Statistic analysis (1)

1. **Stages of CRISPDM**

- 1. Business understanding
  - i) objective
  - ii) constraint
- 2. Data collection
- 3. Data cleansing / wrangling / EDA (Preparing data for further analysis)
- 4. Model building
- 5. Evaluating Evaluating the model
- 6. Deployment

→ 60-80% time it take

total 6 Stages ⟸

2. **Key aspects of Business understanding**

S – Specific
M – Measurable
A – Achievable
R – Relevant
T – Time Bounded

3. **Elements of project Charter** : a formal document (first document). it includes:

- Defined objective
- Success criteria
- Resources planning
- CRISPDM → timeline
  → Risk at each phase

4. **4th stage of CRISPDM – What approach do we have**

4th Stage – Model Building

If want to predict or what to predict ⇒ Supervised

Unsupervised ⇐ Pattern matching

5. Types of Analytics

i) Descriptive → Describing the data. What has happened?

~~ii) Predictive~~

ii) Diagnostic → Why happened?

iii) Predictive → What may happen?

iv) Prescriptive → If any thing is automatic. There is no intervention of human.

6. Sources of Data Collection

mostly data is not readily available.

i) Primary data collection
- Survey data
- Design of Experiment
- web scrapping
- social media extraction

pros: Get data for exact variable of interest.

cons: costly & time consuming

ii) Secondary Data Collection
RDBMS, CRM, SCM etc.

pros: Data is available

cons: Data may or may not have variable of interest.

- Qualitative Data: can't perform math operation
  - Categorical
  - Fixed set of value / Non-numeric data

forex:- Blood Group, flavour of icecream.

8. if you have numbered data, does it necessarily make our data Quantitative?

→ NO, the data on which we can perform the Mathematical operation.

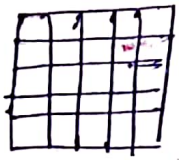
9. Balanced data

- equal proportion of all the category should be there
- ~~classes > 3~~ All the classified data > 30%.

10. Timeseries Vs Crosssectional Data

Date & Time & Sequence is where we have data matters in Timeseries while in cross sectional sequence may vary.

11. Example of Structured Vs Unstructured Data:



⇒ structured data

- images
- videos
- Text messages
- Tweets
- Reviews

} unstructured data.

12. which measurement level is least preferred.

- Continuous
  - Interval — subjective - forex: temp
  - ratio — Objective - forex: weight, Height
- Discrete
  - Categorical
    - Binary
    - multiple
  - Count

* Nominal data are least preferred ratio data are mostly preferred.

12. **Operation performed on ordinal data**

→ • Mode operation only.

13. **Can we represent nominal data?**

through Graph or graphical representation.

- Bar graph
- Pie chart

14. **Data cleansing process involves**

- Process of converting data from 1 form to another.
- other names are → Data munging, Data preparation, Data wrangling, Data organization

i) Type casting.

ii) Handling duplicates

iii) outlier analysis/Treatment

- 3R Technique: Rectify, Retain, Remove
- Sampling (detecting non-outlier or outlier)
- winsorization → Trimming

iv) Zero or Non zero variance-

Ignore the columns with the same ~~name~~ entries or throughout or if the majority of the entries all same.

for ex:- All entries of a column called country shows the same name as U.S.A.

v) Missing values-

Variants:
- Missingness at Random (MAR)
- Missingness Not At Random (MNAR)
- Missing completely at Random (MCAR)

Technique:
- Deletion
  - Row wise
  - List wise
- Single imputation
  - Mean
  - Median
  - Mode
  - Random
  - Hot deck
  - Regression
  - KNN
- Model Based
  - Maximum Likelihood (EM Algo)
- Multiple imputation

vi) Dummy variable creation

- one hot encoding
- Label encoding
- Dummy coding Scheme
- Bin counting Scheme
- Feature hashing scheme
- Effect coding Scheme

vii) Discretization / Binning / Grouping

→ Conversion of continuous data into discrete data

- Binarization - converting continuous data into two categories.
- Rounding - Rounding to nearest value
- Binning < Fixed width, Adaptive

viii) Normalization / standardization - making the data scale free & unitless.

- Standardized: $(X - \text{Mean})/\text{Stddev}$; mean = 0, Stddev = 1
- Min-max (Range method): $\frac{X - \min(x)}{\max(x) - \min(x)}$; Min = 0, Max = 1
- Robust Scaling: $\frac{X - \text{median}(x)}{IQR(x)}$

ix) Transformation
- Log
- Exponential
- Sqrt
- Reciprocal
- Box Cox – lambda value
- Johnson
  - lambda = 0 ⇒ log transformation
  - lambda = 0.5 ⇒ sqrt transformation

x) String Manipulation
- tokenization
- Stemming / Lammatization
- Stopword removal
- Document similarity
- Topic models
- word Embedding

15. Does imputation really makes data artificial and does it give us any false result?

→ NO. mostly it does not impact.

