Knowledge
↑ → Pattern Evaluation
Data Mining
|
Data transformation
|
Data warehouse
Data Integration
|
Data cleaning
|
Data Sources
Flat files DB warehouse

KDD (Knowledge Discovery in Database)

Data Mining is a step in KDD which is called as extraction of knowledge

Data Mining Technologies:

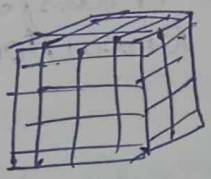
i) DBMS ii) ML iv) Statistics
iii) OLAP (Online Analytical Processing)

OLAP: (Warehouse)
> OLAP Multidimensional data i.e) cubic nature (Data cubes)
operations of OLAP → Roll up, Drill down, slice, Dice, Pivot (or) rotate

Roll up ⇒ low to Higher level
eg city → country

Drill down ⇒ Higher to lower
eg; Year → Month

Data cube:



ML → Supervised → Labels classification
ML → Unsupervised → clustering similarities

STAGES OF DATA MINING PROCESS:

1. Understand the problem & data
2. Prepare the data
3. Build the model
4. Evaluate the Result
5. Implement changes & Monitor

Data Mining Techniques :

- classification
- Regression
- clustering
- Association rule
- Anomaly Detection
- Time series Analysis
- Neural networks
- Decision Tree
- Ensemble methods
- Text Mining

Classification :

> For categorical data it is used
> It is a type of supervised

Association rule :

> If x then y

Regression : > Continuous data

Neural n/ws : > Backtracking

Ensemble : combine multiple models.

Applications of DM :

- Research
- Education Sector
- Transportation
- Market Basket Analysis
- Business Transactions
- Intrusion Detection
- Scientific Analysis
- Finance & Banking Sector
- Insurance & Healthcare

Data Preprocessing :

- Data cleaning
- Data Integration
- Data Transformation
- Data Reduction

Applications of DM :

i) Education sector :

Educational institutions use DM to analyze student performance, identify at-risk students & improve edu programs.

> Stu performance Analysis → Decision Trees or Random Forest Alg
> Personalized Learning → collaborative filtering techniques

ii) Research :

> Genomics & Bio Informatics → Hierarchical clustering
- group genes with similar patterns

> Climate modeling → Time series Analysis (ARIMA)
- Historical climate data to predict future trends

iii) Transportation :

> Traffic flow prediction → Time series Analysis (ARIMA, LSTM)
predict future conditions based on historical data

> Incident Detection → Anomaly Detection (Isolation Forest)
detect unusual patterns indicates accidents
enabling faster Response Time.

iv) Market Basket Analysis :

> Retail sales optimization → Apriori Alg
identifying frequent item sets

> Product Recommendations - Association Rule Mining
Understanding customer purchase patterns

v) Business Transactions :

> customer Segmentation → K Means clustering
- Grouping customers with similar purchasing patterns

> Credit scoring → Logistic Regression
- Assessing the credit worthiness of customers to
make informed lending decisions

> Risk Mgmt → Decision Trees
Assessing the risk level of transactions.

vii) Intrusion Detection :

> Deep Packet Inspection - Deep Learning (convolutional Neural N/W's)
- to analyze packet content and identify hidden malicious patterns.

> Hybrid Intrusion Detection systems - Ensemble learning (eg Bagging, Boosting)
- Analyze both host & n/w data

viii) Scientific Analysis :

> Astronomy → k-means clustering
- classifying stars based on their characteristics

> Physics → Principal component Analysis
- Reducing the dimensionality of particle collision

viii) Finance & Banking sector :

> Fraud Detection - Anomaly Detection (eg Autoencoders) detecting unusual transaction patterns

> Anti Money Laundering - Association Rule Mining
Identifying patterns of transactions that are indicative of money laundering, such as series of small deposits followed by large withdrawals.

ix) Insurance & Healthcare :

> Medical Image Analysis → CNN
to analyze X-Rays MRIs, CTs

> Customer retention in Insurance - Support vector machines (SVM)
Identifying factors that influence policyholder retention to improve customer loyalty.

Attribute : Field, feature, dimension

Types : Nominal, Binary, Ordinal, Discrete, Continuous
Numeric Ratio scale

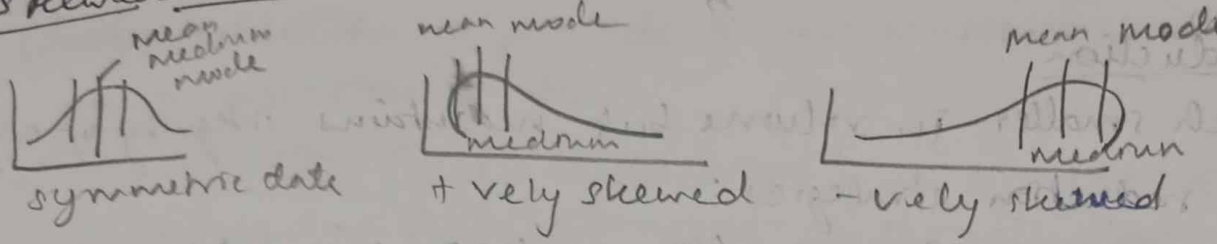
Discrete → classification → Supervised

Continuous → Regression

3 areas of statistical description :

> central tendency > Measuring dispersion of data

Skewed Data :



Major Tasks in Data Preprocessing :

i) Data cleaning → incomplete (missing), incorrect, inconsistent (noisy)

ii) Data Integration

iii) Data Reduction

iv) Data transformation & data discretization

cleaning :

Noisy → Binning, Regression < Linear / Multiple Linear, clustering

Discrepancy detection → Tools : Data scrubbing, Data Auditing

Integration :

> Schema identification prblm, Entity identification prblms

> integrating data from diff sources

> Removing Redundancy → correlation, covariance Analysis

Correlation Analysis :

> x^2 (chi-square test) $X^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$

↳ For Nominal Data

> For Numeric data → Pearson's product moment coefficient

Correlation coefficient $r_{A,B} = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum_{i=1}^{n} (a_i b_i) - n\overline{AB}}{(n-1)\sigma_A \sigma_B}$

$r_{A,B} > 0$ +vely correlated | < 0 -vely correlated | $= 0$ Independent

Correlation (linear relationship):

$$a'_k = \frac{(a_k - \text{mean}(A))}{\text{std}(A)} \qquad b'_k = \frac{(b_k - \text{mean}(B))}{\text{Std}(B)}$$

