NAME: Nilkanth Jadhav - 8087082261

STD: _____ DIV: _____ ROLL NO.: _____

SUBJECT: R with Data Science

INDEX

SR. NO.	DATE	TITLE	PAGE NO.
	30/7/18	Exploratory Data Analysis & Visualization	
	02/8/18	Central limit theorem	
	23/8/18	Machine Learning	
	24/8/18	Regression	
	28/8/18	Ridge & Lasso Regression	
	30/8/18	Time series, Stepwise regression	
	03/9/18	Classification	
		- KNN	
		- Logistic Regression	
		- KNN for regression	
		- Naive Bayes Classification	
	05/09/18	Decision Trees	
	06/09/18	Unsupervised Learning	
		Clustering	
		- K-means	
		- Hierarchical	
		- DB scan	
	19/9/18	SVM	
	20/9/18	ANN	
	26/9/18	Market Basket Analysis.	

Monday
30/07/18

Exploratory Data Analysis and Visualization (EDA)

- Understand the problem (Project charter) optional.
- collect data related to problem.
  what is current status about problem?
- Brain-storming session with management team.
  discuss about problem and find why problem?
  define metadata, take surveys about sub-factors.
- Collect data from warehouse.
- Process data - extraction, integration, clean-up.
- Transforming data - making data ready for analysis, filtering, sorting, numeric → category (descretization), column combination → new column
  - Box-cox Transformation & Johnson transformation un-normalized data to normalized data.
- Understand data - mathematical summery + visualization summery
  - column wise analysis.
  - pairwise comparison, $y \to x_1$, $y \to x_2$ - - - $y \to x_{25}$

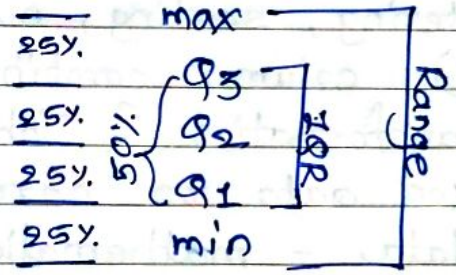
Google (Search) Fishbone pareto Co-ordinate systems

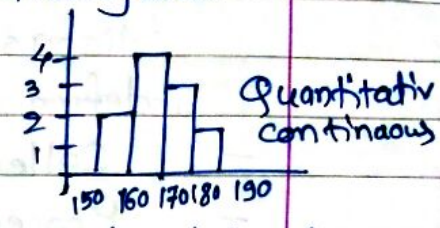
univariate & bivariate

Univariate analysis - column by column.
- Categorical
- Quantitative

Date:
Page No.:
YOUVA

Univariate	Mathematical Summary	Visualization
Categorical	Count, proportion, percent ratio, mode	Pi chart, bar Chart, ~~histogram~~, box plot. dot plot
Quantitative	mean, median, mode. std. deviation, range, (min, max, percentile Q_1 Q_2 Q_3) 5 no. summery IQR ($Q_3 - Q_1$) distribution of 50% middle elements.	Histogram



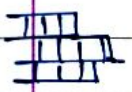


Class interval
how data is distribut order

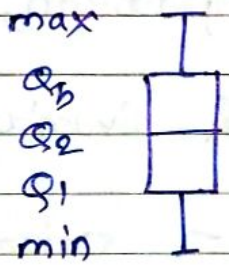
- Information loss in summarization (stem-leaf plot)

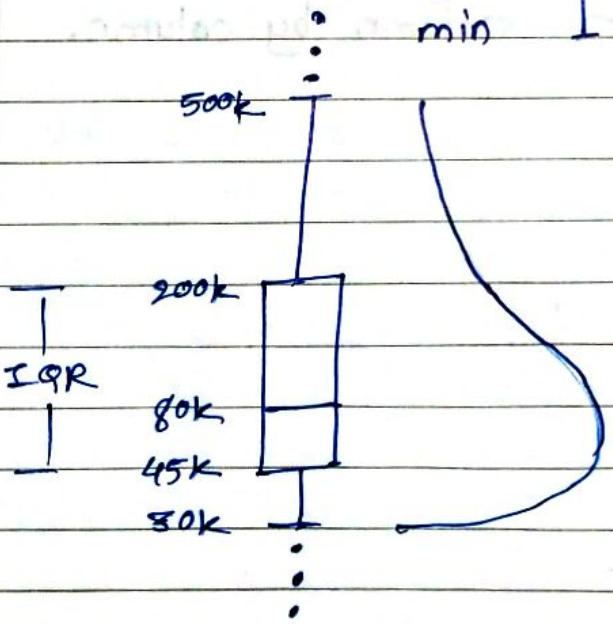
Sort data & partition into stem - leaf.

stem	leaf
15	1
15	3
16	1
16	5
17	0
17	9



Box & Whiskess plot:





Outlier detection
1.5 * IQR will limit bar graph shooting out of paper, outliers of higher values are represented by dots.

$\mu \pm 3\sigma$ are outliers.

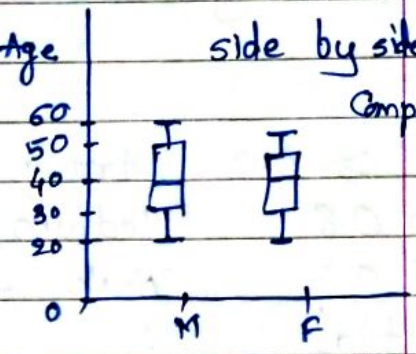
01/08/18

Bivariate | Mathematical Summary | Visualization Summary

✓ C → Q ↦ location → Salary, weather Cond" → Accd"., Gender vs. Hemoglobin
× Q → C
✓ C → C ↦ Smoking → Cancer, location → Gender, season → dieses.
✓ Q → Q ↦ experiance → salary, ½ ⇐ ~~rank~~ no. of hr. study, time → no. of product sold., distance → Time
salary → saving, petrol → avg.

C → Q Comparitive five number summary

	Male	female
Max	60	50
Q_3	45	35
Q_2	35	30
Q_1	27	25
Min	20	20



side by side Boxplot
Company Service

C → C

Smoking	Cancer
Y	Y
Y	N
N	N
N	Y
Y	Y
⋮	⋮

Gender

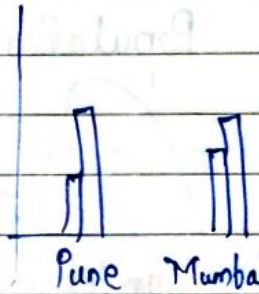


Table - (Two-way contingency)

Gender	M	F
B	0.47	0.53
P	0.45	0.55

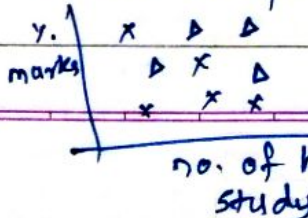
location	M	F	
B	334	471	= 705
P	762	916	= 1680

Comparitive proportion.
- Gender equality independent of city
- Salary distribution for medium band is dependent.

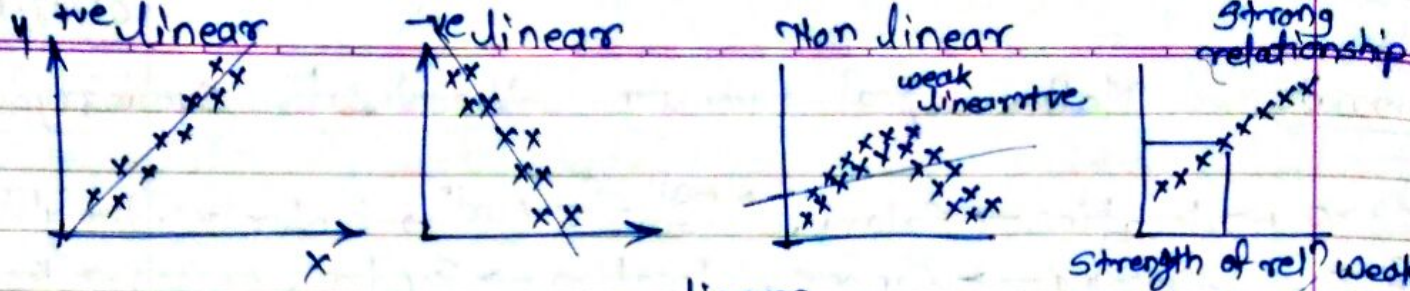
Salary	L	M	H
B	0.8	0.15	0.05
P	0.9	0.05	0.05

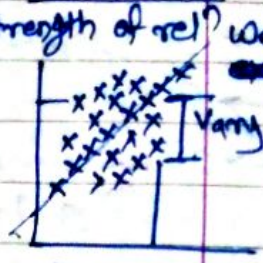
Q → Q Experiance vs salary

Scatterplot (Q → Q)



Stratified Scatterplot.

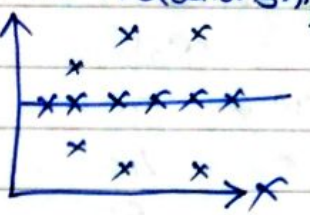




x & y has
- Direction of rel?
- linearity of rel?
- Strength of rel?

} linear Co-relation coefficient. (r) $-1 < r < +1$

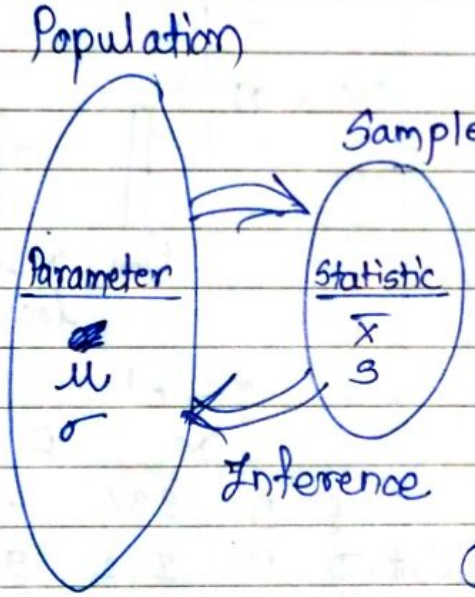
$r = 0$ No relationship.



$|r| \geq 0.8$ Strong rel?
0.5 to 0.8 Medium —"—
< 0.5 Weak —"—

Inferential

Census every 10 years. जनगणना
Tiger Census



Same
$\mu \rightarrow \bar{x}$
$\sigma \rightarrow s$

Inference is
$\bar{x}$ estimates μ
s estimates σ

- We estimate population parameters using sample statistic

① — Point estimate $\mu = \bar{x}$ there will be error

② — Interval estimate

confidence interval = $(\bar{x} \pm$ margin of error$)$ — μ varies in interval

③ — Test of hypothesis.

Date:
Page No.:
YOUVA

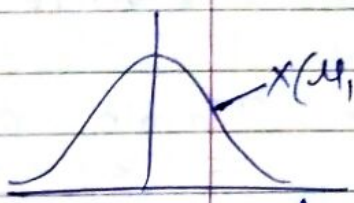
02/08/18

- Population parameters remains same as population does not change. Sample statistic may change as sample changes due to random items.

Study well

Central Limit Theorem -

- Centering, sprode, shape.



- mean of X and mean of $\bar{x}$ are centered at same point

  mean X → μ

  mean $\bar{x}$ → $\bar{\bar{x}}$

  when no. of samples are high

$$\bar{\bar{x}} = \frac{\Sigma \bar{x}}{N}$$

N - no. of sample takens

-

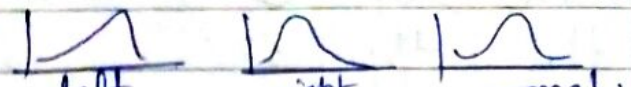
$$sd._{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

n - no. of sample pts

- For large enough sample size of $\bar{x}$ will tend to be normal irrespective of distribution of X

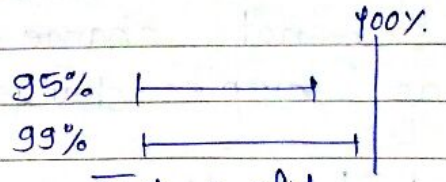
Sample size, n = 30 > its normal

~~जितना देख उतना सच~~



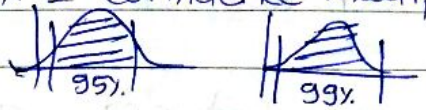
Interval Estimate –

- Confidence level/interval



$= \bar{x} \pm$ margin of error. $= \bar{x} \pm$ confidence multiplier $* \frac{\sigma}{\sqrt{n}}$

% Area will be covered

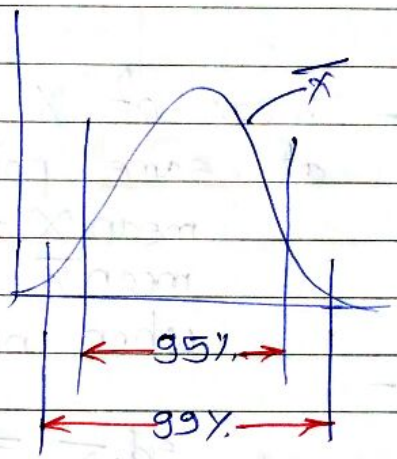


- Margin of error.

$= \frac{2 \times \sigma}{\sqrt{n}}$

$= \frac{3 \times \sigma}{\sqrt{n}}$

2.064



How many ~~std. deviations~~ range ~~center~~ times is away from std. deviation

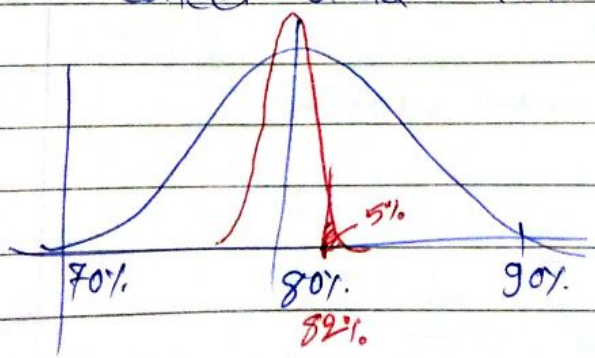
Test of hypothesis –

- Domain claims & management claim
  alternative hypothesis. null hypothesis (H_0)
  change No change

$H_0 : \mu_p = 80\%$

$H_a : \mu_p > 80\%$

- Collect data and test data for hypothesis.



25 samples.

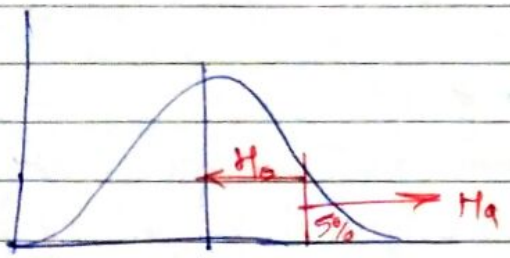
$\bar{x} = 80\%$.