16. what pre process does Histogram ~~does~~ apply to show data?

→ - binning
- univariate

17 An analysis that can be applied to one variable at a time to plot or analyze it called as?

→ univariate
- Histogram
- Bar chart
- box plot

18. Why data cleaning plays an important role in analysis

19. what is probability?

→

$$P(e) = \frac{\text{\# interested event}}{\text{total no. of event}}$$

20. Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

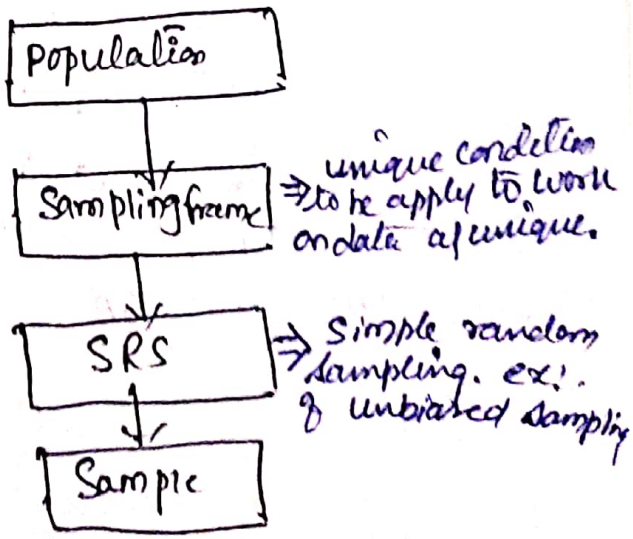
21. Probability Distribution :-

understanding the chance of each probability.

Probability Distribution
- Continuous
- discrete

22. Inferential Statistic or Sampling funnel?

→ getting result on sample data and drawing inference for population is called inferential statistics.



23. Sampling method

- Biased sampling :- you are selecting ~~Option~~ sample on your choice.
- unbiased / blind fold sampling ← most favoured technique
  - ex: Simple random sampling

Example of Biased sampling
- Judgemental
- Snowball
- Convenience ⇒ based on proximity

Example of unbiased
- SRS
- cluster
- Stratified sampling

24. Business Moment ? (Forth)

whatever management wants to see the insight of data.

- 1st Business moment decision
  mean, median, mode
  measures of central tendency
- 2nd Business moment decision
  range, variance, standard deviation
  measure of Spread
- 3rd Business moment decision

Skewness
- Negative skew or left skewed
- positive " or Right skewed
- if skewness = 0 then Normally distributed.

- 4th moment Business decision :-

Kurtosis
- = 3 means mesokurtic distribution / Normal distribution
- Negative kurtosis / platykurtic / -kurtic distribution ← wide Peak, Thin tails
- positive kurtosis / leptokurtic distribution ← Thin peak, Thick tail

25. can we apply mode on continuous data?

we may or may not apply the mode depending upon how much data is.

26. purpose of dummy variable Creation.

a) in model building certain algo can not take string values (Categorical data)

b) To give a numerical identity to categorical data.

27. when to do level encoding and when to do one hot encoding

- label encoding creates only 1 column in place of categorical variable. It is done when data is ordinal. (but we must take care, it gives labeling purely on alphabetically order.

28. Why do we need discretize the data?

⇒ Algo like tree based models can work only when the data is discretized.

29. What happens if we run a Decision tree with continuous input.

While the DT model will intelligently discretize the data. By performing EDA we can choose the different bin lengths to have more diverse input for the said model.

30. Which is better binarization or binning? (doing for o/p).

- Doing this to a continuous output to convert a regression model to a classification model or is this being don to the input.

It's a context based depends on requirement.

31. which is better binarization or binning in case of input data

→ Plotting with pairplot may decide otherwise there is no thumbrule.

but mostly we prefer for binarization cost is

32. Why normalization/standardization is done?

- make the data scalefree.

33. How does it matter if the data is having scale and unit?

→ Perform distance measure, predictive algo's, dimensionality reduction.

Performance distance measures

High unit values or high magnitude are going to influence the distance measure.

Follow up question: will the pair with shortest distance change because of normalization.