$\Rightarrow$ correlation $(A, B) = A' \cdot B'$

Correlation coefficient $\Rightarrow r_{A,B} = \frac{\text{Cov}(A,B)}{\sigma_A \sigma_B}$

$$\text{Cov}(A,B) = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n}$$

Data Reduction:

> much smaller in volume but maintains integrity of original data

> Data reduction strategies:

- Dimensionality reduction → DCA, wavelet transforms, Attribute subset selection
- Numerosity reduction → parametric & Non-parametric, sampling
- Data compression → lossy, lossless

Data Transformation:

Methods → smoothing, Attribute/feature construction, Aggregation

Normalization techniques → min-max, z-score, decimal scaling

↳ range -1 to 1 or 0 to 1

> Min-max $v' = \frac{v - \min}{\max - \min} (\text{new\_max} - \text{new\_min}) + \text{new\_min}$

> Z-score $v' = \frac{v - \mu}{\sigma}$ > Decimal-scaling $v' = \frac{v}{10^j}$

Data Generalization by Attribute Oriented Induction (AOI):

> Data Generalization → replacing relatively low level values with higher level concepts
> Concept description → form of data generalization
> concept description → generate description for data characterization & comparison. Some times called class description (specifically while objects of class usage).
> characterization → provide a concise summarization of given collection of data.
> discrimination (class comparison) → provide description by comparing two (or) more data collection.

Why AOI? → complex data types and aggregation
→ user control Vs automation

> These are the reasons for using AOI than OLAP

AOI for Data characterization:

→ proposed in 1989
→ data cube based on materialized view of data (offline aggregation)
→ AOI based on query oriented, generalization based online data analysis technique.
→ It 1st called task relevant data and perform generalization by attribute removal (or) attribute generalization.
→ Aggregation is performed by merging identical tuples.

Query: use Big-University-DB
mine characteristics as "Science Students"
in relevance to name, gender, major, birthplace, birth-date, residence, phone#, gpa
from Student
where status in "graduate"
(or) where status in {"MSc", "MA", "MBA", "PhD"}

General
Specific

Two essential operations of AOI :

or 1. attribute removal

case 1 : There is large distinct values for an attribute and there is no generalization operator

case 2 : Large distinct values and its higher level concepts are expressed in terms of other attribute.

or 2. attribute generalization :

large set of distinct values with generalization operator

approaches to control generalization

i) attribute generalization threshold control

ii) Generalized relation threshold control

eg : name - op1 , gender - op2, birthplace op1 (case 2),
DOB - op2 , major subj - op2, residence - op2,
phone no - op1 , GPA - op2

Efficient Implementation of AOI :

efficiency is analyzed as follows :

1. Relational query collect task-relevant data into working relation, efficiency depends on query processing
2. collect statistics on working relation
3. drives the prime relation

Algorithm :

1. W (DMQuery) → I/P (Data Mining queries)
2. Prepare for generalization
   * collect distinct values
   * removed (or) compute its level
3. P ← generalization (W)
   prime relation → final O/P

[consider Reg University DB
I/P → consists the names, GPA, residence, etc.
O/P → GPA, residence, DOB, etc]

Attribute oriented Induction for class comparison :

⇒ discrimination (or) comparison ↓
distinguish target class from its contrasting class

How class comparison performed ?

1. Data collection
2. Dimension relevance analysis
3. synchronous generalization
4. Presentation of derived comparison.

1. Data collected by query processing
2. select only higher relevant dimensions
3. Generalization performed on target class controlled by user (or) expert
4. Visualized in the form of tables, graphs & rules.

Example :

use Big_University_DB
mine comparison as "grad vs UG"
in relevance to name, gender, major, birth_place, birth_date, residence, phone #, gpa
for "graduate_students"
where status in "graduate"
versus "UG"
where status in "UG"
analyze count %.
from student

For ← Target class (prime relation)

Major	age	gpa	count %
SC	21-25	good	5.53
SC	26-30	good	5.02
Business	over 30	Beneath	4.68

Contrasting class → For UG

major	age	gpa	count %
SC	16-20	fair	5.53
SC	16-20	good	4.53
SC	26-30	good	2.32
Business	over 30	Ex	0.63

Statistical Measures : understand the distribution of data

- Measure of central tendency ← mean, median, mode
- Measuring dispersion ← variance, std deviation

Association Rules : If then strnt

- Market Basket Analysis motivation eg
- Frequent pattern

Representation $a \Rightarrow b$ support - together, confidence - if that then

- support(A) = $\frac{\text{No. of transactions in which A appears}}{\text{total no. of transactions}}$
- Confidence $(A \Rightarrow B) = \frac{\$(A \cup B)}{\$(A)}$

Closed Freq Itemset : Immediate superset have same support

Maximal Frequent Itemset :

For identifying frequent itemset, remove the items that are less than min support count.

Algs for Frequent Itemsets : mining freq itemsets

1) Apriori : (with candidate generation) ↔ Join step, Prune step
2) FP Growth (without candidate generation)

Apriori eg :

Tid	List of items
T1	I1, I2, I5.
T2	I2, I4
T3	I2, I3
T4	I1, I2, I4
T5	I1, I3
T6	I2, I3
T7	I1, I3
T8	I1, I2, I3, I5
T9	I1, I2, I3

⇒

At last

Items	Freq
1, 2, 3, 5	1

so we took

before Freq set

	Freq
1 2 3	- 2
1 2 5	- 2

where min support count → 2

I_{Id}	Items
101	A, C, D
102	B, C, E
103	A, B, C, E
104	B, E

min support count → 2

A - 2
B - 3
C - 3
× D - 1
E - 3

AB - 1 ×
AC - 2
AE - 1 ×
BC - 2
BE - 2
CE - 2

ABC - 1 ×
BCE - 2
ACE - 1 ×
ABE - 1 ×

BCE - 2

freq itemset

1) For $I_1\ I_2\ I_5$:

From 1st eg take Freq Itemdataset

$I_1\ I_2\ I_3$ - {2}

Generate subsets I_1 I_2 I_3 $I_1 I_2$ $I_1 I_3$ $I_2 I_3$

Associations rules :

(A) (B)

1) $I_1 \rightarrow I_2, I_3$
2) $I_2 \rightarrow I_1\ I_3$
3) $I_3 \rightarrow I_1 I_2$
4) $I_1, I_2 \Rightarrow I_3$
5) $I_1 I_3 \rightarrow I_2$
6) $I_2 I_3 \rightarrow I_1$

$I_1 \rightarrow I_2, I_3$

$S(A \cup B) = 2/9$.