$sd_{\bar{x}} = \frac{5}{\sqrt{25}} = 1$

$\sigma = 5$

your eg. H_0 = coin fair H_a = unfair α = threshhold, level of significance

$>$



$P(x \geq a)$

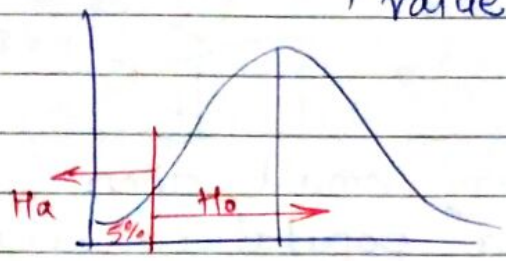
5% = α = 0.05

level of significance

$P(x \geq b)$ = p-value

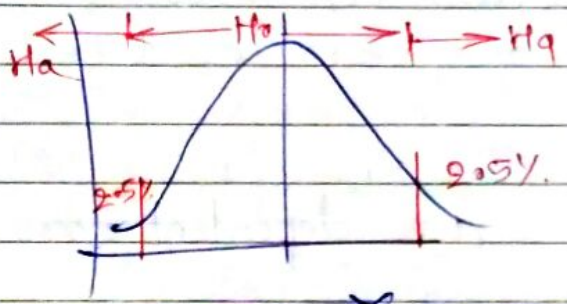
P value < α (0.05) → H_a is accepted

$<$



$P(x \leq a)$

$\neq$



Types of hypothesis tests

Univariate

Categorical

proportion

$H_0: p = 0.20$

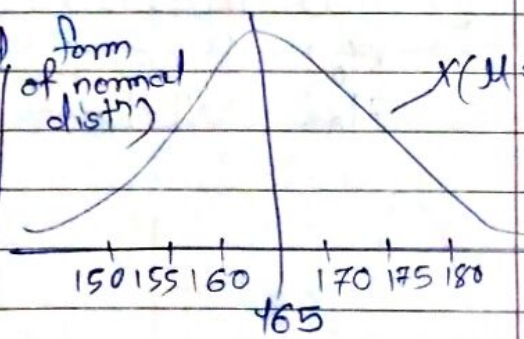
$H_a: p < 0.20$

① z-test for population proportion.

Transformation (standerdized form of normal distn)

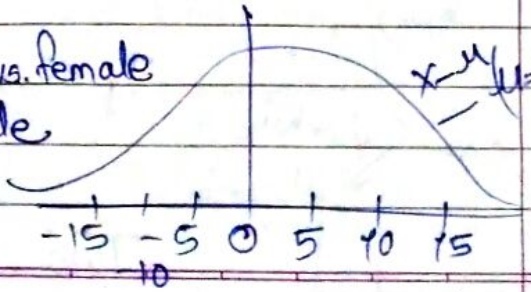
$z: x \rightarrow x - \mu \rightarrow \frac{x-\mu}{\sigma}$

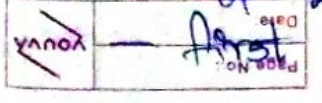
where $\mu = 0, \sigma = 1$



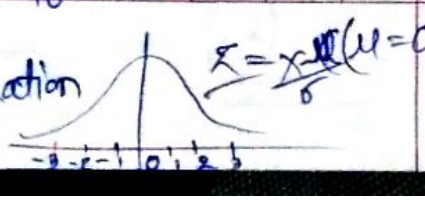
Application of z tron.

- Comparing positions.
  eg. board marks male vs. female
  height male vs. female
- Two normal distn can be compared with help of z transform.





Standerdization



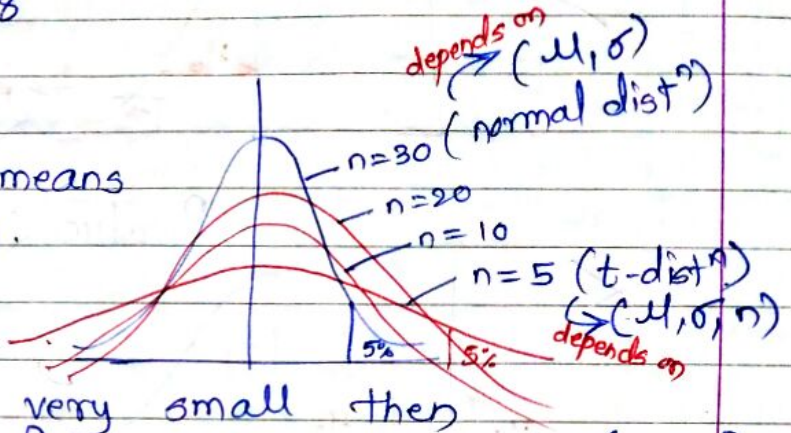
library(dplyr) MASS
dplyr
xlsx

Quatitative

$H_0: \mu_p = 0.8$

$H_a: \mu_p > 0.8$

t-distribution
- less no. of sample means uncertainty is high.



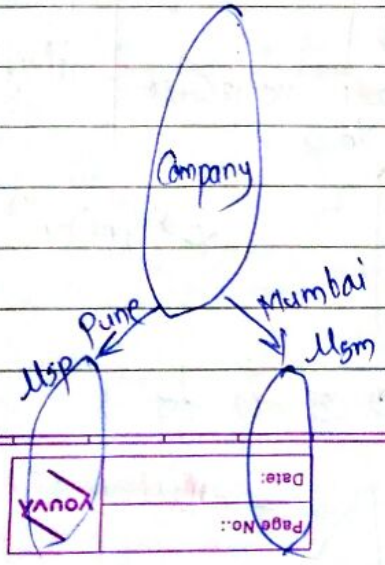
- If sample size is very small then we use t test for population mean ($n<30$)
- z test of population mean. (only t test is considered., one sample t-test known as)

Bivariate

X	Y	
C → Q		Domain claim - X and Y are dependent or not.
C → C		
Q → Q		

H_0: X and Y are independent.
H_a: X and Y are dependent.
Smoking and Cancer. } dependent ideally.
Hr. of study and marks. }

C → Q 1) Two independent samples.
H_0: Pune & Mumbai salaries are diffrent.
H_0: Wethear & Car breakdown are dependent



$H_0: \mu_{salary\,Pune} = \mu_{salary\,Mumbai}$

$\mu_{sp} - \mu_{sm} = 0$

$H_a: \mu_{sp} < \mu_{sm}$

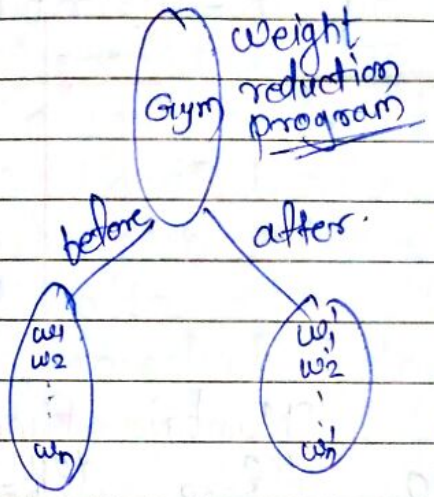
$\mu_{sp} - \mu_{sm} < 0$

Two sample t-test.

eg. p value = 0.063 no diffrence > 0.05

2) Two dependent samples - Two elements are connected with each other

eg. fitness club joining for weight loss.
- weight should be measured before and after experimentation.
- same group of people / machines / things should be under consideration before and after experimentation.
- sales befor & after training.



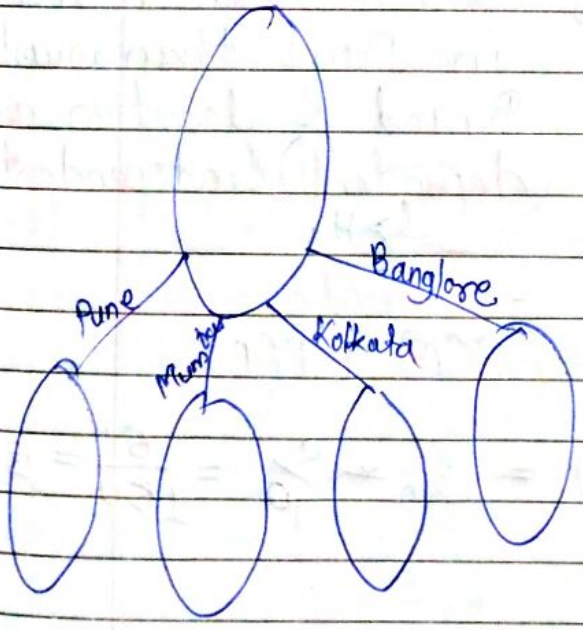
$H_0: \mu_d = 0$ paired t-test
$H_a: \mu_d > 0$

μ_d

$d_1 = w_1 - w_1'$
$d_2 = w_2 - w_2'$
⋮
$d_n = w_n - w_n'$

p-value = 0.0463
H_0 is true.

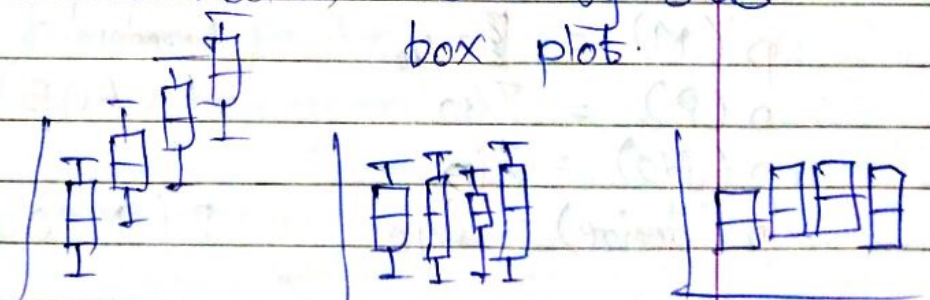
3) More than two independent sample.



$H_0: \mu_{sp} = \mu_{sm} = \mu_{sk} = \mu_{sb}$

H_a: Atleast one of the μ is diff.

Visualization : side by side box plots.



X
f test - extreme 5%

V
above 5%.

May be
check

To compare means in box plot, f test.

$$f = \frac{\text{Variability between samples}}{\text{Variability within samples}} > 1$$

F statestics > 1 should be

ANNOVA (Analysis of variance) F test.

<0.05 good. Overall model significance given by F-distribution
Individual significance given by t-distribution.

C → C

priority vs seviarity
wethear vs delay
brand vs location

Brand	Location
j10	Pune
desire	Mumbai
i	M
d	P
d	P
:	:
:	:
:	:

	Mumbai	Pune	
j10	100 E=60	100 E=140	200
desire	200 E=240	600 E=560	800
	300	700	

H_0 : desire sale more in Pune than mumbai
Brand & location are (dependent)/independent
↳ H_a

variables are independendent

$P(A \cap B) = P(A) * P(B)$

$p(M) = 3/10$
$p(P) = 7/10$
$p(j10) = 2/10$
$p(desire) = 8/10$

$P(M \cap j) = 3/10 * 2/10 = \frac{6}{100} =$

Chi Square Statestics —

$$\chi^2 = \sum \frac{[\text{Actual} - \text{Expected}]^2}{\text{Expected}}$$

$$\chi^2 = \sum \frac{[A-E]^2}{E}$$

$$= \frac{[100-60]^2}{60} + \frac{[100-140]^2}{140} + \frac{[200-240]^2}{240} + \frac{[600-560]^2}{560}$$

Chi Square Test —

Chi square distribution will have distribution in which 5% value, if is falls then they are dependent otherwise independent.

p-value

Assumptions in ppt check.

Q → Q salary - saving, salary - expenditure, experiance - salary, age - salary, stock market of day, fund value, (Time) stress - strain

$y = mx + c$ c - intercept, m = slope

Salary = m * experience + c.

Salary' = m * (exp + 1) + c

Salary' - salary = m = increment

Sales = m * advertising expences + c

$$y = B_0 + B_1 x$$

H_0 : linear relationship bet^n y and x does not exist $\beta_1 = 0$.

H_a : $\beta_1 \neq 0$

β_1 - t distribution should be considered.

↳ t-test for population slope

Judgement eg.

	Truth: Innocent	Truth: Guilty
Judge I.	✓	Type 2 (β)
G	Type 1 (α)	✓

evidence
$\alpha \downarrow \beta \uparrow$
$\alpha \uparrow \beta \downarrow$

→ More sample size more accurate judgement.

Supplier eg.

Pc.		defective
500	Acceptable →	25
↓		
100	Acceptable →	5

1) Producers' risk (α) Type 1

500	Innocent →	20
↓		
100	Guilty →	6

Power of test — $1-\beta$
β should be as less as

2) Consumers' risk (β) Type 2

500	Guilty →	30
↓		
100	Innocent →	4

AB testing ($\alpha\beta$ testing)

YAMAHA Date: Page No.:

Machine Learning

23/08/18

SDLC ⇒ Requirement - Analysis - Design - Dev (Coding) - Testing - Implementation - maintainance & support

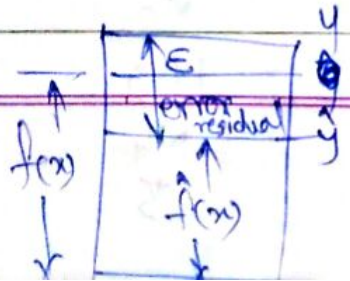
- Business needs are identified and converged to software company
- Analysis and design - break tasks into tasks modules
- Code modules done - testing - Implementation -

- Machine learns on own based on experiance without being explicitely programmed.
- If we provide more experiance to machine the performance of machine will improve.

eg Attrition (y)
- Understand (y)
- define scope of y.
- Baseline y - what is current situation
- Define causes of y. → x
- define metadata of x
- Collect data of x
- Process data - cleaning transformation
- EDA & visualization
- Inference which are important x (Remove unwanted x)
- 10th ML

$y = f(x) + \varepsilon$ - relationship bet^n x and y.
signal noise ε is error factor.

predict ↑ $\hat{y}$ ← estimate machine tries to learn best value of $f \Rightarrow \hat{f}$; $\hat{y} = \hat{f}(x)$ (hat ^)



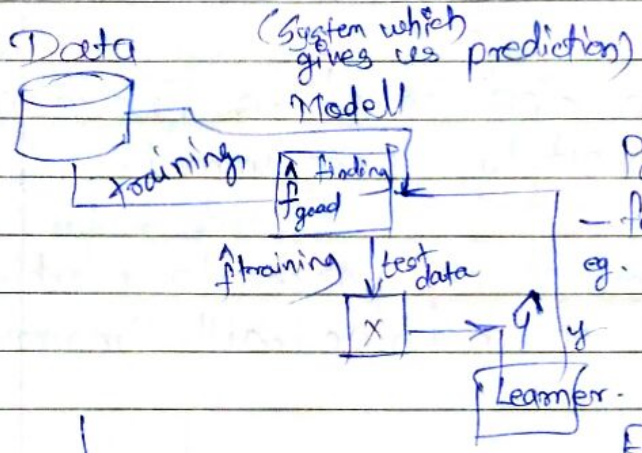
error / residual

$$y - \hat{y} = f(x) + \varepsilon - \hat{f}(x)$$

$$= [f(x) - \hat{f}(x)] + \varepsilon$$

Reducable error.