→ No, it will not change it but the scale of measure or the difference b/w the pair measure would be influenced by magnitude.

for ex

unnormalized data

A - B → 24,349

F - L → 24,347

B - ϕ → 265,012

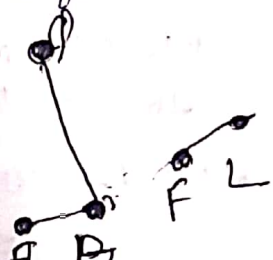


fig-1:- here distance b/w A to B and F to L looks similar

Normalized data

A - B → 3.6

F - L → 3.78

B - ϕ → 4.18

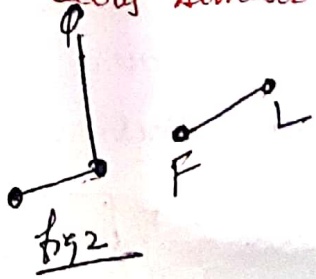


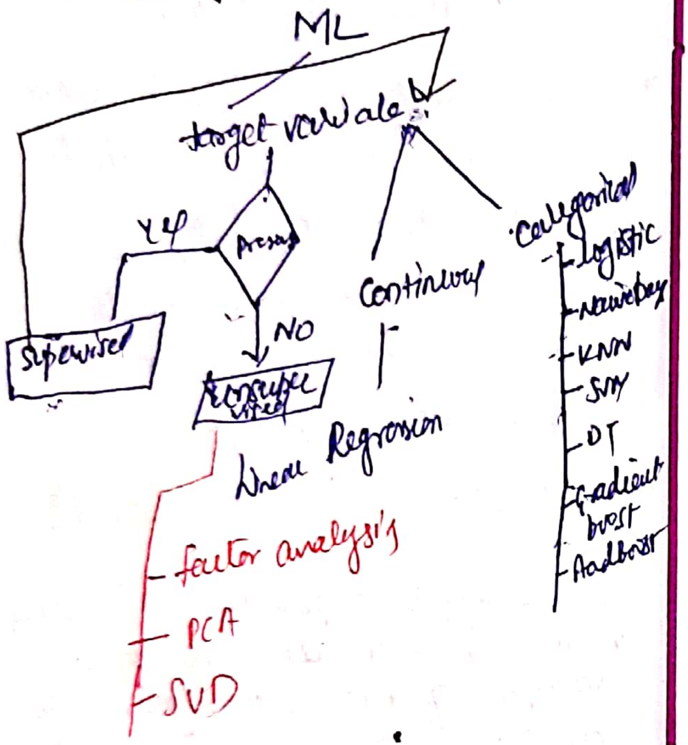
fig 2

by refferring 1 and 2 fig u can see there is no change. both are look similar

Machine Learning

1. What is ML learning :-

2. Different types of training model in ML?



Reinforcement Learning :- hit and try

3. How do you select the important variable while working on data?

→ There are many ways:-

1. Identify and discard correlated variables then finalize.
2. Variable selection w.r.t p-value.
3. Forward selection, backward elimination, stepwise selection
4. Lasso Regression.
5. Random forest and variable importance plot (VMP) based on info gain

3. How can we determine which algo should be used?

→ Depends on kind of data we are using.

Refer Ans of Question - 2

Supervised	unsupervised	Reinforcement learning & Decision making
• classification	• clustering	• Q-learning
• Regression	• K mean	• R-learning
• Neural n/w	• Mixed	
• SVM		
• Baysian Classifiers		
• Ensemble		

4. Difference b/w Correlation & Covariance.

If we do Standardization ⇒ Correlation

Note: Correlation does the Standardization

& > 0.85

5. When does regularization technique come into play?

→ To compensate overfitting

6. Bias & variance trade off?

→ Bias :- very ~~strong~~ limited flexibility

variance - vast sensitive very small fluctuation may also.

If we add more features, it will add more complexity. So bias will not be ~~change~~ loss but variance will be ~~affected~~ Rise.

So, in order to maintain the optimum error, we perform the trade off b/w Bias and variance.

Scanned with CamScanner

~~Low Bias & Low variance~~

7. High variance in data is good or bad?

→

high variance means feature has high variation and splitting is very high.

8. It is bad and to deal with it we do

9. How will evaluate the model if the accuracy is 90% (for ex). but minority data is less than 30%.

71 | 29

majority minority

→ on such imbalanced data set we should use —

1. Recall (sensitivity) => True positive rate
2. Precision (Specificity) => True negativity rate

To make data balance

- we have to deal with imbalance data to make balanced data.
- deal with anamoly data

10. use the right evaluation matrix :- Precision, Recall, F-1 score etc.

- Precision / ~~Sensitivity~~ Specificity = How many selected or instances are relevant.
- Recall / Sensitivity = how many relevant instances are selected
- F-1 score = harmonic mean of Precision & Recall

2. Resample the Training set

- undersampling : Balance the dataset by reducing the size of majority class.
- Oversampling :- it is used when the quantity of data is insufficient. it

3 use K-fold Cross validation in the right way :- do this before oversampling the data.

4. Ensemble different Resample dataset :-

one easy best practice is building n models that use all the samples of rare class (minority) and n-differing sample of abundant class (majority class).

for ex Given that you want to ensemble 10 models. you will keep eg. the 1000 case of the minority and randomly sample 10000 case of the abundant majority class then you just split the 10000 cases in 10 chunks and train 10 different models.

