SML

12m

① Statistical Modelling	Machine learning
→ depends on mathematical equation and finding the relation b/w expression	→ depends upon model and algorithm rather than mathematical eqn.
→ depends on the shape of the model	→ does not depends on shape of the model.
→ predicts the o/p with 85% accuracy & 90% confidence	→ predicts accuracy not more than 90%.
→ perfoms diagonistic parameters like pvalue, etc.	→ does not perform any diagonistic test.
→ Data splitup: training : 70%. testing : 30%	→ Data splitup: training : 50 %. Testing: 25%, validation - 25%
→ can be developed on single dataset called training data	→ can be developed on two datasets called training and validation data.
→ Used for Research purpose	→ used in production environment
→ from the school of statistics and maths	→ from school of CS.

② Steps in ML model development and deployment:

(i) Collection of data
↳ data can be collected directly from source data, web scrapping, API, chat interaction and so on.
↳ ML can work on both structured and unstructured data (voice, img and text)

(ii) Data preparation
↳ Data is formatted as per ML Algo.
~~processing values needs to be replaced by missing~~
↳ Missing values treatment need to be performed by replacing the missing and outlier values

(iii) Feature Extraction
↳ Analyse the data in order to find hidden patterns and relation b/w variables
↳ Helps to solve 70% of the problem
↳ New features are created

(iv) Training Algo and validation data
↳ Data will be divided into three chunks i.e. training, testing, validation & ML is applied on Trng Data
↳ validation are done to avoid overfitting.

(v) Test the Algorithm
↳ After training the model, ~~evaluation~~ its performance will be checked against unseen data
If performance is good enough, we can move further for next step.

(vi) Deploy the Algorithm

↳ Trained ML algo will be deployed on live streaming data to classify the outcomes

(3) Bias vs Variance Trade Off

↳ Every component has both Bias and Variance error components

↳ Bias : Adjust the Algo to fit into the model

↳ Variance : If it accepts the 50% of Model

↳ Bias and variance are inversely proportional to each other.

$$\text{Bias} \propto \frac{1}{\text{variance}}$$

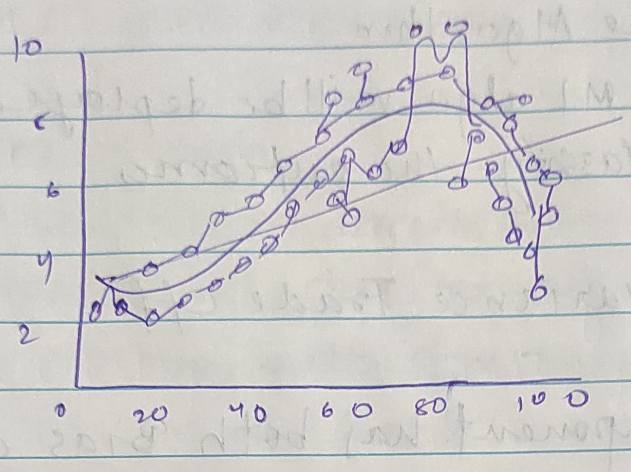
↳ If one component reduces, other component will increase.

↳ Errors from bias component can cause algo. to miss the relevant relation b/w features & target o/p.

↳ Errors from variance can cause overfitting problems

↳ Example for High Bias : Linear Regression

Example for High variance : Decision Tree.



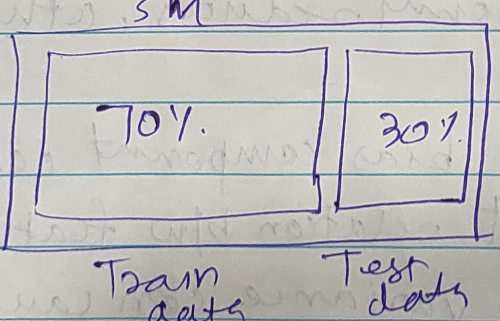
High Bias : Underfitting

~~Best~~ High Variance : Overfitting

Low Bias & Variance : Best fitting

Train and Test Data

↳ Data is splitted into 70-30 into train and testing data.