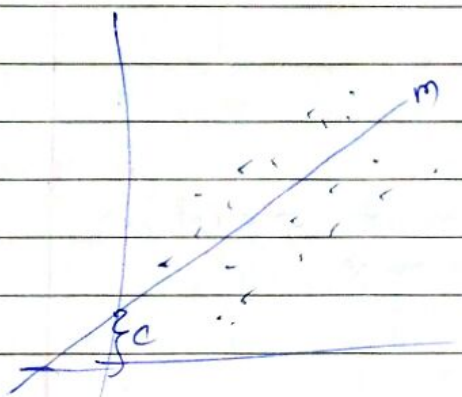
Irreducable error.



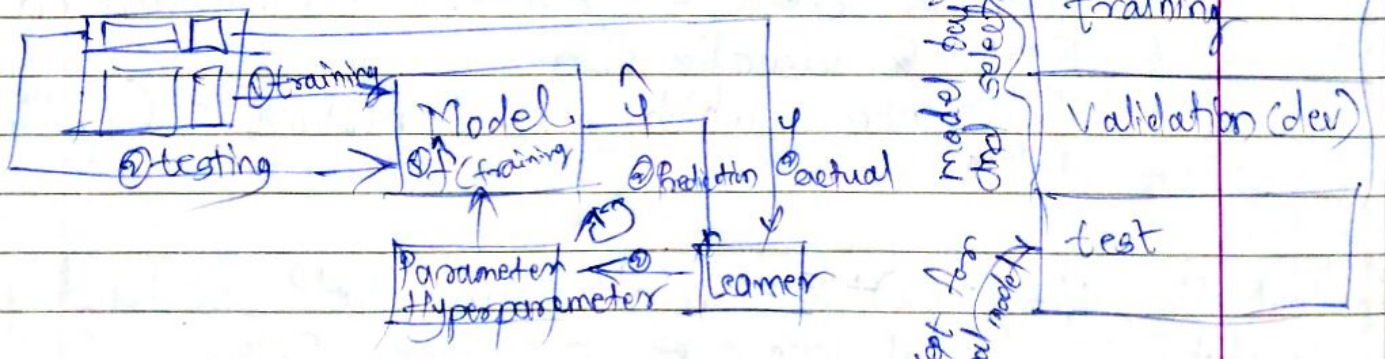
Model

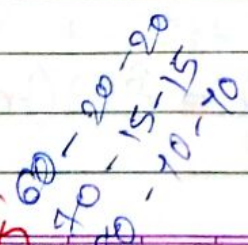
Parametric
- form of model
eg. $y = mx + c$
parameter m & c

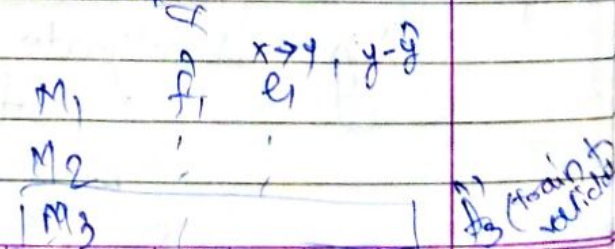
Non parametric
- form decided by algorithm itself.
('hyperparameter)



Existing data divided in two parts randomly into training and testing data.



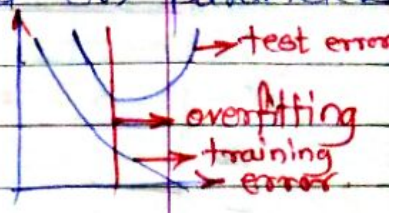




Training samples %

24/08/18

Machine learning example —
Predicting weight of person based on parameters like height, fat %, etc.

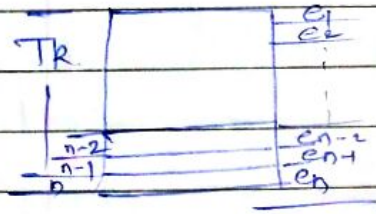


- $f(x)$ = height —①
- $f(x)$ = height + % fat —②.

- eq^n ① gives less information about weight as it only considers height.
- eq^n ② will be more optimised as it also considers % fat in function equation for predicting weight.
  - $f(x)$ = height + % fat + bone_density —③.
- eq^n ③ will be more appropriate to cosider but introducing to many parameters may lead to introduction of error in prediction (problem of overfitting)

- Cross validation approach uses complete dataset to train model and test model.
- Validation set approch uses % of data.

LOOCV — Leave one out cross validation ($k=n$)



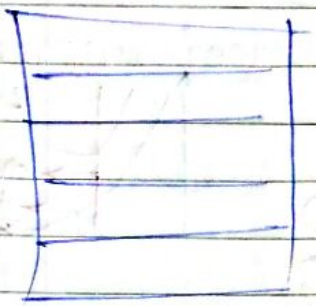
mean = e_{m1}
e_{mn}

$M_1 \rightarrow$ $e_1, e_2, \ldots, e_n$ $\}$ e_{m1}
$M_2 \rightarrow$ $e_1, e_2, \ldots, e_n$ $\}$ e_{m2}
$M_n \rightarrow$ $e_1, e_2, \ldots, e_n$ $\}$ e_{mn}

- has significant cost over computation
- any two succesive models will have almost same as majority rows are same during training such models are known as correlated models.

In correlated model predictions & errors are much similar.

K-fold cross validation.



	Tr	test
M1	1,2,3,4	5
M2	1,2,3,5	4
M3	1,2,4,5	3
M4	1,3,4,5	2
M5	2,3,4,5	1

Suggested value of k = 10

$M_1 e_{mean} = \frac{e_1 + e_2 + \cdots e_5}{5}$

$M_2 e_{mean}$

$M_n e_{mean}$.

Which ever model has e_{mean} least is selected.

Cross validation - Sampling without replacement

Bootstrap - Sampling with replacement

100 sample rows → 5 test data of 100 sample row
5×100 = 500 rows required.
duplication is allowed here

Weather Prediction – Categorical data.
Sunny (S), Cloudy (C), Rainy (R)

	Predictions	Actual
0	S	S
1	S	C
1	R	C
0	R	R
0	C	C
0	S	S
1	C	R
0	R	R
0	S	S
0	C	C

30% error.
Mis classification = 3/10

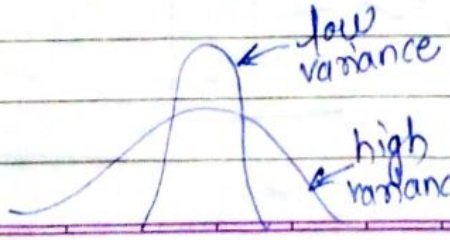
(Mean squared Error)
– Quantitive = $\frac{\sum (y-\hat{y})^2}{n}$

– Categorical has misclassification

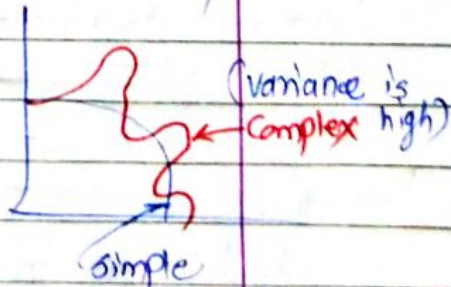
– Residual sum of squares
$= \sum (y-\hat{y})^2$

$y - \hat{y}$ = Residual = Reducable + Irreducable

Bias error
– fitting function in favour of some parameters in simple way.



Variance (σ) goes high
– Complecating function arises due to introduction of excessive parameters



– Variance should be minimum so that function will not turn over around in space.

Model Selection – Performance

Error = 1 – accuracy

- Accuracy
- Time to train & test
- Computation cost
- Scalable
- Simple
- Interpretible (One should understand model whats going on inside)

Regression

① SLR – Simple Linear Regression

$TSS = \sum (y - \bar{y})^2$

Total sum of square (More error)

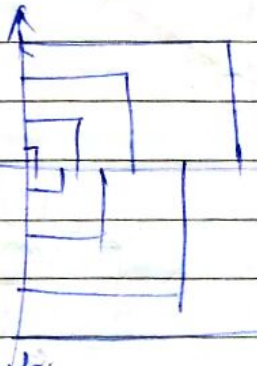
Reidual should be minimum, zero mean & constant variance

Ht	wt (y)	(Mean) $\bar{y} = 72$
152	55	17
162	63	9
170	71	1
185	84	-12
163	59	13
171	68	4
179	75	-3
190	89	-17
180	77	-5
181	80	-8

Residual sum of square

$RSS = \sum (y - \hat{y})^2$

Minimize



$TSS = (y - \bar{y})^2$

Let's say

TSS = 1000

RSS = 200

Reduction in Error / Total Error

$R^2 = \frac{\text{Reduction in error}}{\text{Total Error}}$

$R^2 = \frac{TSS - RSS}{TSS}$ – R^2 gives idea how model is good.

Greater R^2 → Good model

⑤ MLR - Multiple Linear Equation (X power is 1 and all X are additive)

$$y = \beta_0 + \beta_1 X$$

$$wt. = \beta_0 + \beta_1 * ht.$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{\sum (x-\bar{x}) * \sum (y-\bar{y})}{\sum (x-\bar{x})^2}$$

$$\frac{\partial RSS}{\partial \beta_0} = 0 \quad \& \quad \frac{\partial RSS}{\partial \beta_1} = 0$$

RSS

1000

↓

2000 - 1 predictor (ht)

↓

100 - 2

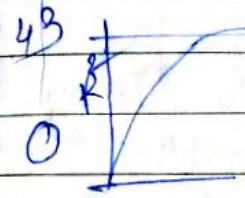
↓

50 - 3

↓

45

43



$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$wt = \beta_0 + \beta_1 * ht + \beta_2 fat$$

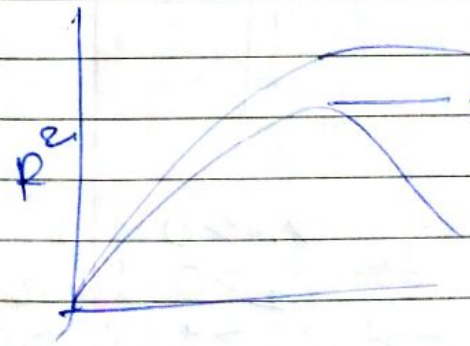
$$wt = \beta_0 + \beta_1 * ht + \beta_2 fat + \beta_3 boned^{n}$$

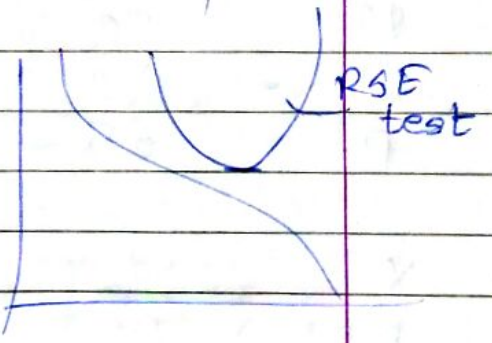
RSE - Residual Standard Error.

$$RSE = \sqrt{\frac{RSS}{n-p-1}}$$

n = no. of rows.

p = no. of predictors.





Residual Std. Error.

RSE calculation

100

10-1-1

50

10-2-1

45

10-3-1

10-4-1

① - null hypothesis independent
- alternative - dependent.

② Normality test -
Curve is normal - null hyp.
Curve is not normal - alternate hyp.

③ Col: Summary, visualization, test of hypothesis

$p < 0.05$

Page No. 43
Date
YOUVA

for project –

① Y – X Univariate analysis for all X and Y
Search for outliers.

② Y – X Summary and visualization will give appearent conclusion

③ Test of hypothesis – $p < 0.05$ dependent
$p \geq 0.05$ independent

Variable will be

28/02/18

X is categorical

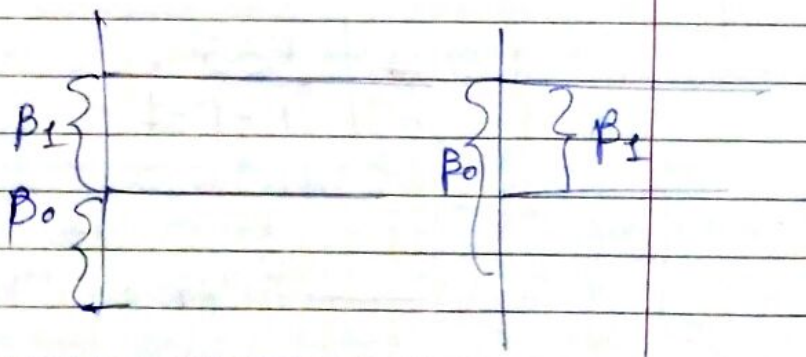
① Location (X) Salary (Y)

X	Y
P	30
M	35
P	70
M	65
P	85
M	90
M	70
M	50
P	65

$Y = \beta_0 + \beta_1 X$ X is dummy variable

Baseline $Y = \beta_0 \Rightarrow X \to 0$ Pune

$Y = \beta_0 + \beta_1 \Rightarrow X \to 1$ Mumbai



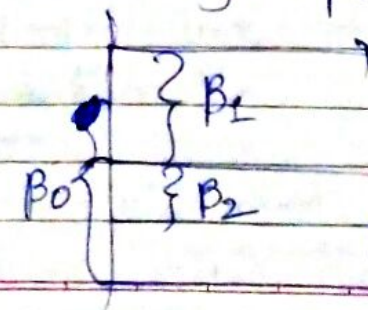
② X Y

X	Y
P	30
M	35
B	70
B	65
M	85
M	90
P	70
B	50
M	65

$Y = \beta_0 \Rightarrow X_1 \to 0, X_2 \to 0$ Pune

$Y = \beta_0 + \beta_1 \Rightarrow X_1 \to 1, X_2 \to 0$

$Y = \beta_0 + \beta_2 \Rightarrow X_1 \to 0, X_2 \to 1$



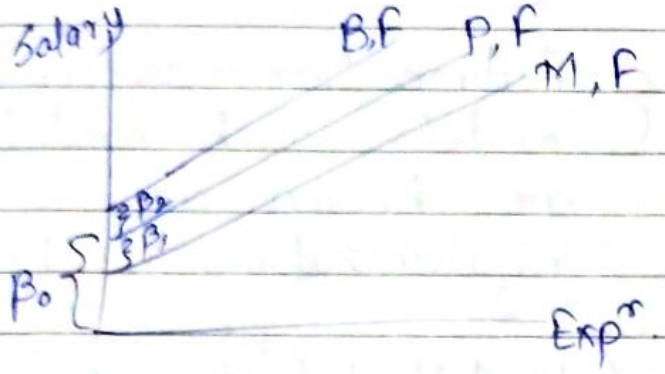
	X_1	X_2
P	0	0
M	1	0
B	0	1

$\beta_0, \beta_1, \beta_2$ eq^n.