5. Resample with different ratios 1:1, 1:3 as even ?:)

6. Cluster the majority class : cluster the abundant class in n groups. with n being the no. of cases in n.

7. Design your models :- XGBoost is a good starting point use tweaking.

9. Difference b/w random forest and gradient boosting.

→

1. Random forest are no. of decision tree and taken the average.

Gradient boosting develop one tree at time. its a greedy algo and takes benefit from regularization.

2. Random forest never work on entire data

10. Why you go for ensemble learning if we have decision tree?

→ decision tree is prone to overfitting.

11. curse of dimensionality:-

→ refers to n-no. of features and we have to optimize the number of input.

Solution is PCA.

*note:- if we have too many features it may lead to overfitting. ⇒ does dimensionality reduction. it select all those features only which are orthogonal or not correlated.

Y	X (continuous / Discrete)	Algo	Remark
continuous	Single	Simple Linear Regression	
Continuous	Multiple	Multiple Linear Reg	
Discrete (2-Category)	Single	Simple Logistic Regression	
Discrete (2-Categories)	Multiple	Multiple Logistic Regression	
Discrete (>2 Category)		Multinomial Regression	Nominal
Discrete (>2 Category)		Ordinal Regression	ordinal
Count		poission Regression	mean = variance
Count		Negative Binomial Regression	variance > mean Overdispersion
Excessive zeros		ZIP or ZINB or HURDLE	
Overfit	Regularization Technique	Lasso Regression (L1) Ridge Regression (L2) Elastic Net Regression	

Error = predicted - actual

Model 1
Sales

	Predi	Actual	Error
Month 1	99	101	|-1|
Month 2	99	100	|-1|
Month 3	101	100	|1|
Month 4	100	99	|1|

Mean error = 0

MAE / MAD = 1

Model 2
Sales

	Pred	Actual	Error
	100	110	|-10|
	99	109	|-10|
	109	99	|10|
	110	100	|10|

MAE / MAD = 10

ME = 0

Mean Absolute error / Mean absolute Deviation

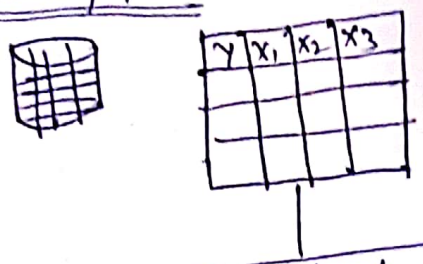
Mean squared error (MSE)

RMSE

Mean percentage error = it is beneficial when we want to represent the error unit less.

Mean absolute percentage error (MAPSE)

Machine learning primer



y = default / Non default

Balanced VS Imbalanced data.

atleast 50% data for Training

Random

information leak

Train

Test

80%.

70%.

60%.

20%.

30%.

40%.

Build the model ⇒ Algo learn the pattern find out the relation (Y & X) (using some math eqn)

$Y = \beta_0 + \beta_1 X + \beta_2 X_2 + \beta_3 X_3 + \cdots$

⇒ for ex: regression model was built.

Hyperparameter tuning

	Pred	Actual	
①	D	D	✓
②	ND	D	✗
:	:	:	:
200			

Accuracy = $\frac{\#\text{correct result}}{\text{overall result}}$

$= \frac{1600}{2000} \times 100 = 80\%$

Test Acc = 80%

Test Acc = 85%.

Test acc = 90%.

Principle of parsimony = Simplest model is the best model

PyChon Code

- load data
- Data
- Split data
- Build model (& Hyper parameter tuning)
- Test model

Test acc = 91%.

Train acc ↑ & Test ↓ → overfit / variance

↳ do Regularization Technique

Train acc ↓ & Test ↓ → underfit & Bias

↳ Get more data / Transformation / Feature

when we have less Hyperparameter do Train and Test only if we have more Hyperparameter use Train - Validation - Test.

when Training acc & Test acc both are high and close to each other ⇒ it Right fit

Problem with R^2.

- R^2 increases with every predictor added to a model. R^2 always increase and never decrease. It can appear to be a better fit with more terms we add to the model. This is completely misleading.
- if our model has many terms and too many ~~parameter~~ order polynomial we can run into the problem of Overfit. leads to wrong prediction.

Adjusted R^2

→ it is used to determine how reliable the correlation is b/w x and y. on addition to high correlated variable it will increase where as for variable with no correlation with dependent variable, the adjusted R^2 will decrease.

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right]$$

n = no. of points in our dataset

k = no. of indep endent - regressors. or no of input columns.

R^2 = Adjusted R^2

ROC curve

for regularization in R we use glmnet() library.

* to combat the right model dilemma we have ensemble technique.

* combine multiple Base learn
  - Homogeneous ensemble → applying multiple time M/c learning algo on same ~~model~~ algo
  - Heterogeneous ensemble → on different algo

* ensemble
  - Stacking → heterogeneous ensemble
  - Bagging → Homogeneous ensemble
  - Boosting → Homogeneous ensemble

* for the bagging & boosting usually we used tree based algo. non-tree may be used.