Train & Test split :

```
>> import pandas as pd
>> from sklearn.model_selection import train_test_split
>> original_data = pd.read_csv("cars.csv")
>> train_data, test_data = train_test_split(original_data,
    train_size = 0.7, random_state = 42)
```

④ Statistical Terminology for model building and validation:

(i) population : Includes the entire data which is collected

(ii) Sample : It is subset of population, small portion of population

(iii) parameter vs statistics : any measure that is calculated on population is parameter
: any measure that is calc. on sample is statistics

(iv) Mean : Arithmetic application to calculate average which is computed by the total sum divided by the count.

(v) Median : Mid value of the data.
: calc. by ascending or descending order.

(vi) Mode : Most repetitive point in data.

(vii) Range : Diff. b/w max. and min. of value

(viii) variance : Mean of the squared deviation from mean

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

$x_i \rightarrow$ data pt.

$\mu \rightarrow$ mean

(ix) Standard Deviation : Square root of variance.

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}$$

(X) Quantities: Identifies the fragments of the data.
* percentile: x% of data below the value of orig: data (50%)
* decile: 10% of the whole data
* Quartile: 25% of the whole data
* Interquartile range: 25-75% of whole data

(XI) Hypothesis Testing: process of making ~~overall~~ inference about overall population by conducting test
* Type I Error: rejects Null hypo, for True
* Type 2 Error: accepts Null hypo for false

(XII) Chi-Square Test of independence.

$$\chi^2 = \sum_i \frac{(O_i - e_i)^2}{e_i}$$

Ans

① Confusion Matrix : ~~Matrixe of actual versus predicted~~
↳ Matrix of the actual versus the predicted

(i) True Positives (TPs) : predict the disease as yes but actually does~~not~~ have disease

(ii) True Negative (TNs) : predict the disease as no but actual does not have disease.

(iii) False Positive (FPs) : predict the diseas as yes, actually does not have disease

(iv) False Negative (FNs) : predict the diseas as no, actually does have disease

(v) Precision : TP / TP+FP

(vi) Recall : TP / TP+FN

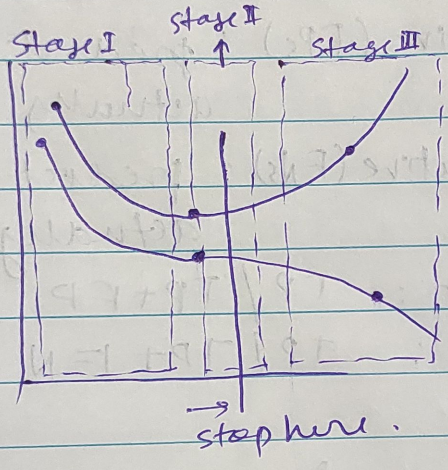
	Predicted: Yes	Predicted: NO
Actual: Yes	TP	FN
Actual: NO	FP	TN

② When to stop tuning ML models:

(i) Stage 1: Underfitting stage
↳ high train and high test errors

(ii) Stage 2: ~~Best~~ Good fit stage
↳ low train and low test error

(iii) Stage 3: Overfitting stage
↳ low train and high test error



Stage 1: underfitting
Stage 2: good fit
Stage 3: ~~over~~ fitting

③ Supervised Learning: Relationship between other variables and target variables.

: Major segments used is
→ classification problem
→ Regression problem

(i) House price: need data about houses square feet, no of rooms, features, etc.
: need to know the prices of these houses
: from thousands of data of house price we will classify the features and price and train to predict new house price

(11) Weather Prediction : need diff parameters like historical temp, wind humidity, etc.
: using the regression problem, we can use the the output labels to predict whether it is going to rain or not.

(4) Machine learning losses.
↳ the loss function in ml is a function that maps the values which represents some cost with the var. values
↳ Zero-one loss, value of loss is 0 for $m \geq 0$
value of loss is 1 for $m < 0$
error is diff. b/w predicted and actual o/p.
↳ Squared loss : loss fns that can be used while Predicting real value var. based on root mean sq. error
↳ Log Loss ; most imp classification based on probabilities
: used to compare diff models for better prediction
↳ Hinge loss : los fns used for training the classifiers for maximum margin.