$\beta_0 \rightarrow$ Pune, female.
$\beta_1 \rightarrow$ M, f.
$\beta_0 + \beta_2 \rightarrow$ B, F
$\beta_0 + \beta_3 \rightarrow$ P, M
$\beta_0 + \beta_1 + \beta_3 \rightarrow$ Mumbai, M
$\beta_0 + \beta_2 + \beta_3 \rightarrow$ B, M

	X_1	X_2
P	0	0
M	1	0
B	0	1

	X_3
F	0
M	1



④ Polynomial Regression.

In MLR we have eq^n.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots$$

where x have power of 1 and all terms are additive

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2^3 \cdots$$

polynomial regression. powers are >1

⑤ $$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

Interaction - eg. assembly line example where X_1 & X_2 together value add.
no changes in production { new assembly line without workers.
new workers without assembly line
- new workers on new assembly will increase production of company

Completed ⑤ models of regression

First ⑤ models minimizes RSS.

RSS - Residual Sum of squares.

↓

Irreducible ε

Reducible $(f - \hat{f})$

- Bias
- Variance

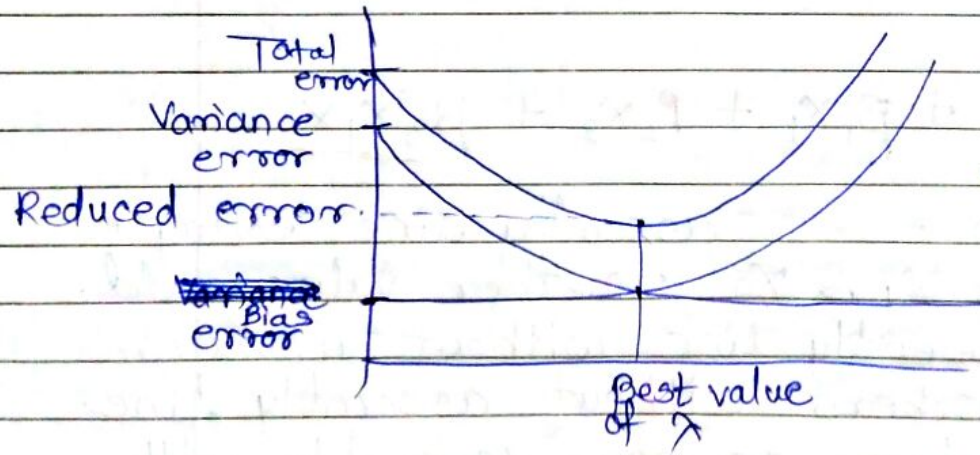
⑥ Ridge Regression ⑦ Lasso } Reduce Variance from model. such models are known as regularization methods.

- RR $\xrightarrow[\text{min}]{}$ $\left[RSS + \underset{\text{Tuning parameter}}{\lambda} \underbrace{(\beta_1^2 + \beta_2^2 + \beta_3^2 + \cdots \beta_p^2)}_{L_2 \text{ norm}}\right]$

- Lasso $\rightarrow$ $\left[RSS + \lambda \underbrace{(|\beta_1| + |\beta_2| + \cdots |\beta_p|)}_{L_1 \text{ norm}}\right]$

- Lp norm $\Rightarrow |\beta_1|^p + |\beta_2|^p + \cdots + |\beta_p|^p$

- λ is a tuning parameter.



- ε is negligible
- When variance error is more in any model this techniques are useful.

- If no. of parameters are more variance errors is more

- Imperical model. (multiple readings are taken trial & error)

is not connected with λ

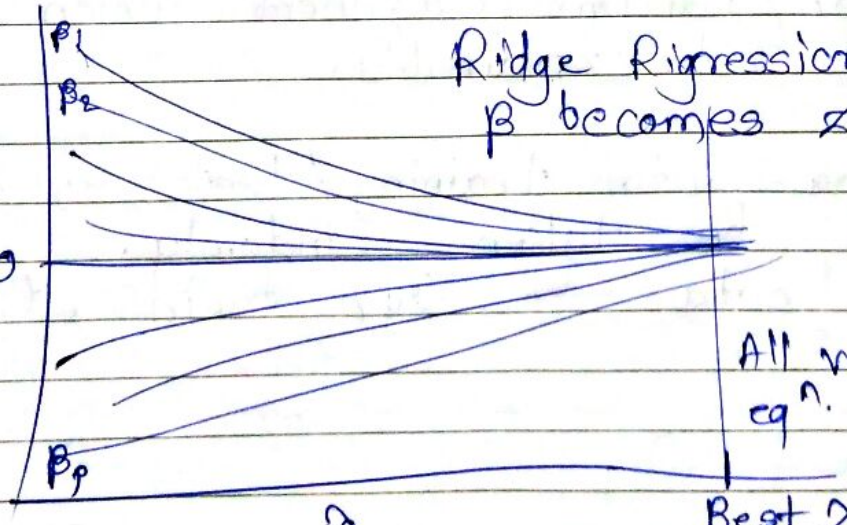
$G\beta_0$ $\lambda \uparrow$ $\beta_1 \downarrow \beta_2 \downarrow \beta_3 \downarrow \beta_4 \downarrow \cdots \beta_p \downarrow$

2
0.9
0.2
10^{-1}
10^{-2}
10^{-4}
10^{-9}

- β's are shrinking.
- Ridge Rigression & Lasso are shrinkage method
- Read RR & Lasso Theory

1) Create a vector of λ values, very small to very large.
2) for each λ
   a) Form the model. (we obtain $\beta_0, \beta_1 \cdots \beta_p$ by minimizing eq^n)
   b) Test the model. $\rightarrow y - \hat{y} \Rightarrow$ error

   continue for all λ
3) Plot error vs. λ
4) Find best λ (minimum error)
5) Find β corrosponding to best $\lambda \rightarrow$ (Best model)

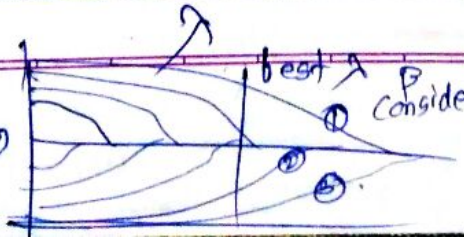


Ridge Rigression → none of β becomes zero.

β's are shrinking

All values of β in eq^n.

Best λ



Lasso

consider all β which are zero do not count in eq^n. at best value of λ

Model	Error	# of predictors.
MLR	10126	20
Ridge	9018	20
Lasso	9216	7

Choosing of model depends on Business domain
eg. card fraud - error must be minimum.
Insurance, policy - error can be tolarated

Ridge & Lasso gives improvement in error.

⑧ Time series. Y vs T for forcasting

Finding one or more value in future.

Time series has four parts.
1) Trend. - upward / downward.
2) Seasson - regular pattern. eg. electricity use monthy, weekly, quarterly; < 1yr.
3) Cycle - > 1yr trend., period not constant needs lots of data around 25-30yr
4) Error / Random component which is not explainable.

- Subsampling into training / test is complicated, can not be taken randomly.
eg. 5yr data → 4yr train, 1 yr test.

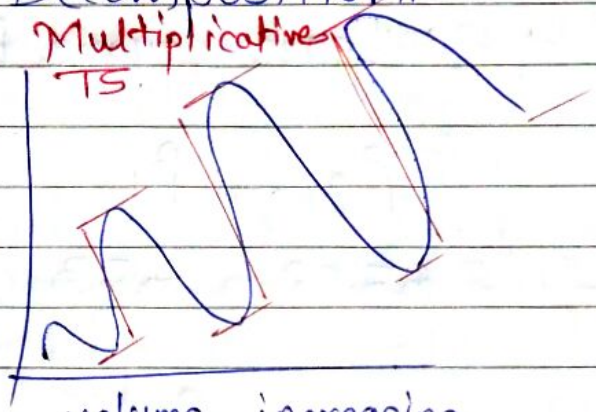
Do from ppt. - Time series.

- static & dynamic time series.

- Trend analysis. - should not contain season

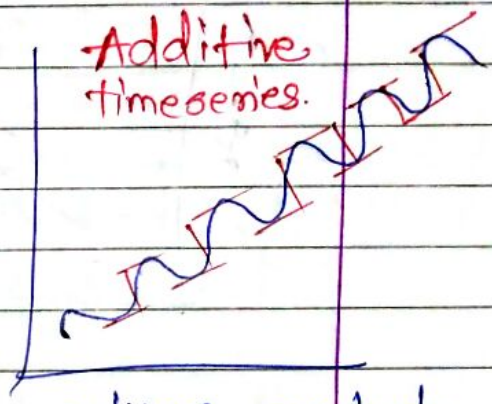
- Decomposition.

Multiplicative TS.



volume increasing

Additive timeseries.



volume constant.

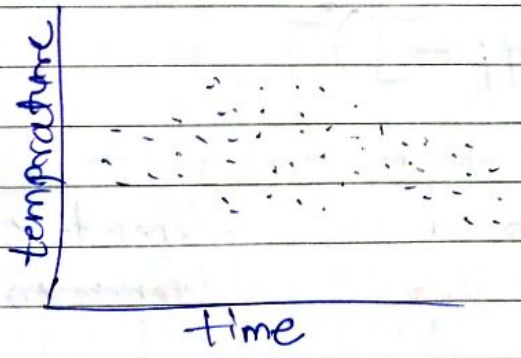
$y = trend * season * error$

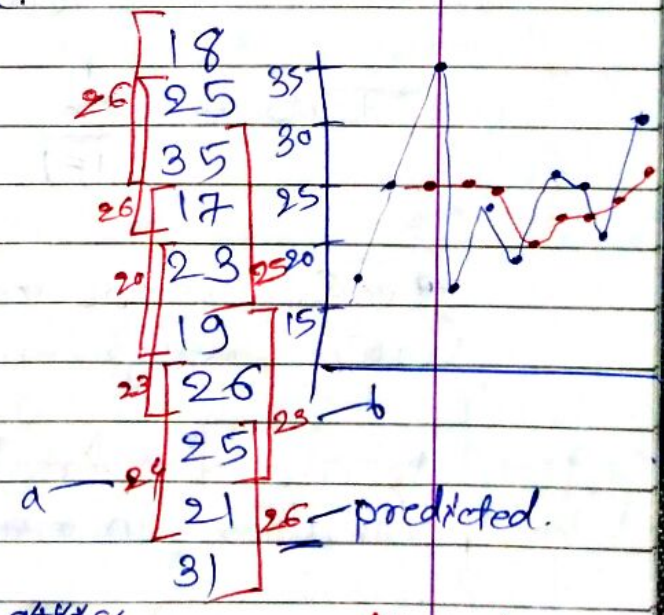
season & trend are interactive - dependent.

$y = trend * season + error$.

trend & season are independent

- Moving Average.
eg. temprature data.





Smoothing of curve - define span, eg. 3.

~~AB*C/2~~ = 26 predicted

Exponential smoothing - coefficient reduces exponentially

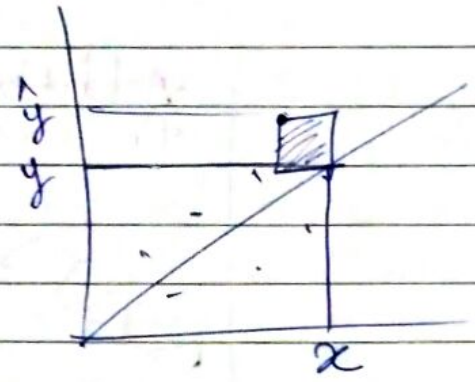
Error analysis.
Winter's method & ARIMA

30/08/18

$$RSS = \sum_{j=1}^{n} (y_i - \hat{y}_i)^2$$

$(y_i - \hat{y}_i)$ is residual

y_i ↑ Actual, $\hat{y}_i$ ↑ Predicted.



Least square Estimate

β_0	β_1
37.28	−5.3

41.1 −6.4 —
37.28 −5.03 —
33.45 −4.2 — } Confidence interval.

plot (x, y) $y \sim x$ convention

$$R^2 = \frac{\text{Reduction in error}}{\text{Total Error}}$$

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Q vs Q slope zero independent null ✓
slop not zero dependent alternate

Degree of freedom decreases as known variables increases

$$RSE = \sqrt{\frac{RSS}{n-p+1}}$$

↑ denominator is degree of freedom

RSS
Summary
plot (modelname, 1)

lm (mpg ~ wt + hp, cars)

Model	RSS	RSE	R^2	R^2 adj.
wt				
wt + hp				

1. ✓ SLR ← lm (mpg ~ wt)
2. ✓ MLR ← lm (mpg ~ wt + hp, cars)

3. Categorical. c → 8

cars $ cyl ← as.factor (cars $ cyl)
model2 ← lm (mpg ~ cyl, cars)

4. (2+3) ✓ lm (mpg ~ wt + hp + cyl, cars)

5. + Interaction term * ✓ lm (mpg ~ wt + hp + cyl + wt * hp, cars)

Identity terms allow only one power term (I)

poly
6. lm (mpg ~ hp, cars)
7. lm (mpg ~ hp + I(hp^2), cars)
8. lm (mpg ~ poly (hp, 3), cars)

$y = \beta_0 + \beta_1 hp + \beta_2 hp^2 + \beta_3 hp^3$

(5+ poly) 9. lm ✓ (mpg ~ wt + hp + cyl + wt * hp + I(hp^2), cars)

10. lm (mpg ~ ., cars)
11. ✓ lm (mpg ~ . + wt * hp + I(hp^2), cars)

RSS } Training error.
R^2 }

- min - RSE - error in model.
  max - R^2 adj. - Reduction in error

- anova (carsmd1, carsmd2, ...)

- predict (carslm 2, data.frame (wt = c(2.5, 3.5),
  ht = c (100, 200)
)
)

- Stepwise model.

Anova -

Fstat pvalue will tell how good model

p < 0.05, model value should be less than 0.05.

In anova, models should be progressively increasing, then only it can give comparitive evaluation.

Stepwise Regression—

30/08/18

y X_1 X_2 X_3 X_4 X_5

- First take SLR of
- min p value gives the most dependent variable X

$y\ X_1$
$y\ X_2$
$y\ X_3$ — 1st min. found.
$y\ X_4$
$y\ X_5$

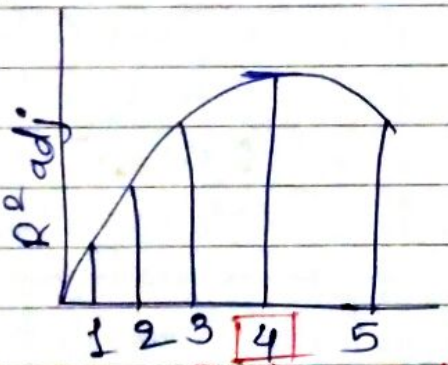
- Now pair (y, X_3) and remaining.