* through bagging & boosting we resolve issue of Bias and variance

Combining different models in Ensemble :-

* ensemble technique is used for Cross sectional & time series data as well.

1. Voting / max voting

- Hard voting → Go with majority
  - Yes
  - No ⇒ Yes (majority wins)
  - Yes
- Soft voting → Go with the average of probability of prediction.
  - Yes probability 90%.
  - No " 70%.
  - Yes " 20%.
  - Average = 40% ⇒ Soft voting

P<=50%.	P>50%.
NO	Yes

50% → Risk neutral

Note:- if we have classification problem go with voting.

2. Averaging :-

- Simple
- Weighted → whichever algo is given more accuracy give more weightage

	MPG	weighted
Linear	12.5 mg	.3
KNN	13.5	.45
SVM	13	.25
	13 (Simple average)	1

sum of weight should always be equal to 1

Note, it is used to handle Regression.

Sampling technique

- Probability Sampling technique
  - → SRS
  - → Stratified Sampling
  - → Systematic Sampling (Picking sample on regular interval)
- Non-probability Sampling

- Sampling error =

Total Data → 70% ND, 30 Default

Train → 68%, 32% ; Test → 71, 29

- Sampling Bias :

D. alt → 70, 30

Train → (ND) 38%, (D) 62% ; Test → 91% (ND), 9% (D)

Here samples are completely skewed. When the samples are completely skewed, it is sampling Bias.

Generalization error

means the error generated when testing the model on test data.

Data → Train → Build ; Data → Test → Generalization error (test error)

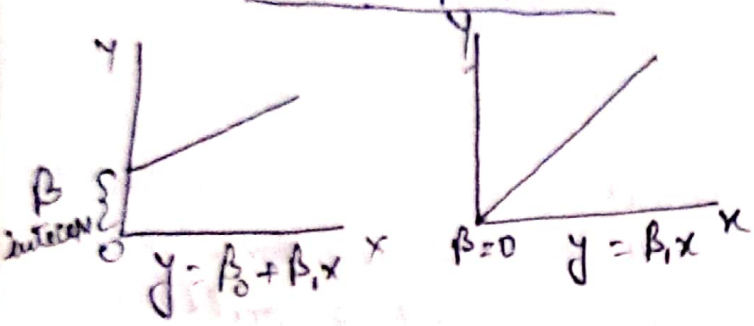
* Test data which is completely new is called out of sample / out of bag.
* Training data is called in sample / in bag.

Prediction.

- R^2 is also called Goodness of fit or coefficient of determination

$$R^2 = \frac{SSR}{SST}$$

- NO intercept model means if the line starts from zero means NO intercept model.



- Trivial Regression model:-

Regression model without coefficient or intercept. So, NO intercept model is Trivial Regression model.

- $\beta_0, \beta_1, \beta_2 \ldots$ are called as Partial coefficient or Regression Co-efficient.

- confounding variable:- a confounder is a third variable that affect variable of interest and make them seem related when they are not.

- Partial Correlation co-efficient:-

Cook's distance

- Any value > 0.5 is influential observation when we have plot with b/w standardize Vs residual Vs model.

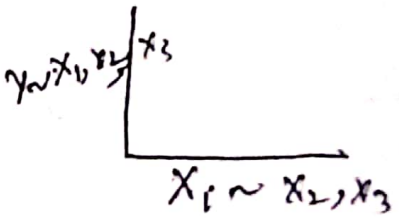
- variance Inflation factor:- if $VIF \geq 10 \Rightarrow$ co linearity

$VIF = \frac{1}{1-R^2}$ So, $VIF \propto \frac{1}{R^2}$

- Interaction variable - when two or more feature/variable combined have a significantly larger effect on a feature as compared to the sum of individual variable alone. This effect is called interaction effect & variable are called interaction variable.

- Confound Vs mediation variable a confounder is a third variable that affect variable of interest and make them seem related when they are not. mediator is a mechanism of a relationship b/w 2 variables. It explains the process through which they are related.

- Added variable plot:- It determines the magnitude of the co-efficient value of new value. its a plot b/w



- StepAIC - Akaike information criteria. it is used for model comparison. Model with least AIC value is best model.

Subset Selection

- Best selection – 2^P – model
- Forward Stepwise selection
- Backward Stepwise selection, Backward elimination Technique

→ total model = $\frac{1}{2} + \frac{n(n+1)}{2}$

i) if no of observation < no. of variable
→ Best selection & Backward selection will not work

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_n x_n + \epsilon$

explained — unexplained

- Logit Transformation (Link function) : use to make logistic into linear form.

- Techniques used to find the slope and intercept of a linear regression line which best fits the model?

→ ~~Olds~~ ordinary least squares (OLS)

Explain OLS in brief?