$I_2 \rightarrow I_1, I_3$

$S = 2/9$.

$I_3 \rightarrow I_1, I_2$

$S = 2/9$.

(1) Support (A ∪ B) for all association rules will be same as $\frac{2}{9}$

(C) Confidence level = $\frac{\text{support}(A \cup B)}{\text{support } A}$ If $A \rightarrow B$ is association rule

For Rule 1 : $C = \frac{S(I_1 I_2 I_3)}{S(I_1)} = \frac{2/9}{6/9} = \frac{2}{6} = 33\%$.

For Rule 2 : $C =$

> Similar process for $I_1 I_2 I_3$ which is in PPT

Taking Min support count → 1, 2, 3, ... select 2
1, 3, 4, ... select 3
4, 5, 6, ... select 4

> FP Tree (Frequent pattern tree)

> To overcome the disadv of candidate gen in Apriori, FP growth is used

eg 1:

TID	Items bought
1	f a c d g i m p
2	a b c f l m o
3	b f h j o
4	b c k s p
5	a f c e l p m n

Given Min support count = 3

f - 4	l - 2
a - 3	o - 2
c - 4	e - 1
d - 1	h - 1
g - 1	b - 3
i - 1	s - 1
m - 3	k - 1
p - 3	n - 1
	j - 1

Find desc ord → f c a b m p | l o e d ...
only this to consider since count >= 3

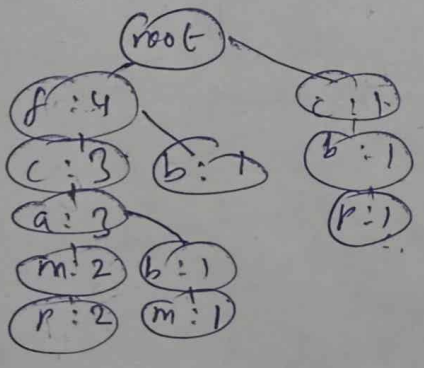
Now change transaction table (based on desc ord)

1	f c a m p
2	f c a b m
3	f b
4	c b p
5	f c a m p

(Construct tree based on these transactions from root { }: make path with child nodes & have a count which shd be same initial count)

> desc order since tree size will be lower

Final tree



⇒

Conditional pattern Base

> Write the ascending order of considered Items & fill the pattern base with the path to reach the item

Items	conditional pattern base	conditional FP tree
p	{{f, c, a, m : 2}, {c, b : 1}}	
m	{{f, c, a : 2}, {f, c, a, b : 1}}	
b	{{f, c, a : 1}, {f : 1}, {c : 1}}	
a	{{f, c : 3}}	
c	{{f : 3}}	
f	ϕ	

→ Count of item 'p' through for p's path

conditional FP tree is 3rd col

For item P => c : 2+1 => (c : 3) ✓, f : 2 ✗ m : 2 ✗ b : 1 ✗

m f, c, a : 2 + 1 => f, c, a : 3 ✓ f : 3, c : 3, a : 3

> after finding for all items

> Draw the final table for frequent pattern generated

Steps : Count → Descending order → Eliminate which < min support count

→ Construct new table → draw FP tree

→ Find conditional pattern Base → Then conditional FP tree → Final table with frequent pattern

Prblm: (✗)

(✗) Min support = 60 %. min confidence = 80 %. Apply Apriori & FP growth. (✗)

If asked to find Freq Itemset then find sets alone no need for finding confidence, support

TID	Items bought
T_1	M O N K E Y
T_2	D O N K E Y
T_3	M A K E
T_4	M U C K Y
T_5	C O O K I E

If asked for strong Association Rule find confidence & support f.s

Works

i) Apriori Alg

M - 3 ✓
O - 4 ✓
N - 2
K - 5 ✓
E - 4 ✓
Y - 3 ✓
D - 1
A - 1
U - 1
C - 2
I - 1

MO - 1
MK - 3 ✓
ME - 2
MY - 2 ✗
~~OK~~
OK - 3 ✓
OE - 3 ✓
OY - 2
KE - 4 ✓
KY - 3 ✓
EY - 2

MKY - 2
MOK - 1
MOE - 1
KYO - 2
MKE - 2
MEY - 1
MOY - 1
OKE - 3 ✓
EOY - 2

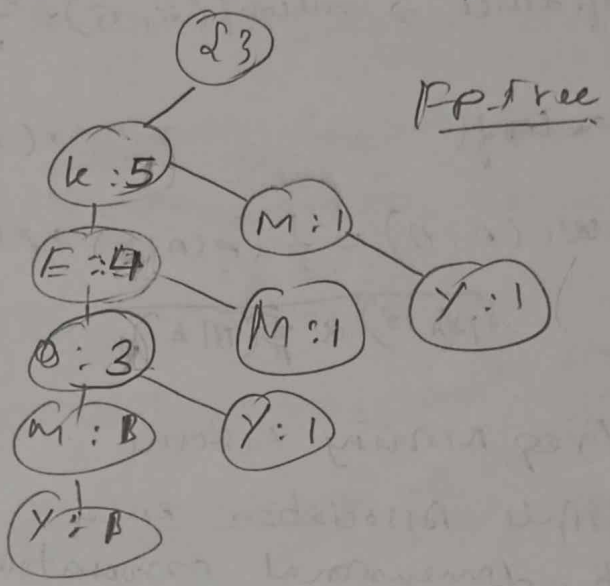
OKE ⇒ 1) O → KE
2) K → OE
3) E ⇒ OK
4) KE → O
5) OE → K
6) OK → E

1)

ii) Fp growth :

K, E, O, M, Y K-5, E-4, O-4, M-3, Y-3

T_1	K E O M Y
T_2	K E O Y
T_3	K E M
T_4	K M Y
T_5	K E O



Fp tree

> All confidence → $allconf(A,B) = \frac{sup(A \cup B)}{max(sup(A), sup(B))}$

> max conf($= max(P(A|B), P(B|A))$

> kulc $(A,B) = \frac{1}{2}(P(A|B) + P(B|A))$

> cosine : $\sqrt{P(A|B) \times P(B|A)}$

Adv Freq Mining Patterns :

> Multiple Association rules

> Multi-dimensional Association mining. (Interdimensional) ↑ no repeated predicates

Single-dimensional " " also known as intra dim

> Repeated predicates - Hybrid dimensional Asso rules.