$y\ X_3\ X_1$
$y\ X_3+X_2$ — 2nd min.
$y\ X_3\ X_4$
$y\ X_3\ X_5$

Stepwise Regression - Forward
Backward
Mixed.

Model having R^2 adj - maximum is good.

$y\ X_3\ X_2\ X_1$
$y\ X_3\ X_2\ X_4$
$y\ X_3+X_2+X_5$ — 3rd min
$y\ X_3\ X_2\ X_5\ X_1$ — 4th min
$y\ X_3\ X_2\ X_5\ X_4$



Best suitable model

$^5C_1$ $^5C_2$ $^5C_3$ $^5C_4$ $^5C_5$

5 10 10 5 1 = 31

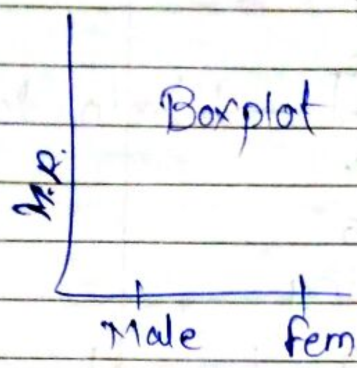
R library for stepwise regression — Search

library (leaps)
regsubsets

Variable Reduction and Model Selection.R

Project -

Regression y = monthly rate (- Attrition)



Boxplot. test - two Independent sample. pr

pr(F) more than two Ind. sam

paired t-test.
training before - after
weight loss before - after

Data preparation -
- Decide Q and C
- Datatype check. if required to change to factors.

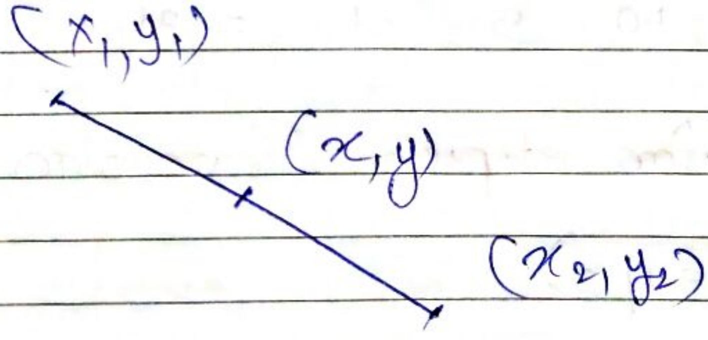
x_1 x_2 x_3 x_{100}.

β_0 β_0' β_0'' β_0'''

⋮

β_{19}

$x \uparrow \quad \beta \downarrow$



$$\frac{y - y_2}{y_1 - y_2} = \frac{x - x_2}{x_1 - x_2}$$

To find β value at certain x_x.
predict function is used.

youva
Date:
Page No.:

Classification -

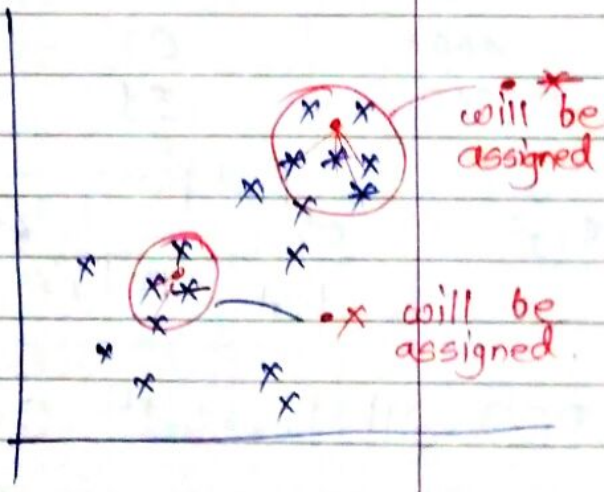
- Prediction of quantitative data is regression.
- " " qualitative " " classification

- Classification -
  Characteristics of animal. - will classify.
  Mail spam or not spam
  Employee rating.
- Based on training data classifier will classify the new data.

① KNN — K nearest neighbour.

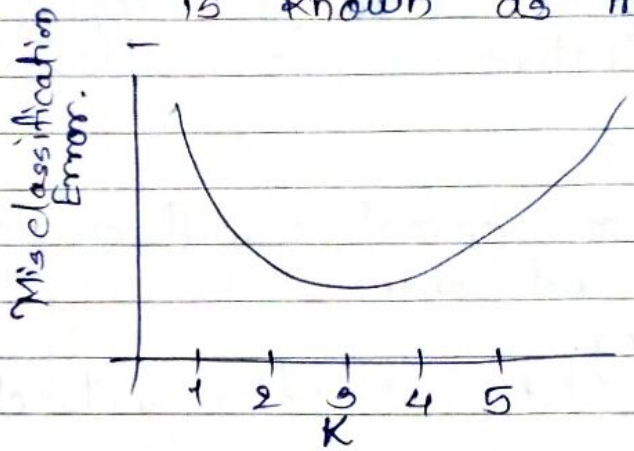
- K is user given input.
- K decides no. of neighbouring sample points
- Mode - Highest frequency.

Age	Exp.	Salary
		H
		L
		H
		L
		M
		H
		H



- Here, we can not make training & testing separately. They go in parallel.

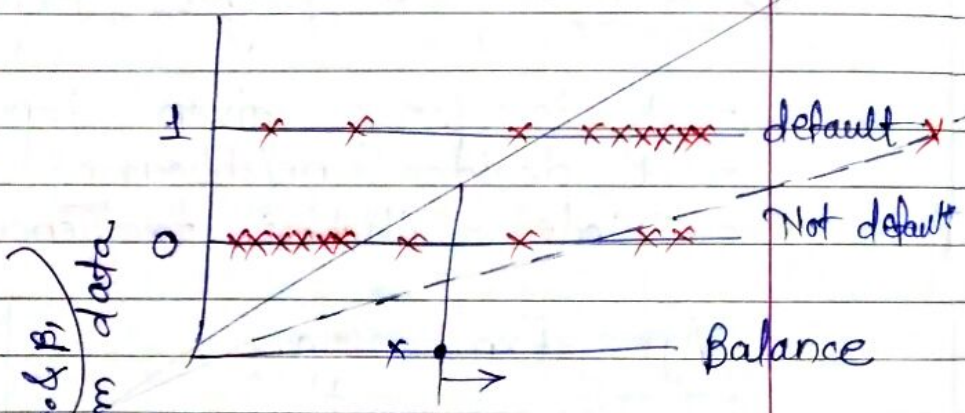
- Misclassification error
  - Any point which is not predicted correctly is known as mis-classification error.



small k - high variance
high k - low variance

② **Logistic Regression** - (Only for two classes)

X	Y
Credit Card Bill Amt.	Person is default/not
7,000	0
40,000	1
3,000	0
50,000	1

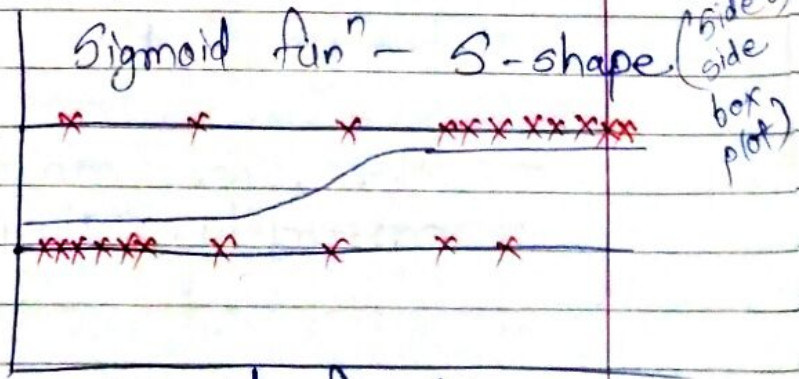


β_0 & β_1 from data

$$P(X) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

- Outliers impact the regression line hence linear regression is not useful in this scenario.

P(X) will be bet^n 0 to 1



large X - 1 den^r = 0 - Sigmoid function.
small X - 0 den^r = ∞

$$P(\text{balance}) = P(\text{default}) = \frac{1}{e^{-(\beta_0+\beta_1 X)}+1} \quad \text{OR} \quad \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

$P(Y/X)$

X Test	y	<0.5 cutoff	<0.8	<0.4
15,000	0.6	1	0	1
11,000	0.4	0	0	1
7000	0.35	0	0	0
20,000	0.63	1	0	1
25,000	0.71	1	0	1
12,000	0.45	0	0	1

$$P(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

$$\frac{P(x)}{1 - P(x)} = \frac{\frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}}{1 - \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}}$$

$$\frac{P(\text{Success})}{P(\text{failure})} = \frac{P(x)}{1 - P(x)} = e^{\beta_0 + \beta_1 x} = \text{odds}$$

$$\text{odds} = \frac{9}{4} = 2.25 = \frac{9/13}{4/13}$$

$$P(\text{Success}) = \frac{P(x)}{1 - P(x)}$$

$$\frac{9}{4} = \frac{P(x)}{1 - P(x)} \qquad \therefore P(x) = \frac{9}{13}$$

If we increment x by 1., odds will increase

$$\text{odds}' = \text{odds} * e^{\beta_1}$$

$$\text{odds}' = e^{\beta_0 + \beta_1 (x+1)}$$
$$= e^{\beta_0 + x\beta_1 + x}$$
$$= e^{\beta_0 + x} * e^{\beta_1}$$

$$y = \beta_0 + \beta_1 x$$
$$y + \beta_1 = \beta'$$

$$odds = e^{\beta_0 + \beta_1 x}$$

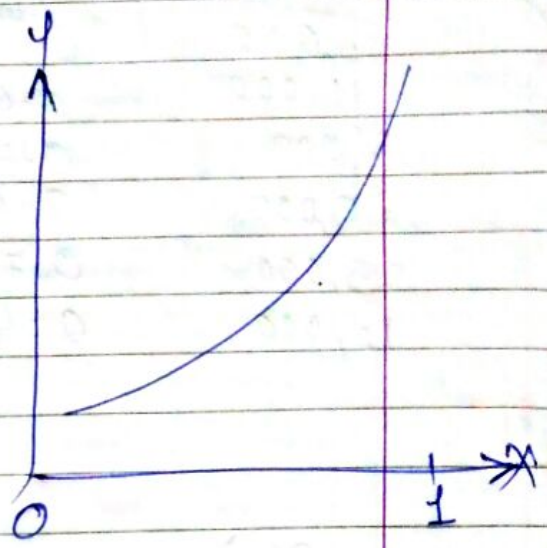
ranges between 0 to 1

$$\log(odds) = \beta_0 + \beta_1 x$$

$$\log\left(\frac{P(x)}{1-P(x)}\right) = \beta_0 + \beta_1 x$$

diff. Names
- logit
- logistic Regression

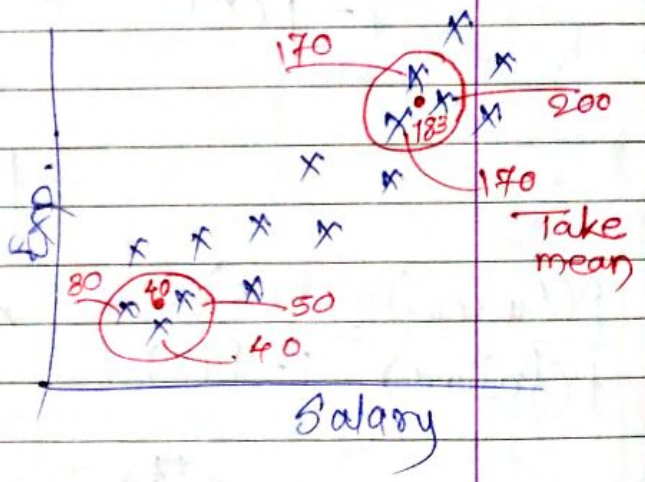
$$y = \beta_0 + \beta_1 x$$



⑤ KNN for regression —

Exp	Age	Sal(k)
32	55	200
4	25	60
:		
30	55	?

- Take $k = 3$



cl/Rg — KNN		SVM	ANN
cl — Logistic		cl	cl/Rg
cl/Reg — Decision Tree			
— Classification Tree			
— Regression Tree			
— Bagging			
— Random forest.			
cl — Naive Bayes Classification			

④ Naive Bayes Classification

Conditional Probability — Statistical / Conditional / Baysian Probability

$P(A \cap B) = P(A) * P(A/B)$ — (1)

$= P(B) * P(B/A)$ — (2).

eg. Effect of smoking on cancer.

Smoke	Non Smoke	Cances	
100	100	200	Yes
200	600	800	No.

Prospective Study → Group formed already

$P(\text{cancer}/\text{smoke})$ P(Person will have cancer when smoke)

Retrospective Study → Study continues for years.

$P(\text{smoke}/\text{cancer})$ P(Person smokes and has got cancer)

Using Bay's Theorem.

∴ from eq" (2) $P(A|B) = \frac{P(A)}{P(B)} * P(A/B)$

$P(\text{Cancer}/\text{smoke}) = \frac{P(\text{smoke})}{P(\text{cancer})} \cdot P\left(\frac{\text{smoking}}{\text{cancer}}\right)$

eg. Heart attack Prediction.

Exr.	BMI	High BP	high Cholesterol	Heart attack
Y/N	H/M/L	Y/N	Y/N	Y
N	H	Y	Y	1
Y	M	N	N	0
N	L	Y	Y	1
Y	L	N	Y	0
X_4	X_3	X_2	X_1	Y
N	H	Y	N	?

$$P(B/A) = \frac{P(B) * P(A/B)}{P(A)}$$

$$P(HA/X) = \frac{P(HA)}{P(X)} * P(X/HA)$$

Prior probability — Postirier Probality.

$P(X_{HA=y}) = P(Ex=N \cap BMI=H \cap BP=y \cap CH=N)_{HA=y}$

Assumption - Naive (simplistic)
- All X's are independent of each other.

$\therefore P(X) = P(Ex=N) * P(BMI=H) * P(BP=y) * P(CH=N)$

- All probabilities can be calculated from respective columns.

$\therefore P(X/HA=y) = P(Ex=N/HA=y) * P(BMI=H/HA=y) * P(BP=y/HA=y) * P(CH=N/HA=y)$

Laplas Correction

H	M	L	
160	40	0	= 200
161	41	1	= 203
$\frac{161}{203}$	$\frac{41}{203}$	$\frac{1}{203}$	

- In Navie Bays it is recommended to convert features to qualitative data.

eg. If it is not possible to convert data to qualitativ

BMI	HA
22	0
23	0
25	1
27	0
30	1
32	1
21	0

Seggregate 0's & 1's.