- It is a statistical method of analysis which will estimate the relation between two or more independent variable and one dependent variable.
- estimate the relationship by minimizing the ~~sum of squares~~ sum of squares of the diff between Observed and predicted values of dependent variable configured on straight line.
- It is used in bivariate model

Limitation of OLS? –

- computationally this is expensive if performs well on small data.
- for large data we need to prefer Gradient descent.

which evaluation matrix should you prefer to use for data having a lot of outliers?

→

- Mean Absolute Error (MAE)
- is a robust technique to deal with outliers whereas RMSE and MSE are susceptible to outliers.

what is the ~~intuition~~ intuition behind R^2?

→

we use linear regression to predict y given some x. But suppose we have to predict y value without corresponding x value

To measure the fit of a line we to calculate the sum of squared residuals – This gives an overall sense of how much prediction error ~~gives~~ a given model has once if we predict – we seem same data with regression.

* Decision tree is an example of greedy algo.

Decision Tree

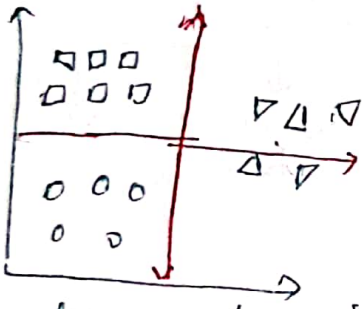
* Expected information content

$$= -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

* Information gain = Entropy before − Entropy after

Advantage
- subset feature selection.

Disadvantage



- it always do axis parallel split.
- for too many features it is biased

C5.0 algo in DT →

Feature Subset selection

it can be done through
- entropy (max Infogain)
- Gini index
- Chi Square
- Gain ratio

Regularization in Decision Tree

it is done through Pruning

Pruning
- Pre-pruning
- Post pruning

- Subtree raising / replacement

* default criterion of decision tree in Python is gini index

Random Forest

+ Regularization parameter

- no. of decision Tree.

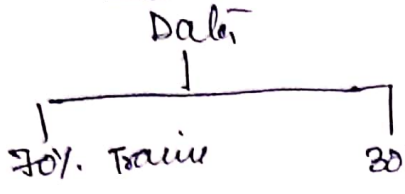
* n_tree - # of decision tree
* m_try for selection of features

- classification: $\sqrt{P}$
- regression: $P/3$

in RandomForestClassifier( )

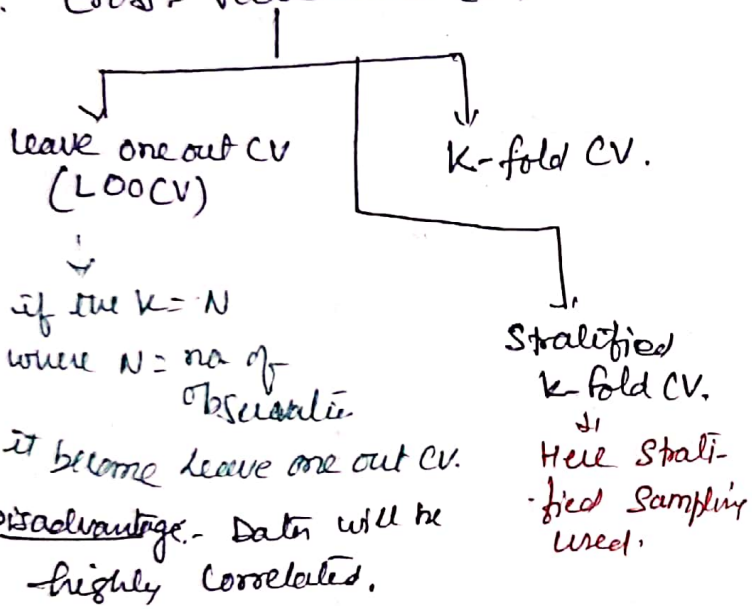
n_jobs = no. of job in parallel
n_estimator = no. of decision tree
~~Criterion~~
max_features = no. of maximum features allowed in a single decision tree.