> Static
> Dynamic — Discretization → converting numeric to nominal values

Single-dimensional association rules :

eg : buys (A, 'car') => buys (B, 'petrol')

Multi-dimensional association rules :

> With no repeated predicates => inter-dimensional asso rules

> with repeated predicates => hybrid-dimensional rules

2 approaches for handling quantitative data

→ discretization using concept hierarchy => Predefined or static (values known)

→ discretization with clustering => dynamic (unknown values)
(dynamic quantitative asso rules)

UNIT-I 5×10=50

> Plbms on Data Preprocessing cleaning & transformation
> Attributes, Distance measures, Removing noise
> Normalization min-max > Mean Median mode, chi square
> Binning, PCA, wavelet > Numerosity, dimensionality reduction
> AOI

UNIT-2 > Association rule mining < FP growth / Apriori

> For FP growth draw tree for each transaction Freq set / Strong Assc rules

Classification vs Prediction:

> For continuous data ⇒ Regression ⇒ numeric prediction
classification ⇒ 1. learning step 2. classification step

Distance Measures b/w numeric data points:

Minkowski distance

$$d(i,j) = \sqrt[h]{|x_{i1}-x_{j1}|^h + |x_{i2}-x_{j2}|^h + \dots |x_{ip}-x_{jp}|^h}$$

p-dimension h - order (user defined)

if $h=1$ means it becomes L_1 norm (or) Manhattan distance

$$d(i,j) = |x_{i1}-x_{j1}| + |x_{i2}-x_{j2}| + \dots + |x_{ip}-x_{jp}|$$

if $h=2$ means Euclidean distance

$$d(i,j) = \sqrt{|x_{i1}-x_{j1}|^2 + |x_{i2}-x_{j2}|^2 + \dots + |x_{ip}-x_{jp}|^2}$$

if $h=\infty$ supremum distance / L_{max} norm / L_∞ norm

$d(i,j) =$

eg: data set given calculate euclidean distance

x_1	12	30	15	30	19
x_2	25	40	23	45	56
x_3	38	11	35	63	42

$$d(x_1, x_2) = \sqrt{|12-25|^2 + |30-40|^2 + |15-23|^2 + |30-45|^2 + |19-56|^2}$$

$$= \sqrt{13^2 + 10^2 + 8^2 + 15^2 + 37^2}$$

$$d(1,2) = \sqrt{169 + 100 + 64 + 225 + 1369} = 43.898$$

$$d(1,3) = \sqrt{|12-38|^2 + |30-11|^2 + |15-35|^2 + |30-63|^2 + |19-42|^2}$$

$$= \sqrt{26^2 + 19^2 + 20^2 + 33^2 + 23^2}$$

$$= 55.272$$

$$d(2,3) = \sqrt{13^2 + 29^2 + 12^2 + 18^2 + 14^2} = 40.915$$

∴ Distance b/w 2,3 is lesser with 40.915

	x_1	x_2	x_3
x_1	0	43.89	55.27
x_2	43.89	0	40.9
x_3	55.27	40.9	0

Min-Max sum:

1) 13, 15, 16, 16, 19, 20, 23, 29, 35, 41, 44, 53, 62, 69, 72. Use min-max alg to solve for age 42. On the range [0.0, 1.0]

Soln:

Given $v = 42$, min = 13, new-min = 0.0, new-max = 1.0

For age = 42 $\quad v' = \frac{v - min}{max - min} \times (1.0 - 0.0) + 0.0$

$$= \frac{42 - 13}{72 - 13} \times 1 = \frac{29}{59} = 0.492$$

Age = 45 $\quad = \frac{45 - 13}{72 - 13} \times 1 = 0.5423$

Note : If v is not given in ques find min-max for all datas given

2) 200, 300, 400, 600, 1000 [0, 1]

Min-Max :

$v = 200$, $v' = \frac{200-200}{1000-200} \times (1-0) + 0 = 0$

$v = 300$, $v' = \frac{300-200}{1000-200} \times (1-0) = \frac{1}{8} = 0.125$

$v = 400$, $v' = \frac{400-200}{1000-200} = \frac{200}{800} = 0.25$

$v = 600$, $v' = \frac{600-200}{1000-200} = \frac{400}{800} = 0.5$

$v = 1000$, $v' = \frac{1000-200}{1000-200} = 1$

$v' = \frac{v - \min}{\max - \min}$ (not)

$v' = \frac{v-\mu}{\sigma}$

Z-score :

$$v' = \frac{v-\mu}{\sigma} \qquad \sigma = \sqrt{\frac{\Sigma (x-\mu)^2}{N}}$$

$$\mu = \frac{200+300+400+600+1000}{5} = \frac{2500}{5} = 500$$

$$\sigma = \sqrt{\frac{(200-500)^2}{5}} \neq \sqrt{\quad}$$

$$\sigma = \sqrt{\frac{(200-500)^2 + (300-500)^2 + (400-500)^2 + (600-500)^2 + (1000-500)^2}{5}}$$

$$= \sqrt{\frac{90000 + 40000 + 10000 + 10000 + 250000}{5}}$$

$$\sigma = 282.842$$

Let v = 550

$$v' = \frac{550 - 500}{282.84} = 0.1767$$

For sum 1) Z-score :

~~μ~~ = 35.133 ~~σ~~ = 19.94

For 42, $v' = \frac{42 - 35.133}{19.94}$

So that values stay in $\left(\frac{200}{1000}\right)$

For sum 2) Decimal-scaling = $\frac{v}{10^j}$ → If v = 200 then min j = 3

0 - 200 / 1000 = 0.2

Sums: ← UNIT-1 → Theory

> Binning Method
> Chi-square test
> Wavelet transfor
> PCA
> 3 Normalization (Min Max, Z-score, Decimal scaling)
> Distance of data points (Distance b/w each & every points) (Default : Euclidean or else specified in ques)

> KDD
> Data Mining
> Types of Attributes
> AOI → Process (Sub divisions) 3M

> Central Tendency

UNIT - 2 ⇒ Assoc rule Mining → Apriori → Pseudocode → Adv & Disadv → & sums (Mention Principles - John Paul)