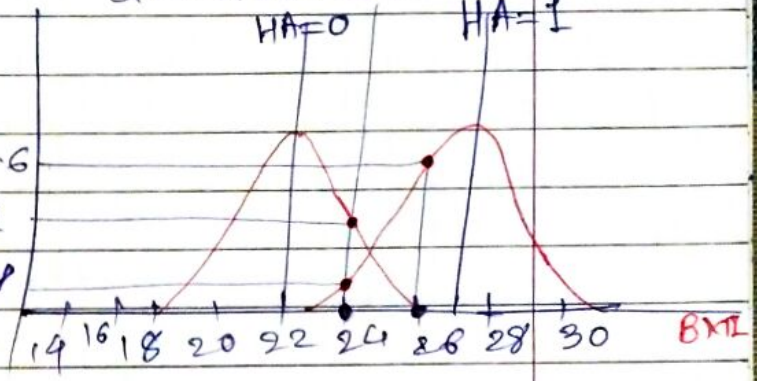
0's	1's
22	25
23	30
27	32
21	
mean = 23	mean = 29
$\sigma = 2$	$\sigma = 2$

- Assume they are normally distributed.

$P(24 / HA=1) = 0.08$

$P(24 / HA=0) = 0.2$

$P(26 / HA=1) = 0.6$

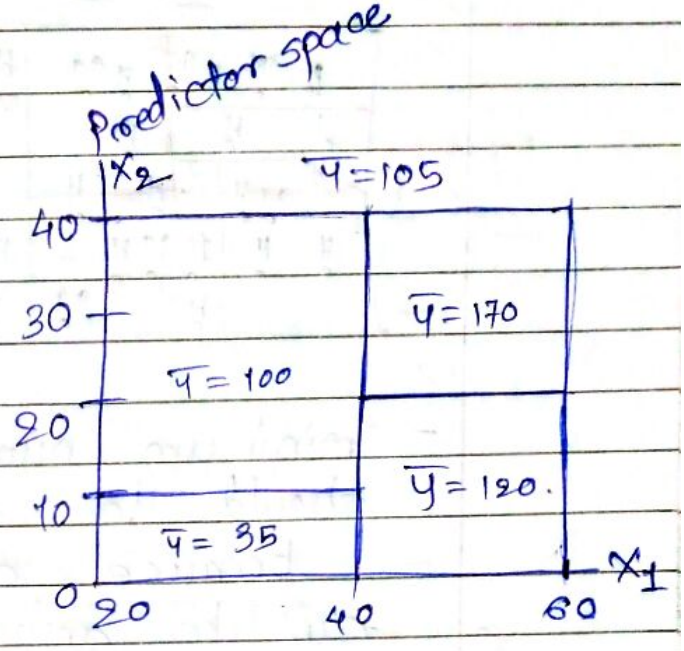


Page No.:
Date: 05/09/18

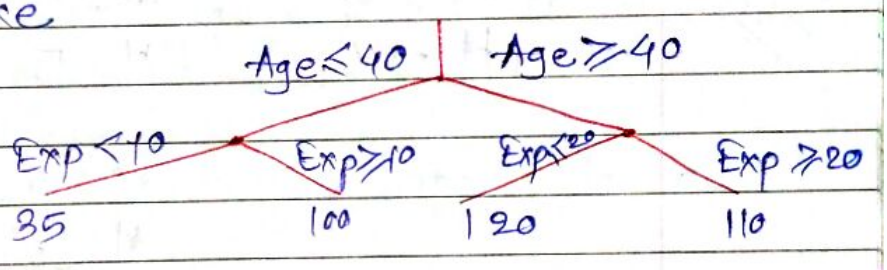
Decision Trees —

classification tree

X_2 Age	X_1 Exp^n.	y Salary(k)
22	1	30
26	3	40
55	30	180
50	27	160
38	15	100
42	19	120
35	14	? (100) from tree

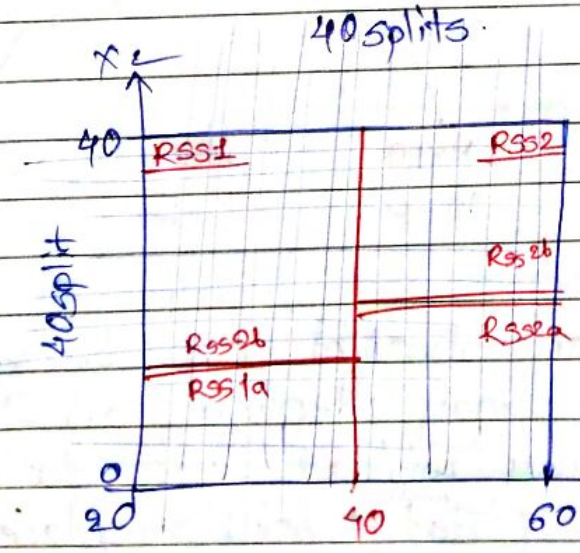


- If there's no function formulation to find relationship we take mean of data.
- $\bar{y} = 105$



- Binary split
- Branches, internal nodes, terminal node (leaf)
- Easy to understand and impliment.
- Split can come hr. or ver^n. where (RSS)$_{sum}$ is min. which will give (RSS1 + RSS2)
- The less error split will be selected

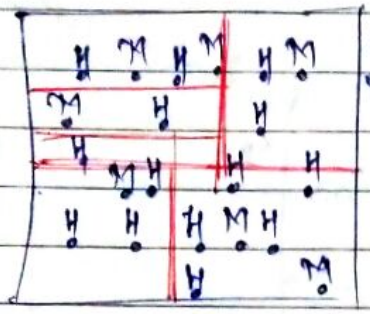
Split can take any order
- vertical
- horizontal



Hierarchy - nested if-else

- Stopping criteria - atleast minimum number of points should be included (eg. 5 pt. in box)
- $\frac{max}{min}$ ratio RSS = 5

- When salary is categorical (H, M, L)



H=6 M=4 | Mode H=14 M=7 | Mis classification = 7

H=8 M=3 | - Node purity more important

- minimum number of point in a region should be slightly high eg. 10
- because a sigle or couple of pts. can be ambiguously classified.

Regression Tree

M - 4 → M - 2 2
H - 10 → H - 5 5

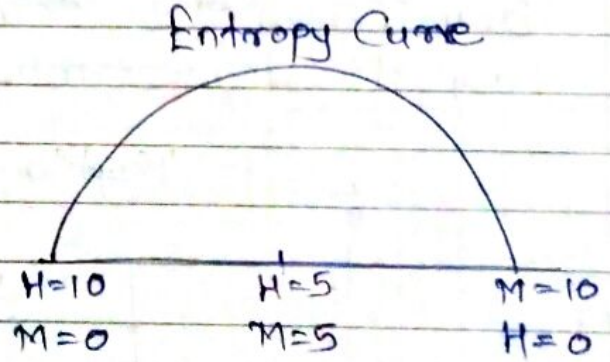
OR

M - 4	0	This split is prefereble as it gives pure node
H - 7	3	

Pure node.

- Cross Entropy
- Gini Index

- Entropy is measure of disorder.



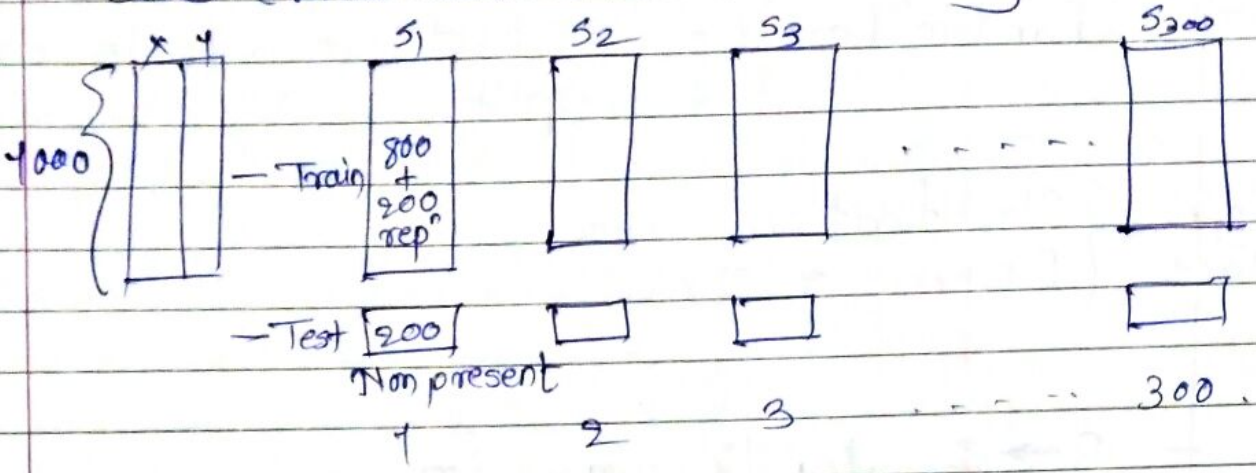
Bagging

- Tree performance might not be good.
- Tree formation may depend on train data, which may mislead test data result
- Using bootstrap (multiple data sample with replacement) is used to improve accuracy

Bootstrap aggregation (Bag of trees)

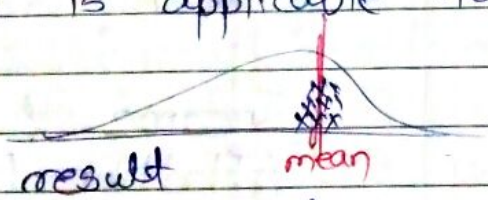
Bagging - Centre limit theorem - take mean of multiple predicted values.

OOB error estimate (Out of Bag)



25 predictors - 1 strong
3 medium
21 Weak
} Predictors

- Because of 1 strong & 3 medium predictors all 300 trees would look alike. Co-related trees.
- Central limit theorem is applicable for spreaded data.



- Bagging will not good result in such scenarios. Hence predictors should be equally important.

Random Forest

To solve above problem, take predictor subset randomly.
- Out of 25 predictor split subset of 5 randomly
- Each tree is looking diffrent hence known as Random forest

P predictors - classification $\sqrt{p}$
(Recomendation) - regression $p/3$

Strong Predictor = $\frac{\Sigma RSS}{\text{no. of occurance}}$

Multiple Tree { Bagging – All predictors in each split
Random Forest – Subset of p in each split

Single Tree { Classification Tree
Regression Tree

- S → C logistic regression
- $\mu \pm 3\sigma$ } Outlier
  1.5 IQR }
- Categorical data has no outlier.

Data Prep. Inference
- ↑ EDA + Visualization ↑ + ML.
- Matrix plot will give good relationship between X's

Income Y and other X
Attrition Y and other X

Page No.
Date: 06/09/18 YOUVA

Unsupervised Learning :- There's no Y

Employee data - each employee is object.

	F.Name	L.Name	Skill	Exp	Salary	Loc	Gender
Emp 1							
Emp 2							
⋮							

Grouping can be done on attributes.
- Clusturing is done on objects.
- Similar type of objects form clustur.
  eg. similar employee, customer, product
- eg. Library is example of having clustures of subjects.
- Clustures are formed based on similarities.
- Similarities needs to be done based on all attributes.
- dissimilarity varies between 0 & 1
  0 → similar.
  1 → dissimilar.

Dissimilarity co-efficent calculation —

	1	2	3
1	0	d_{12}	d_{13}
2	d_{21}	0	d_{23}
3	d_{31}	d_{23}	0

$d_{12} = d_{21}$ $\quad d_{ij} = d_{ji}$

↑ Useful portion.

- cont'd

Nominal data – There is no order and it is a categorical data.

	Tech.	Loc?.	Responsibility
Emp 1	Oracle	Mumbai	Coding
2	Java	Pune	Testing
3	Oracle	UK	Coding
4	R	Pune	Coding.

Emp 1 & Emp 2 $D_{12} = \frac{3}{3} = 1$

	1	2	3	4
1	0	1		
2	1	0		
3	0.33	1	0	
4	0.66	0.66	0.66	0

Ordinal data –

	Ordinal Rating	Rank	Normalized Rank (N.R.)
Emp 1	0.66 G	1 Outst.	0
2	0 Outst.	2 VG	0.33
3	0.33 VG	3 G	0.66
4	0.66 G	4 A	1

$N.R = \frac{Rank - 1}{Max\ Rank - 1}$

	1	2	3	4
1	0			
2	0.66	0		
3	0.33	0.33	0	
4	0	0.66	0.33	0

Binary Data

	Result	High Priority	Delivered
Test Case 1	P 1	Y 1	N 0
2	F 0	N 0	Y 1
3	F 0	Y 1	Y 1
4	P 1	N 0	N 0

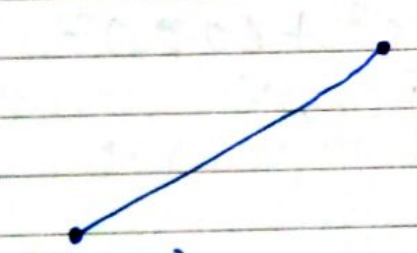
T_{23}

	0	1
0	1	0
1	1	1

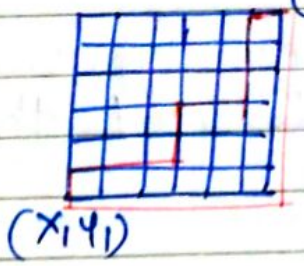
$= \frac{1}{3}$

Numeric Dissimilarity –

– Distance is calculated to find dissimilarity



$$d_e = \sqrt{(y_2 - y_1)^2 + (x_2 - x_1)^2}$$
$$= \sqrt{14^2 + 15^2}$$
$$= 14\sqrt{2}$$ Euclidian dist^n



$$d_m = |y_2 - y_1| + |x_2 - x_1|$$
$$= 14 + 15 = 29.$$
Manhatten dist^n.

45 = 0.8
15 = 0.2
30 = 0.5
5 = 0
50 = 0.9
55 = 1
36 = 0.62

Min. Max normalization – Converts value bet^n 0 & 1 proportionally

$$= \frac{\text{value} - \text{min}}{\text{Max} - \text{min}}$$

$$= \frac{\text{value} - 5}{50}$$

Ordinal data cont? – more than 2 category

As per previous

Rating	Normalised Rank
Outst^n	0
VG	0.33
G	0.66
A	1

Rank	SAL
1	H
2	M
3	L

Normalized Rank

$\frac{1-1}{3-1} = 0$

$\frac{2-1}{3-1} = 0.5$

$\frac{3-1}{3-1} = 1$

	X_1	X_2
Emp 1	O = 0	H = 0
2	VG = 0.33	M = 0.5
3	G = 0.66	H = 0
4	V = 0.33	L = 1

d_{12} (Emp 1, 2) $= \sqrt{(0.33-0)^2 + (0.5-0)^2}$

d_{23} (Emp 2, 3) $= \sqrt{(0.33-0.66)^2 + (0.5)^2}$

$= \sqrt{(0.33)^2 + (0.5)^2}$

Strategy 2 – dist^n = $\frac{\Sigma \text{ Val}}{\text{Count}}$

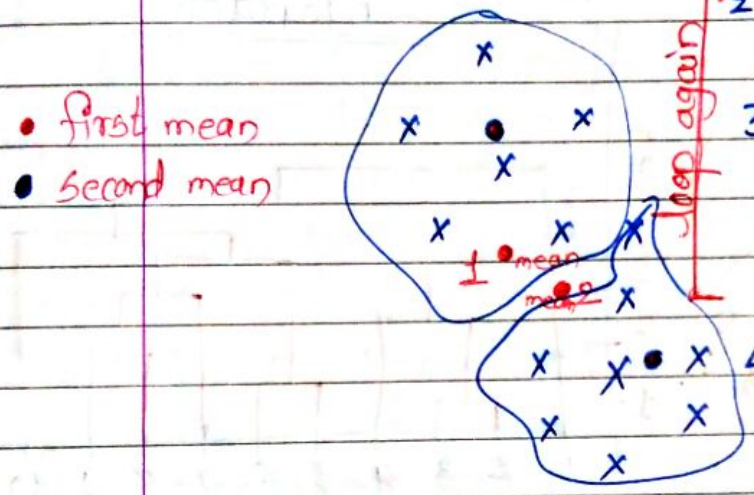
Strategy 3 – If there are many types of data first calculate separately and then take mean.