1. Hold out method

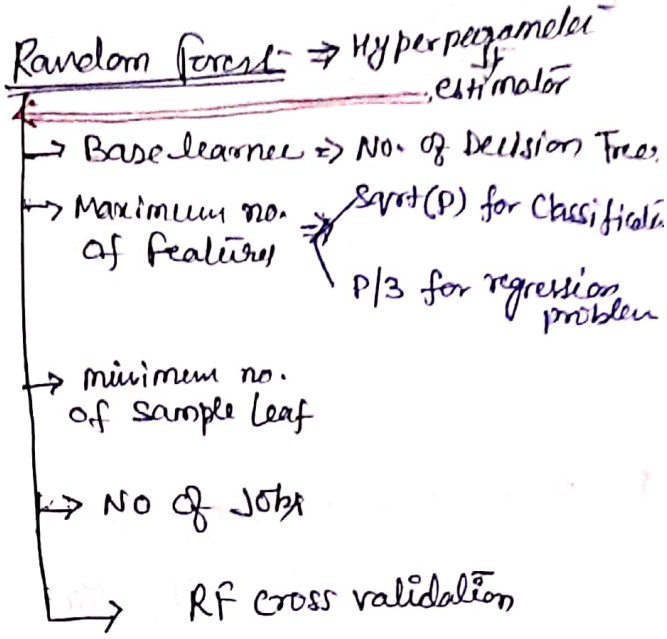


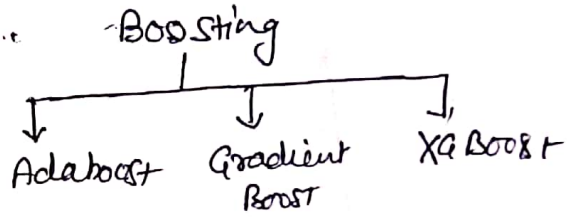
* When we split data in 70-30% ratio, it is called Hold-out method.

2. Cross-validation (CV)



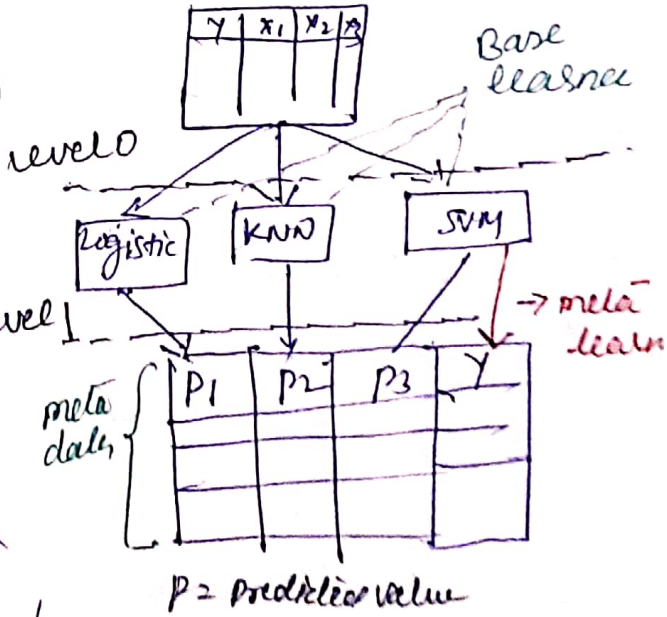
Random forest ⇒ Hyperparameter estimator





Stacking

- First we construct meta data

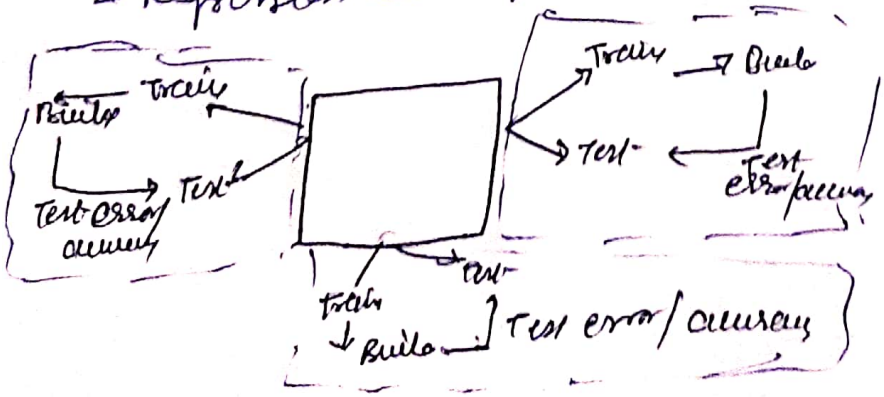


Bootstrap Sampling (Jack-knife method)

- Sampling with replacement.

Bagging (Bootstrap Aggregating)

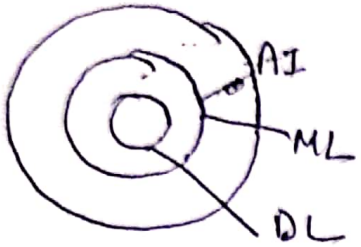
- Usually apply tree based also.
- Can be applied for classification & Regression technique



Deep learning

1. Difference b/w machine learning & deep learning

→

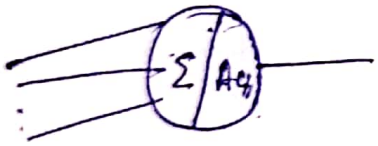


- in M/c learning, we use statistical algo, train to improve the machine with experience
- in Deep learning we try to mimic human brain.

2 what is perception?

→

It receive input from various entities, perform function and give input.



