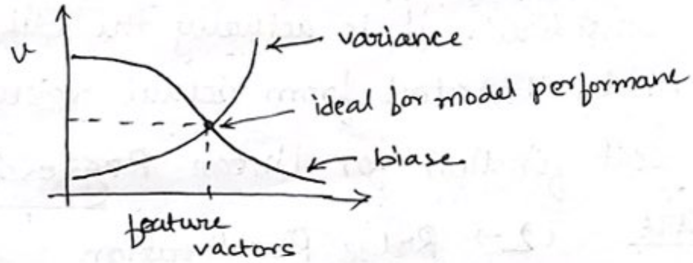
Underfitting means when the model ~~performs well on the~~ ~~te~~· doesn't performs well in both training and testing dataset it's underfit condition. The usual cause is the model is not very able and adaptive to learn the data from training dataset so performs ill in both cases. To prevent it we much induce proper feature ~~ve~~ vectors and their hypeparmeters in the dataset and model (tuning) respectivly.

(7) As the term 'bias' suggests biased towards a particular entity, in machine learining 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 model being biased towards ~~a~~ certain features ~~lerans~~ learns them properly but the rest remained unlearn as a result poor training of the model on entier dataset thus ultimatly underfitting condition.

to reduce bias we must choose models and train them with ~~both~~ dataset a number of time, we may use cross-validation approch for same also ensemble techniqu give better ~~o.~~ results.

Variance in other hand as name suggest 'anomaly' thus uncertain. due to presence of noicy 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 tradeoff is a phenomenan in which. if for a particular model bias increases then variance should decrease and vias-versa.
we can conclude this from ~~owo~~ our observation of overfit-underfit theory thus if bias is high it's underfit so definatly the feature vectors are less thus ~~o~~ variance must be low.



(8) ~~cost function is similar to loss function, it's called~~
~~Cost function when there are many feature considered.~~
at same time

(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 resut

Predicted value	TP	FP
	FN	TN
	actual value	

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

given

TN = 82

FP = 3

FN = 5

TP = 10

10	3
5	82

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

$$\text{recall} = \frac{TP}{TP+FN} = \frac{10}{10+5} = \frac{10}{15}$$

False ~~negative~~ Postive rate $= \frac{FP}{FP+TN} = \frac{3}{3+82} = \frac{3}{85}$

false ~~positive~~ negative rate $= \frac{FN}{FN+TP} = \frac{5}{5+10} = \frac{\cancel{5}}{\cancel{15}} = \frac{1}{3}$

(8) Cost function is the term used for Loss function for a model where more than one features are considered. it is actually the diffrence how the model deviated from actual result.

The cost function for Linear Regression are

L2 → Redge Regulization

L1 → LASSO Regularization

& elastic net

} Using it cost can be Obtaine

for Logistic Regressin are similar.

minimize by Least square principle.