Page No.:
Date:

Clustering
- K-means
- Hierarchical
- DB Scan

Clusture properties -
- Similar pts. in same cluster
- two clusters should be dissimilar

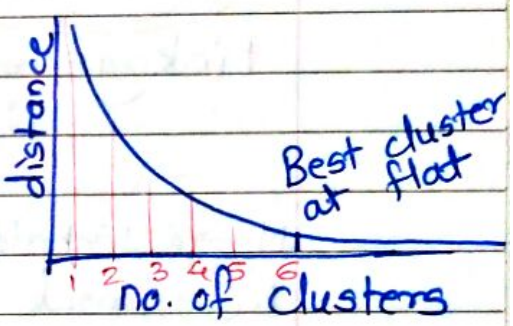
① K-means



1. Randomly assign cluster number to each point.
2. Calculate mean for each cluster.
3. Re-Assign cluster number to each pt. based on minimum distance from calculated mean.
4. When there is no change in assignment of to pts. - stop.

Within cluster
$\Sigma (dist_1)^2 + \Sigma (dist_2)^2$

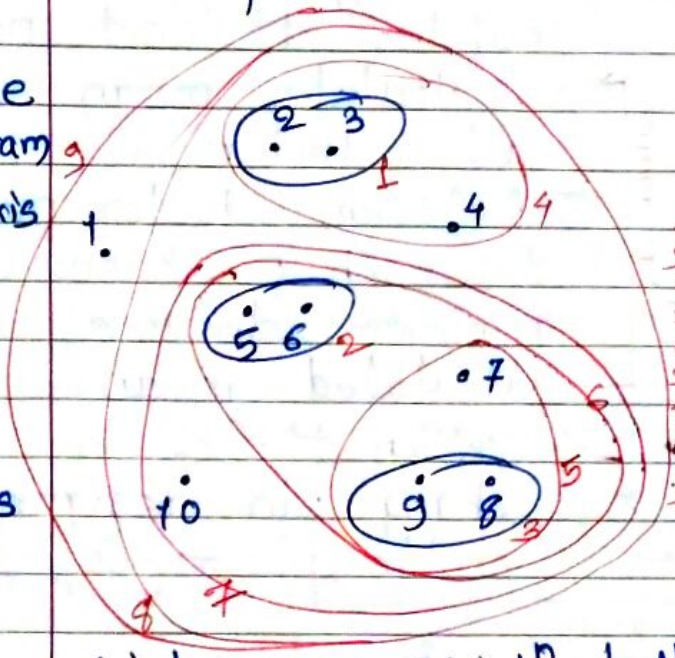
Between $= d^2$



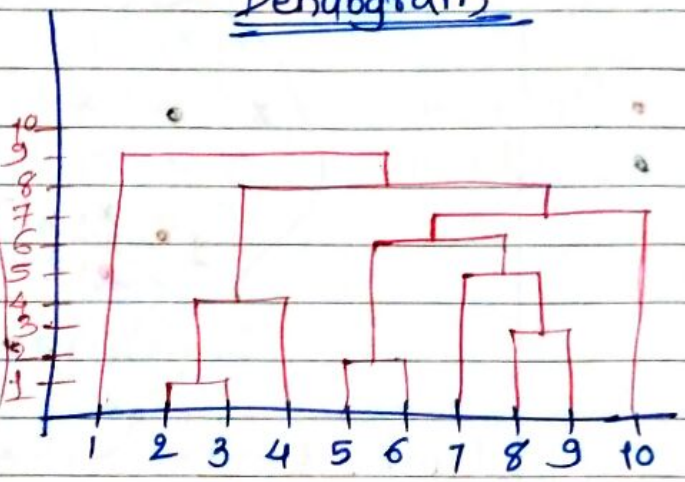
Hierarchical -

- All pts. are clustures individully at the begining
- Close ones form one cluster and repeat.

Cut the dendogram on y axis for that many no. of clusters



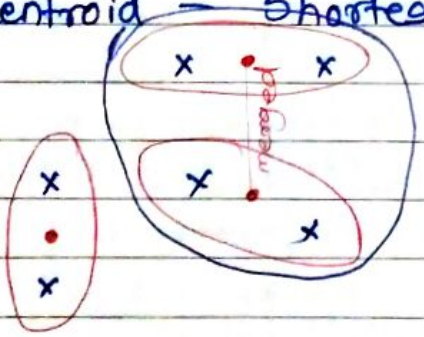
Dendogram



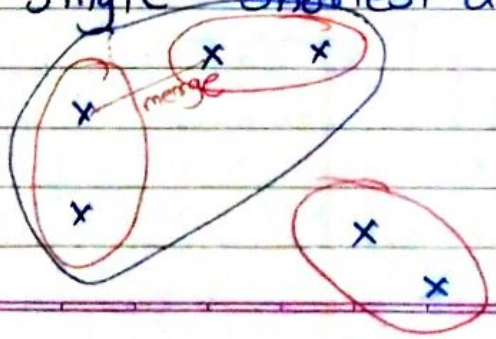
Linkage - Dist^n bet^n pt. & cluster
—u— two clusters

Merge Techniques
1. Centroid
2. Single
3. Complete
4. Average

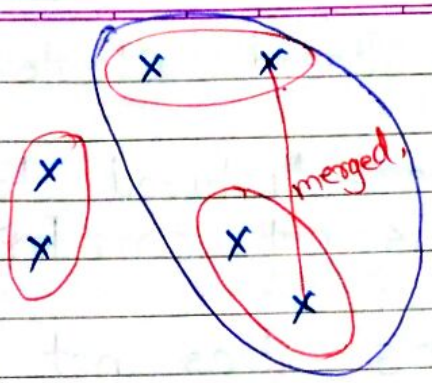
1. Centroid - ~~Shortest~~ centroid dist^n



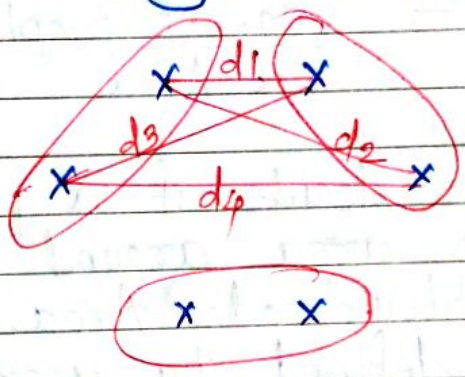
2. Single - ~~shortest~~ distance bet^n any two pts of diff. clusters



3. Complete - Longest distance

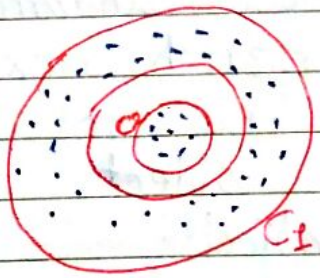


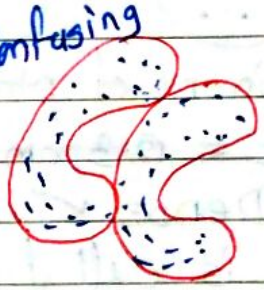
4. Average - $(d_1+d_2+d_3+d_4)/4$ minimum



Disadvantages —

- Spherical cluster - K means
  - Hierarchical
- Confusing





- Convex shap clusters are formed by K means & Hierarchical
- Outliers affects these too as it forces to include every pt.

DB Scan - Density Based Cluster.

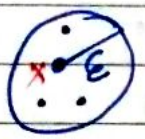
Densed points are clustured together
Sparse region is not considered.

Advantage - outliers does not affect DB scan. Thoes are left behind.

Population Density = $\frac{\text{no. of people}}{\text{Area}}$

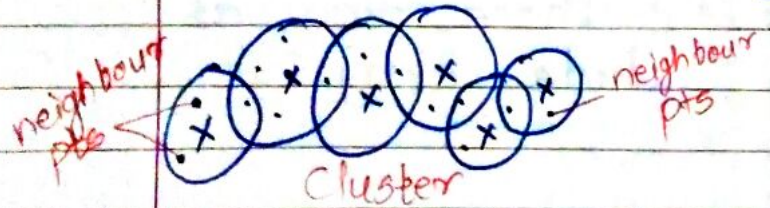
- First we need to decide if a pt. is dense.
- We consider an area around that pt.

ε neighbourhood Area
ε is defined by user
min. pts. are defined by user



min. no. of pts = 3
ε = 0.3 cm
Dense ✓
x is called core object.

- Starting core pt. is selected at random and continue to search next pt.
- Go on until you get non-core object
- Group together all neighbouring core points.
- Core objects and their neighbours form cluster.



Columnwise Grouping - Assosiation

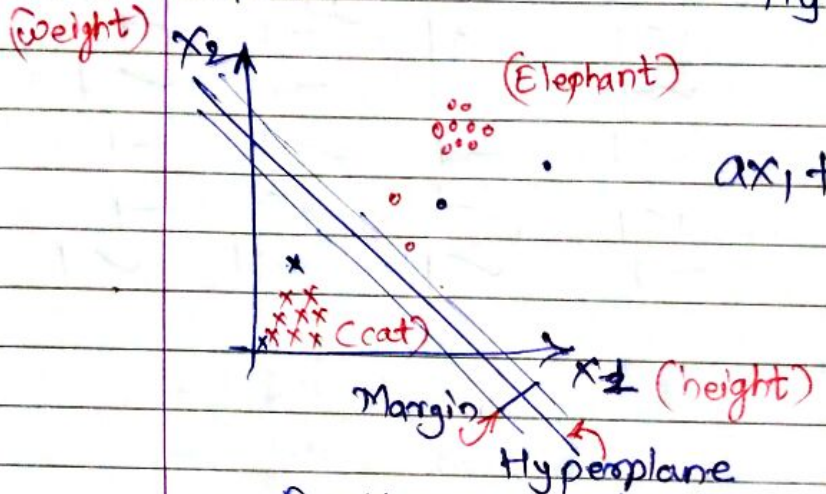
Market Basket Analysis.

Basket	bread	butter	milk	egg	cheese
1	✓1	✓1	✓1	−1	−1
2	✓1	✓1	−1	−1	✓1
3	✓1	✓1	−1	✓1	−1
4	✓1	✓1	✓1	✓1	−1
5	(−1)	✓1	✓1	✓1	−1
⋮	✓1	−1	✓1	✓1	✓1
⋮					

SVM

Page No.:
Date: 19/09/18

Hyperplane is used to segregate classes.
Two diamention - line.
Three —"— - plane.
>3 —"— - Hyperplane



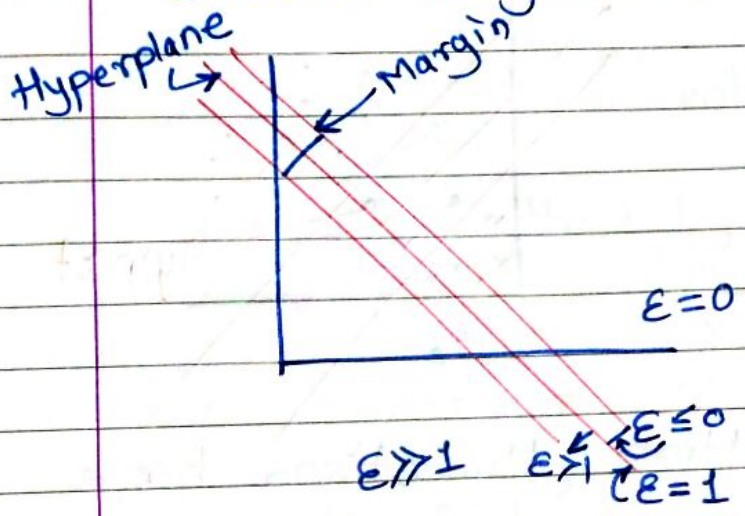
$ax_1 + bx_2 + c = 0$
< 0 x
> 0 .

- Farther points can be said confidently of some class.
- Pts. close to lines are less confident about class
- Try to fit thickest band inbetween edge pts. (outermost points)
- nearest pts to line should be farthest.
- SVM is mostly used for binary classificatio

No overlap of data pts

Maximum margin classifier. is used for perfectly linearly separable classes.
- Due to few points newly introduced margin can be varied.
- points could be messed up, there will be overlap bet^n two classes eg. cat - dog
- At the expense of some misclassification majority should be classified.

ε - Penalty / slack penalty
crossed margin but not hyperplane $0 \leq \varepsilon \leq 1$
crossed margin also hyperplane $\varepsilon > 1$



Tuning Parameter c (budget)

$\sum_i \varepsilon \leq c$

- Budget will control how many pts. will continue to cross the borders.
- Margin decision is optimization problem.

- As value of c increases margin increases.
- Model is build using training data
- Cost $\propto \frac{1}{\text{budget}}$.

cost is less, more pts. will cross.
cost is more, less pts will cross.