3. Which is Better DL or ML?

→

Deep learning	ML learning
- very efficient when large no. of data.	less efficient in comparison of DL

We are

4 Advantage of loss function

→

i) measure of accuracy.

ii) performance of Neural network

↓ loss function ⇒ Cost function ↑

5 Step to use Gradient descent algo?

→

i) weight are getting initialized Bias are also initialized

ii) i/p — 1st layer

iii) ~~Ac~~ ~~Expo~~ Actual value - predicted value

iv) Loss function

v) change the weight to minimize loss fast

These are repeated until we get the ~~Optim~~ better like model.

Epoch:- Reaching all the ~~Obse~~ Observation once.

iteration: Updating the weight. One ~~for~~ fore weight updation mean iteration = 1

* Single layer perception is effective only for linear pattern

~~Gradient descent Optimizer~~:-

Perceptron	MLP
i) not good for non-linear data	i) used for non-linear
ii) less efficient with large data.	ii) heavily good efficient with the large data.

Forward propagation	Back propagation
Sending the value of activation function as an input to other hidden layer.	try to minimize cost function
→	←

6. How can hyper parameter be trained in MLP.

Hyper parameter are to understand parameter

Batch Size

EPOCH

Momentum

Learning rate — parameter denote time require for the parameter updation. (η)

Momentum : it will understand the next consecutive action.

7. Different Gradient descent optimizers?

→

i) Batch gradient descent
Optimizer & 1 EPoch = 1 Iteration

ii) Stochastic Gradient descent
(Randomly) 1 EPoch = n iteration

iii) Mini batch Gradient descent.
No of Observation / batch size = total no. of Batches

$$W_{n+1} = W_n + \eta (AV - PV) x$$

8. Drop out methods in Deeplearning

→ Dropout is used to avoid overfitting.
. Dropout value ↓ - minimal affect on learning
Dropout value ↑ - reduce the efficiency

* Dropouts are happening random.

- Supervise learning in Deep learning.
  - ANN
  - CNN
  - RNN

UNsupervised learning in Deep Learning
- Autoencoders
- Boltzman machine
- self organizing maps.

Deeplearning Libraries
- Tensorflow
- keras
- Pytorch
- MXNet

Natural Language processing

① what is Natural Language understanding in machine learning.

→

It is a subtopic of NLP in AI that deals with M/C reading comprehension. it a comprehension part of NLP.

② Natural language procesing:-

→ it a field in machine learning with the ability of compuler to understand, analyse and manipulate and potentially generate human language.

Raw speech
↓
Sequence of word spoken
↓
Structure of sentence
↓
Sequence of meaning
+ parts of representation
↓
final implementation

③ Applications of NLP?-

- Information retrieval
- Information Extraction
- Machine Translation
- Text simplification
- Sentiment Analysis
- Text summarization
- Spam filter
- Auto predict
- Auto correct.

④ Natural language understanding terminology:-

→

1. Vocabulary: group of term in text
2. Corpus: collection of document eg: movie review
3. Document: body of text and collectively forms corpus
4. Preprocessing: attempt to remove unwanted text, noise, cleaning. This is first step of Text mining
5. Tokenization: process of breaking strings into tokens. Broken token provide meaningful information.
6. Embedding:-
7. N Gram:- sequence of N tokens for a given text.
8. Transformers: deep learning architecture that use to learn long term dependencies and these have the ability to do parallal computing.
9. parts of speech:- They are word functions like Noun, Pronoun, verb etc
10. Parts of speech tagging:- Converting words into different parts of speech.
11. Stop words:
12. Normalization: one of the process to map similar term in canonical form.
for ex: understanding 2 mmrw, 2 mrw, tomrw as tomorrow

13. Lemetization, works based on logical analysis of words.
for this, you should have detailed dictionary :-
for ex: Converts went, gone, ~~going~~ into go.

14. Stemming :- it's a similar to lammetization but without part of speech. in other words segregating words without part of speech.
drawback: it may chop the words without its context.

forex: Effeliin, effected, affecting etc as effect

(5) Steps in Natural language understanding ?

→
i) signal processing.
ii) Syntactic analysis.
iii) Semantic analysis
iv) Pragmatic analysis.

i) signal processing :- enable the SW to apply, modify or analyze the text signal.

ii) Syntactic Analysis : referred to as syntax analysis or parsing. use to draw the meaning by applying formal grammer rules.

iii) Semantic analysis - use to draw the meaning and with the help of grammer rules. It is the process of understand the meaning of text as human try to understand.

iv) Pragmatic analysis :- It's a proces to extract the informatic through some tool.

(6) What is Semantic analysis ?

- through this we try to detect the polarity of text/speech.
- form: how's the review, response intensity etc.

(7) Common NLP Techniques :-

→
i) Named entity recognition
ii) Text summarization
iii) Text modelling
iv) sentiment analysis

(8) Tools used for Training NLP models

→
i) NLTK
ii) pytorch NLP
iii) Open NLP

(9) Opensource library for NLP

i) Textblob
ii) Gensim
iii) scikit learn
iv) NLTK → most popular library for tokenization, tagging, classification, stemming.