FP Growth sums → Pseudocode, adv, disadv

> Solve sums using full explanation & detail
> Preprocessing how can we remove Noisy data
> Sampling ↔ with replacement / without replacement

Classification

1. One R classifier → Rule :

⇒ simple, accurate

→ select the rule with smallest total error as its one rule to construct rule

→ create frequency table

eg : dataset in ppt (P [aying tennis)

eg : Frequent table from the dataset

for → outlook	Play tennis: yes	NO
sunny	2	3
rainy	3	2
overcast	4	0
	+ve = 9	-ve = 5

Based on this eg :

⇒ If SUNNY THEN play tennis = NO

2. Decision Tree Algorithm : (ID3..) or ID3 Alg :

Decision Tree Induction :

3 parameters : i) 'D' → Data ii) Attribute list

iii) Attribute selection measures (to find which node shd be root / leaf / internal)

↳ Information Gain ⇒ or entropy $\Rightarrow Info(D) = -\sum_{i=1}^{m} P_i \log_2(P_i)$

$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$ ← particular attribute

$$\boxed{Gain(A) = Info(D) - Info_A(D)} \text{ or } \boxed{Entropy(A)}$$

Dataset before (playing tennis) set is taken

Yes = 9 NO = 5 $P_{Yes} = 9/14$ $P_{NO} = 5/14$

$$Info(D) = -\frac{9}{14}\log\frac{9}{14} - \frac{5}{14}\log\frac{5}{14} = 0.9403$$

For outlook attribute,

Sunny		Rainy		Overcast	
Y	NO	Y	NO	Y	NO
2	3	3	2	4	0

$$Info_{outlook}(D) = \frac{5}{14}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right)$$

(5 → total no. of sunny; 2 → yes of sunny; 3 → No of sunny)

$$+ \frac{5}{14}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right)$$

(for rainy)

$$+ \frac{4}{14}\left(-\frac{4}{4}\log_2\frac{4}{4}\right) = 0.693$$

(for overcast)

$$Gain(Outlook) = 0.9403 - 0.693 = 0.247$$

For temperature :

	Y	N		Y	N		Y	N
H	2	2	M	4	2	C	3	1

Tot Y = 9 Tot N = 5

$$Info(D) = 0.9403$$

$$Info_{Temp}(D) = \frac{4}{14}\left(-\frac{2}{4}\log\frac{2}{4} - \frac{2}{4}\log\frac{2}{4}\right) +$$

$$\frac{6}{14}\left(-\frac{4}{6}\log\frac{4}{6} - \frac{2}{6}\log\frac{2}{6}\right) + \frac{4}{14}\left(-\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4}\right)$$

$$= \frac{4}{14}\left(\frac{1}{2} + \frac{1}{2}\right) + \frac{6}{14}(0.39 + 0.523) + \frac{4}{14}(0.311 + 0.5)$$

$$= 0.9111$$

$$Gain(Temp) = 0.9403 - 0.911$$
$$= 0.0292$$

Humidity :

	Y	N	total
High	3	4	7
Normal	6	1	7

$$Info_{Humidity}(D) = \frac{7}{14}\left(-\frac{3}{7}\log\frac{3}{7} - \frac{4}{7}\log\frac{4}{7}\right) + \frac{7}{14}\left(-\frac{6}{7}\log\frac{6}{7} - \frac{1}{7}\log\frac{1}{7}\right)$$

$$= 0.4927 + 0.2959$$

$$= 0.7886$$

$$Gain = 0.9043 - 0.7886 = 0.1157$$

Windy:

	Y	N	
Strong	3	3	6
Weak	6	2	8
	9	5	

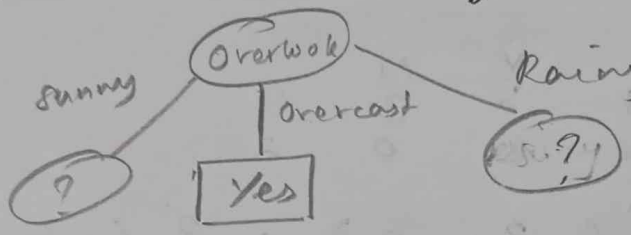
$$Info_{windy}(D) = \frac{6}{14}\left(-\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6}\right) + \frac{8}{14}\left(-\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8}\right)$$

$$= \frac{6}{14}(1) + \frac{8}{14}(0.3113 + 0.5)$$

$$= 0.8920$$

$$Gain = 0.9403 - 0.8920 = 0.0483$$

Among all gain Overlook is high so the root is Overlook



Here from the given dataset overcast has all yes so yes is leaf node of overcast but for sunny, Rainy we have to make separate tables & calculate info & gain.

Outlook	Temp	Humidity	Windy	Playtennis
sunny	Hot	High	weak	No
sunny	Hot	High	Strong	NO
sunny	Mild	High	weak	No
sunny	cool	Normal	weak	Yes
sunny	Mild	Normal	Strong	Yes
Rainy	Mild	High	weak	Yes
Rainy	Cool	normal	weak	yes
Rainy	Cool	Normal	Strong	No
Rainy	Mild	Normal	weak	Yes
Rainy	mild	High	strong	No

For sunny $N = 5$ $Y_0 = 2$ $N_0 = 3$

$$Info_{sunny}(D) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.159 + 0.133$$

$$Info_{D=sunny} = 0.292$$

For temp :

	Y	N	
H	0	2	2
M	1	1	2
C	1	0	1

$$Info_{temp}^{D} = \frac{2}{5}\left(-\frac{2}{2}\log\frac{2}{2}\right) + \frac{2}{5}\left(-\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2}\right)$$

$$+ \frac{1}{5}\left(-\frac{1}{1}\log\frac{1}{1}\right) =$$

For Humidity :

	Y	N	
H	0	3	3
N	2	0	2

$$Info_{hum}(D) = \frac{3}{5}\left(-\frac{3}{3}\log\frac{3}{3}\right) + \frac{2}{5}\left(-\frac{2}{2}\log\frac{2}{2}\right)$$

Gini - Index :

$$\text{Gini-Index}(D) = \left(1 - \sum (P_i)^2\right)$$

Pruning Tree : What, why, categories

↳ prepruning ⇒ halting its construction early

↳ postpruning ⇒ removing subtrees from a fully grown tree

Decision tree issues ⇒ Repetition & Replication

Model Evaluation :

> important for all models / Algs

Confusion Matrix :