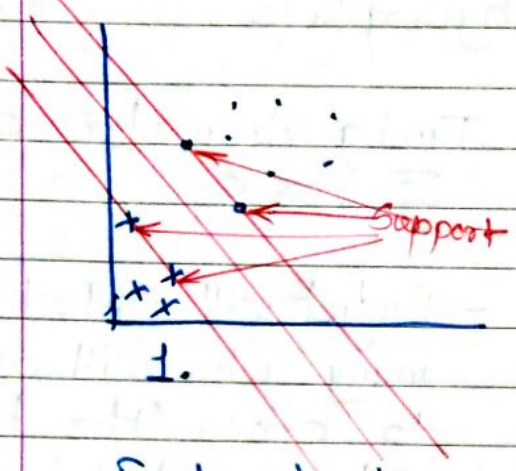
- models are tested using multiple C values on test data and chosen best

Support Vector Machines

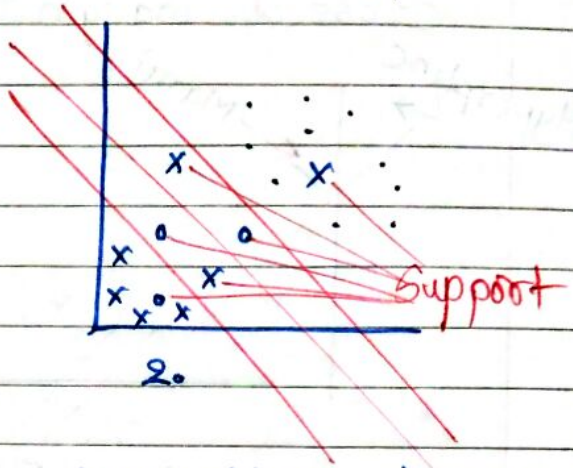
- 1. Support Vector Classifier. (eg. cat-dog)
  - overlap of data points
  - linear classifier.
- 2. Maximum Margin classifier. (eg. cat-elephant)
  - Perfectly separable classes
  - No overlap pts.
- 3. Radial Classifier (PTO) / Polynomial (SVM)
  - Non linear hyperplane
  - overlapping data pts.

Support Vector Machine also known as.

- Margin depends on nearest pts. these pts. are called as support vector

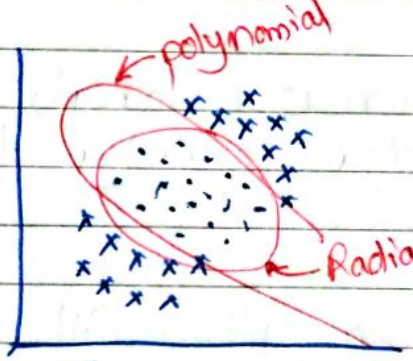


1.



2.

- Each pt. has multiple diamentions hence called as vector.



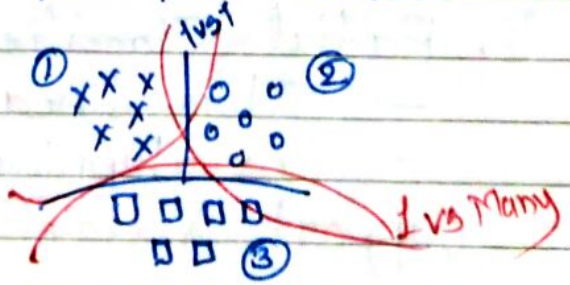
3.

- Such data will require non linear classifier.
- eg. radial classifier

- Kernel is used to simplify solution of polynomial classifiers.
- In 1. & 2. Kernel is used linear.
  3. Kernel is nonlinear

polynomial - degree
radial - gamma γ

- In case of more than two classes (eg. three classes)
  ① 1 vs 1 - take two classes at once



② 1 vs many

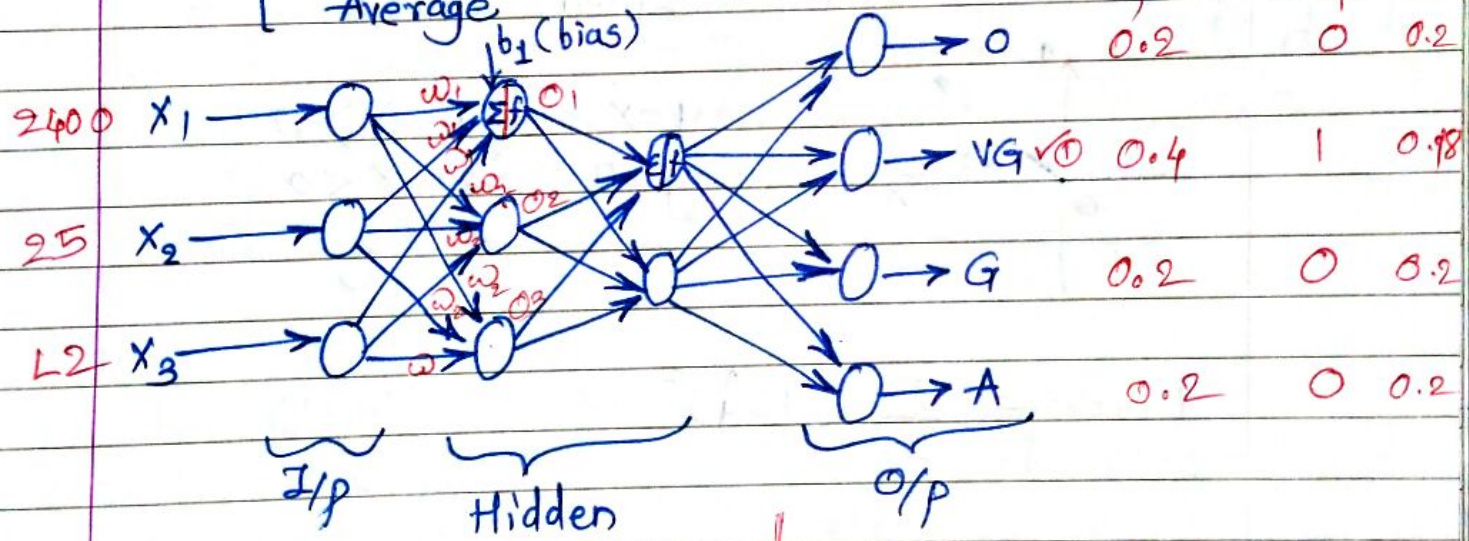
ANN

- Each neuron - processes signal independently
  - passes signal to each connected neuron
  - can enhance / surpass signal

- ANN has three layers (Input - Hidden - Output)

	I/p			O/p
	X_1	X_2	X_3	y
	Work	Defect	Skill	Performance
①	2400	25	L2	VG
②	3000	5	L1	Predict?

O/p: Outstanding, Very Good, Good, Average



	$\hat{y}$	y	error $=(y-\hat{y})^2$
	0.2	0	0.2
①	0.4	1	0.18
	0.2	0	0.2
	0.2	0	0.2

L - layered n/w
L = O/p layers + Hidden layers
= 1 + 2 = 3 layers

Similar MLR equation

$$\Sigma = w_1x_1 + w_2x_2 + w_3x_3$$

f is a function which makes signal to non-linear

$$f(b_4 + w_4O_1 + w_5O_2 + w_6O_3)$$

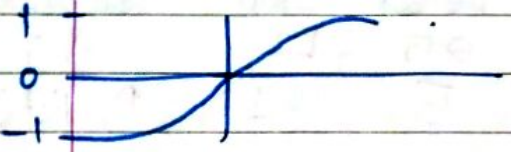
- Outputs will be probabilities.
- f is known as activation function

default function

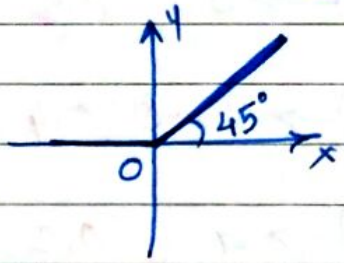
1) Sigmoid function $\frac{e^{b_1+w_1x_1+w_2x_2+w_3x_3}}{1+e^{b_1+w_1x_1+w_2x_2+w_3x_3}}$

2) Gaussian function (Normal)

3) Tan-hyperbolic (tanh)



4) ReLU (Regularized Linear Unit)



$y = x$ for $x \geq 0$

$y = 0$ for $x < 0$

$$\text{Error} = \frac{1}{2}(y - \hat{y})^2$$

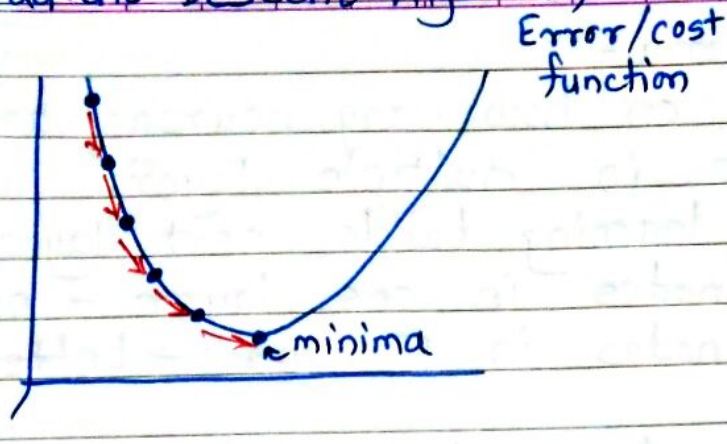
Initially all weights are initialised randomly in between -1 to 1

Forward propogation - signal

Backward propogation - Error

Page No.:
Date:

Slop Go down
Gradient Descent Algorithm



- Maximum slope is calculated by partial diffrentiation
- it will give Δw to adjust current weight so that error will minimize.

- All w's and all biases will try to minimize error.
- When there's no further change in network it is converged. it can take 1000 rows to millions rows to get train.
- learning rate is a parameter which decides convergence steps. As large learning rate might cause overshooting. and might not reach minima.

1. Classification, more than 2 classes
   no. of output nodes = no. of classes.
2. Classification, 2 classes
   no. of o/p nodes = 2
3. Regression (only summation fun?, no activation?)
   no. of o/p nodes. = 1

Page No.:
Date:
YOUVA

Design Criteria -
- Depending on how many neurons are required setup n/w in multiple layers.
Stepwise learning builds next layer.
- eg. 25 nodes in one layer - not good
5 nodes in 5 layer - better

Tensorflow by Google

Reaurrent network - Input is taken from one of the output.
(feedback n/w)

Page No.:
Date: 26/9/18

Market Basket Analysis

- What are group of products that are bought together by customers.
- Business can take decisions about products to keep in store.
- Recommender systems can be built using market basket analysis.

	Cheese	Bread	Butter	Milk	eggs	Potato	Onion
Basket 1	1	1	1	1	0	0	0
2	1	1	0	0	1	0	0
3	0	1	1	1	1	0	0
4	0	1	0	0	0	1	1
:	0	1	1	1	0	1	0
:	1	1	1	1	1	0	0
:	0	0	0	0	0	1	1
8	1	1	1	0	0	0	0
9	0	0	1	0	1	0	0
10	0	1	1	1	1	0	0

Suppose data, not accurate

Support

Support

Bread + Butter = $\frac{6}{10}$ = 60% frequent combinations are calculated in percentage

Cheese + Bread. = $\frac{4}{10}$ = 40%

Bread + Milk. = $\frac{5}{10}$ = 50%

:

Cheese + Bread + Butter = $\frac{3}{10}$ = 30%.

${}^7C_1 + {}^7C_2 + {}^7C_3 + \dots + {}^7C_7$ combinations possible

Good for Business { support ≥ min. support & confidence ≥ min. confidence & Lift > 1

- Lets assume a customer buys only at max. of 150 products in basket.
- First find combinations and frequency occured in basket.
- Bread + Butter is commonly bought together, whereas Potato - onion combination bought less together.
- With help of domain expert, we decide minimum support for combinations.
- For supperrmarkets cutoff will be smaller, for shops cutoff can be higher

Confidence

- Further we evaluate the impact of products on each other.
  eg. A+B, when A sold how many times B was sold. (A ⇒ B)

Bread + Butter = 6/10

(Bread ⇒ Butter) = 7/8 ← Butter, ← bread.

- Further based on confidence level cutoff can be decided eg. 67% or 2/3

Co-relation

Positively corelated > 1
Not corelated = 1
Negatively corelated. < 1
} Lift value

Lift = Confidence (in presence of other) / self (on own)

Lift(B) = (A ⇒ B) / B
(B in presence of A) / (B on own)

70/80 < 1 -ve

90/80 > 1 +ve

$$\text{Support}_{(A \Rightarrow B)} = \frac{n(A \cap B)}{N} = P(A \cap B)$$

$$\text{Confidence}_{(A \Rightarrow B)} = \frac{n(A \cap B)}{n_A} = P(B/A)$$

$$\text{Lift} = \frac{\text{Confidence}}{n_{B/N}} = \frac{\text{conf}(A \Rightarrow B)}{\text{sup}(B)}$$

$P(A \cap B) = P(A) * P(B)$ provided A & B independent

$$\text{Lift} = \frac{P(A \cap B)}{P(A) * P(B)} = \frac{\text{conf}(A \Rightarrow B)}{\text{sup}(B)}$$

- Group of items known as "item set"
- Support $\geq$ min. support. known as "frequent item set"
- Frequent item set mining

A-priori Algorithm (R. Srikant R. Agarwal)

A+B+C ≥ 50%.

then A+B ≥ 50%.
B+C ≥ 50%.
A+C ≥ 50%.

TID	Transaction
1	I1, I2, I5
2	I2, I4
3	I2, I3
4	I1, I2, I4
5	I1, I3
6	I2, I3
7	I1, I3
8	I1, I2, I3, I5
9	I1, I2, I3

Item	Count
I1	6
I2	7
I3	6
I4	2
I5	2

min-supp = 2
min-conf = 50%
Lift > 1.

$I_1\ I_2$ – 4
$I_1\ I_3$ – 4
~~$I_1\ I_4$ – 1~~ (minsup condition not meet)
$I_1\ I_5$ – 2
$I_2\ I_3$ – 4
$I_2\ I_4$ – 2
$I_2\ I_5$ – 2
~~$I_3\ I_4$ – 0~~
~~$I_3\ I_5$ – 1~~
~~$I_4\ I_5$ – 0~~

$I_1\ I_2\ I_3$ – 2
$I_1\ I_2\ I_5$ – 2
~~$I_1\ I_3\ I_5$~~
~~$I_2\ I_3\ I_4$~~
~~$I_2\ I_3\ I_5$~~
~~$I_2\ I_4\ I_5$~~

Check sub combinations

~~$I_1\ I_2\ I_3\ I_5$~~

min. Confidence = 50%.

2/6 ~~$I_1 \Rightarrow I_2, I_3$~~ $I_2, I_3 \Rightarrow I_1$ 2/4 ← (I_1, I_2, I_3)
2/7 ~~$I_2 \Rightarrow I_1, I_3$~~ $I_1, I_3 \Rightarrow I_2$ 2/4 ← (I_2, I_3)
2/6 ~~$I_3 \Rightarrow I_1, I_2$~~ $I_1, I_2 \Rightarrow I_3$ 2/4

		1/(6/9) = 1.5		
2/6	$I_1 \Rightarrow I_2 I_5$	1/(7/9) = 9/7	$I_2 I_5 \Rightarrow I_1$	2/2
2/7	$I_2 \Rightarrow I_1 I_5$	1/(4/9) = 2.25	$I_1 I_5 \Rightarrow I_2$	2/2
2/2	✓ $I_5 \Rightarrow I_1 I_2$	1/(4/9) = 2.25	✓ $I_1 I_2 \Rightarrow I_5$	2/4

$$\text{Lift} = \frac{\text{confidence}}{(\text{Sup.}/\text{Total})}$$

Therefore $I_5 \Rightarrow I_1 I_2$ or
$I_1 I_2 \Rightarrow I_5$ is considerable.