Actual class \ Predicted class	yes	no	
yes	TP	FN	P
no	FP	TN	N
	P'	N'	

Evaluation Measures :

$$\text{Accuracy} = \frac{TP+TN}{P+N}$$

% of test set tuples that are correctly classified

$$\text{Error rate} = \frac{FP+FN}{P+N}$$ opposite to accuracy

Class imbalance prblm :

specificity, sensitivity

specificity $= TN/N$

sensitivity $= TP/P$ = Recall

$$\text{Precision} = \frac{TP}{TP+FP}$$

(Measure of exactness)

$$\text{Recall} = \text{sensitivity} = \frac{TP}{TP+FN} = \frac{TP}{P}$$

(Measure of completeness)

$$F_1 \text{ score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}$$

$$\text{Weighted } F_1 \text{ score } F_\beta = (1+\beta^2) * \frac{(\text{Precision} * \text{Recall})}{(\beta^2 * \text{Precision}) + \text{Recall}}$$

Assessing accuracy:

> Holdout > Randomsubsampling > Cross validation > Bootstrap

Holdout → training & test data | Cross validation → k-fold, N-fold

Randomsubsampling → Continuous holdout | Bootstrap → sampling with replacements

Bayesian classifier:

> statistical classifier > Bayes theorem $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$

B → evidence A → hypothesis

For buys-computer dataset:

$P(\text{buys yes}) = \frac{9}{14}$ $P(\text{buys no}) = \frac{5}{14}$

	yes	no
$P(\text{age} > 40)$	$\frac{3}{9}$	$\frac{2}{5}$
$P(\text{income-medium})$	$\frac{4}{9}$	$\frac{2}{5}$
$P(\text{student-no})$	$\frac{3}{9}$	$\frac{4}{5}$
$P(\text{credit-fair})$	$\frac{6}{9}$	$\frac{2}{5}$

→ [If for eg age > 40 No is zero/5 then it shd be $\frac{1}{6}$ & all other below Nos is like $\frac{2}{6}$ (income-medium) $\frac{4}{6}$, $\frac{2}{6}$ ... ↓ shu no ↓ credit fair

& also total buys no => $\frac{6}{15}$ yes => $\frac{9}{15}$]

$$\frac{9}{14} \times 0.032 \times \frac{5}{14} \quad (\times) \quad 0.051$$

$$0.021 \quad > \quad 0.018$$

Yes > No

∴ X = {age > 40, income-medium, student = no, credit = fair} — Yes

[If probability to zero, add +1 for every attribute count] & also add the total count in samplespace value]

→ Problem: Use Euclidean | Theory: Manhattan for eg

KNN Algorithm: (k Nearest Neighbor)

[k → no. of neighbours considered]

→ Non-parametric Algorithm
→ both classification & Regression
→ Lazy learner Algorithm

eg:

x	y	class
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

New instance (3,7)
Predict the class

distance (Manhattan b/w new instance & given x, y

Bad	4+0 = 4 (2)
Bad	4+3 = 7 (4)
Good	0+3 = 3 (1)
Good	2+3 = 5 (3)

Rank based on distance

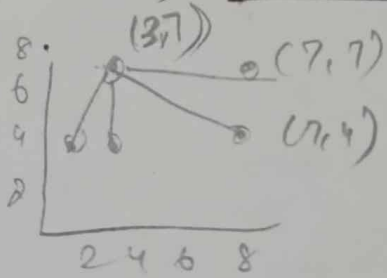
∴ if k = 3 given then (3,7) is

	class	
if k = 1	good	Majority good
k = 2	Bad	
k = 3	Good	

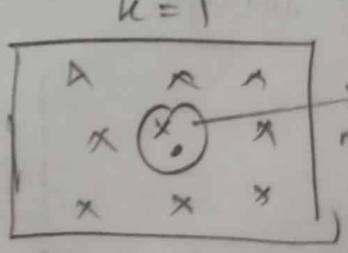
Limitations: > what k value, distance metrics
> computational complexity

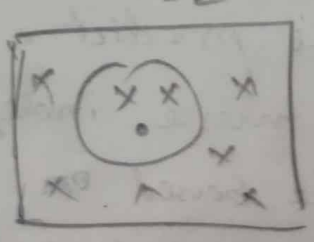
In kNN ⇒ $k = sqrt(N)$

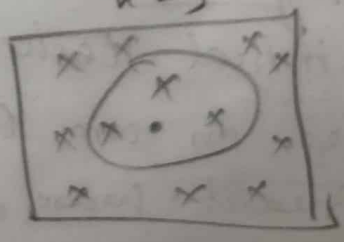
N → size of the dataset
(k shd be odd)



(Based on k value we will choose the neighbor)







(k → only odd coz
There may be same number of class
if even like 2 Good, 2 Bad)

$k = 5$ (given)

Height (cm)	Weight (kg)	Class	Distance
167	51	Underweight	$3 + 4 = 7$ (5)
182	62	Normal	$12 + 7 = 19$
176	69	N	$6 + 14 = 20$ (6)
173	64	N	$3 + 9 = 12$
172	65	N	$2 + 10 = 12$
169	56	UW	$1 + 1 = 2$ (1)
173	58	N	$3 + 3 = 6$ (4)
170	57	N	$0 + 2 = 2$ (2)
174	56	N	$4 + 1 = 5$ (3)
170	55	?	Distance

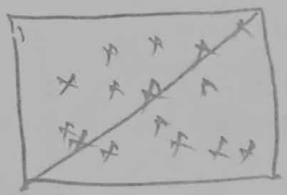
if $k = 1$ UW
$k = 2$ N
$k = 3$ N
$k = 4$ N
$k = 5$ UW

(even though for $k = 5$ UW
Majority is Normal
so for (170, 55) is Normal)

Numeric Prediction:

↳ Linear Regression

> Linear Models (Regression Model) → Under Supervised Learning
> Straight one through a set of data points.
> Straight one represents best fit prediction equation.



Linear Regression:

→ statistical technique predict a continuous values based on one (or) more independent attributes

eg: Predict house price based on living area
Predict future income of student based on college, major & GPA

Least square Method sums

1) g:

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
1	2	−2	−2	4	4
2	4	−1	0	0	1
3	5	0	1	0	0
4	4	1	0	0	1
5	5	2	1	2	4
$\bar{x}=3$	$\bar{y}=4$			6	10

$$m(\text{slope}) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2} = \frac{6}{10}$$

$y = mx + c \rightarrow$ intercept

let $y = 4$ ($\because \bar{y} = 4$) $\bar{x} = 3$

$$4 = \frac{6}{10} \times 3 + c \Rightarrow 4 = \frac{3}{5} \times 3 + c$$

$$\therefore \boxed{c = 2.2}$$

2)

	x	y
	2	50
	3	60
	5	80
	7	90
	9	95
mean	5.2	75

Predict for $x = 4$ $y = ?$

$$m = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
−3.2	−25	80	10.24
−2.2	−15	33	4.84
−0.2	5	−1	0.04
2.2	15	33	4.84
3.8	20	76	14.44
		221	34.4

$$m = \frac{221}{34.4}$$

$$m = 6.424$$

$$\bar{y} = m\bar{x} + c$$

$$75 = 6.424 \times 5.2 + c$$

$$c = 41.5930$$

$$\boxed{c = 41.6}$$

Now if $x = 4$, $y = 6.424 \times 4 + 41.6 = 67.29$

$$\boxed{y = 67.3}$$

3). Find the Regression line for

x_i	y_i	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
1	4	-3.0	-7.0	9.0	21
3	10	-1.0	-1.0	1.0	+1
5	14	1.0	3.0	1.0	3
7	16	3.0	5.0	9.00	15
~~16~~	~~44~~	~~0~~	~~0~~	20	40

$$m = \frac{40}{20} = 2.0 \Rightarrow \boxed{y = 2x + c}$$

$$\bar{y} = 2\bar{x} + c \Rightarrow 11 = 2 \times 4 + c \Rightarrow c = 3$$

$$\therefore \boxed{y = 2x + 3}$$

Linear Regression :

$$Y = \beta_0 + \beta_1 x + E$$

$$E(Y) = \beta_0 + \beta_1 x$$

[Best line for Regression shd cross through Mean value]

I) **least square error:** $\bar{y} = b_0 + b_1 \bar{x}$

⇓ Minimize least square error

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

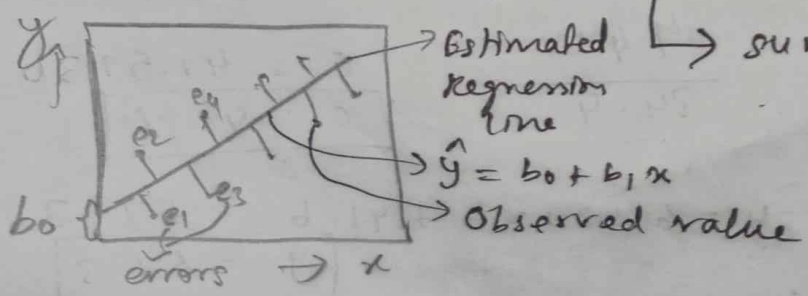
sum of squares

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\text{Slope}(m) = \frac{S_{xy}}{S_{xx}} \qquad SSE = S_{yy} - \frac{S_{xy}}{S_{xx}}$$

↳ sum of squares of errors



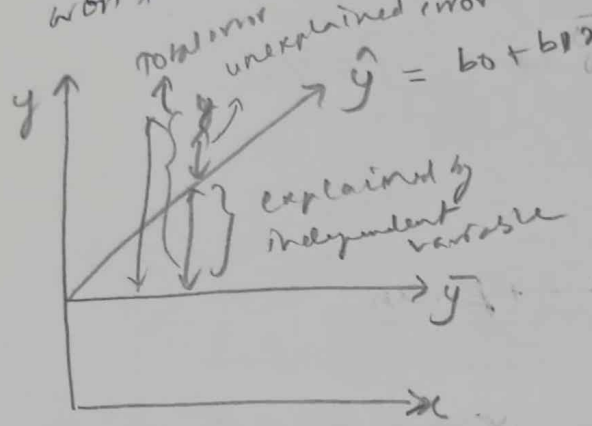
II.) R^2 - coefficient of Determination

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

SSR → sum of square of Regression
SST → Total sum of squares

$$SST = SSR + SSE$$

This is used when x value (Independent variables) won't have relationship with dependent variables.



$$R^2 = \frac{\text{Explained error}}{\text{Total error}}$$

interval b/w 0 to 1

Total error

$$\Sigma(y-\bar{y})^2 = \Sigma(\hat{y}-\bar{y})^2 + \Sigma(y-\hat{y})^2$$

Explained error, $\Sigma(\hat{y}-\bar{y})^2$

Total error, $\Sigma(y-\bar{y})^2$

Sum:

	x No. of ads	y No. of car sold		$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
1	1	14	314	-1	-6	6	1
4	3	24	72	1	4	4	1
4	2	18	36	0	-2	0	0
1	1	17	17	-1	-3	3	1
4	3	27	81	1	7	7	1
24	10	20	220			20	4

$$m = 20/4 = 5$$

$$\Rightarrow \bar{y} = m\bar{x} + c \Rightarrow 20 = 5 \times 2 + c \Rightarrow \boxed{c = 10}$$

$$\hat{y} = mx + c \Rightarrow \boxed{\hat{y} = 5x + 10}$$

$$\frac{5 \times 220 - (10 \times 70)}{5 \times 24 - 100}$$

For R^2

$(y-\bar{y})^2$	$\hat{y}$	$\hat{y}-\bar{y}$	$(\hat{y}-\bar{y})^2$
36	15	5	25
16	25	5	25
4	20	0	0
9	15	5	25
49	25	5	25
114			100

$$\left[R^2 = \frac{\Sigma(\hat{y}-\bar{y})^2}{\Sigma(y-\bar{y})^2} \right]$$

$$\therefore R^2 = \frac{100}{114} = 0.877 = 88\%$$

Multiple Linear Regression :

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

sums :

house	Size (x_1)	No. of bedroom (x_2)	Price (Y)
1	1500	3	300
2	1600	3	320
3	1700	3	340
4	2100	4	480
5	2300	4	500

Find regression line & test the significance :

1)

X (hours studied)	Y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
1	50	-2	-13	26	4
2	55	-1	-8	8	1
3	65	0	2	0	0
4	70	1	7	7	1
5	75	2	12	24	4
15	315			65	10

$$m = \frac{65}{10} = 6.5$$

$Y = 6.5x + c$ For c $\Rightarrow$ $63 = 6.5(3) + c$

$\Rightarrow c = 43.5$

$\Rightarrow \boxed{y = 6.5x + 43.5}$

if $m = 7$ is taken $\Rightarrow \boxed{\hat{y} = 42 + 7x}$ $\Rightarrow$ (Take this)

Test the significance

Null Hypothesis $H_0 : \beta_1 = 0$ (No relationship b/w x & Y)

Alternate hypothesis $H_1 : \beta_1 = 1$ (Same value of x used to predict Y)

Calculate std error of slope $(\beta_1) = \frac{S}{\sqrt{\Sigma(x_i - \bar{x})^2}}$

S → std dev of residuals

Sum of square of residuals, $SSR = \Sigma(y - \hat{y})^2$

$SSR = 1^2 + 1^2 + 2^2 + 0 + 2^2$ $\left[y - \hat{y} = 1 \ -1 \ 2 \ 0 \ -2\right]$

$= 10$

residual variance $S^2 = \frac{SSR}{n-2} = \frac{10}{5-2} = 3.33$

Std dev, $S = \sqrt{3.33} = 1.825$

$$SE(\beta_1) = \frac{1.825}{\sqrt{10}} \left(\frac{S}{\sqrt{\Sigma(x-\bar{x})^2}} \right) = 0.577$$

$$t\text{-statistic} : t = \frac{\beta_1}{SE(\beta_1)} = \frac{7}{0.577} = 12.13$$

$t \leq$ table value $\rightarrow$ accept H_0

$t >$ table value $\rightarrow$ reject H_0

$\because$ table value for $\alpha = 0.05$ is 3.182

$12.13 > 3.182$

$\therefore$ Reject H_0

2)

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$\hat{y}$	$y-\hat{y}$
2	81	-9	-11.7	+105.3	81	85.09	
4	93	-7	0.3	-2.1	49	86.78	
6	91	-5	-1.7	8.5	25	88.47	
8	97	-3	4.3	-12.9	9	90.16	
10	85	-1	-7.7	7.7	1	91.85	
12	87	1	-5.7	-5.7	1	93.54	
14	89	3	-3.7	-11.1	9	95.23	
16	99	5	6.3	31.5	25	96.92	
18	102	7	9.3	65.1	49	98.61	
20	103	9	10.3	92.7	81	100.308	
11	92.7			279	330		

$$(\beta_1) = \frac{279}{330} = 0.8454$$

$$\bar{y} = \beta_1 \bar{x} + c \Rightarrow 92.7 = 0.8454 \times 11 + c$$

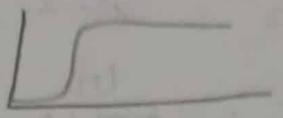
$$\boxed{c = 83.4}$$

$$\boxed{\hat{y} = 0.8454x + 83.4}$$

Logistic Regression :

odds = $\frac{P}{1-P}$; Sigmoid fun.

Statistical ~~model~~ method used to model the relationship b/w one (or) more Independent variable and many dependent variable.



[Sigmoid func $\hat{P} = \frac{1}{1+e^{-z}}$]

Binary classification :

eg : Predicting whether patient has a disease or not whether loan will be approved.

$z = \beta_0 + \beta_1 x_1 + \cdots \beta_n x_n$

Sum: Calculate the probability if a student studied 34 hrs he is pass or fail?

hrs x	Pass / Fail y
29	0
15	0
33	1
28	1
39	1

0 - 0.3 - 0.5 : 0.5

0 - 0.5

-0.5

-0.5

+0.5

+0.5

Probility prediction $\hat{p} = \frac{1}{1+e^{-z}}$

Soln 1 (Logistic regression)

Given $x = 34$ hrs

$odd(z) = -64 + 2 * hrs + 0.3$

$\beta_0 = 0.5$

$odd(z) = -64 + 2 \times 34$

$= -4$

$\beta_0 = 0.5$

$\beta_1 = 28$

Ques Final $\hat{p}$ will be given like 95%. Find the hours to get 95%.)

$\therefore \hat{p} = \frac{1}{1+e^{-4}} \Rightarrow \frac{1}{1+0.0183}$

$\therefore \boxed{\hat{p} = 0.98}$ (98% pass percentage)

Soln 2 (Linear Regression):

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
29	0	0.2	-0.6	0	-0.12
15	0	-13.8	-0.6		8.28
33	1	4.2	0.4		1.68
28	1	-0.8	0.4		-0.32
39	1	10.2	0.4		4.08
$\bar{x}$ 28.8	$\bar{y}$ 0.6				13.6

type 2 sum: $\hat{p} = 95\% = 0.95$ (given) hrs = ?

$$\frac{1}{1+e^{-z}} = 0.95 \Rightarrow \frac{1}{0.95} = 1+e^{-z}$$

$$\Rightarrow 0.052 = e^{-z} \Rightarrow \ln(0.052) = \ln(e^{-z})$$

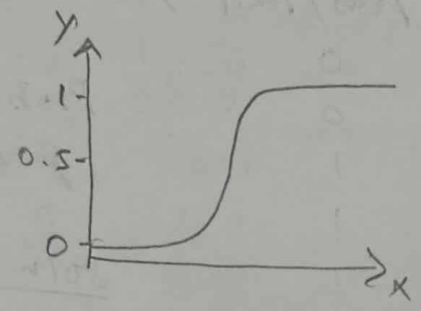
$$\Rightarrow -z = \ln(0.052) \Rightarrow \boxed{z = 2.95}$$

$\Rightarrow$ given $z = -64 + 2 * hrs$

$$\Rightarrow \frac{2.95 + 64}{2} = hrs \Rightarrow \boxed{hrs = 33.475 \text{ hrs}}$$

33.475 hrs required to get minimum 95% probability

Logistic regression:



For theory

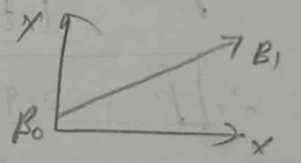
Derivation of sigmoid func:

logit (odds) = $\frac{\text{Probability of an event happening}}{\text{Probability of an event not happening}}$

↓

odds of success

logit function ← odds $(\theta) = \frac{p}{1-p}$



$$Y = \beta_0 + \beta_1 x$$

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

Take log on odds

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

exponentiating both sides

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

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

Let $y = e^{\beta_0+\beta_1 x}$

Then $\frac{P(x)}{1-P(x)} = y$

i.e) $P(x) = y(1-P(x))$

$P(x) + y(P(x)) = y \Rightarrow P(x) = \frac{y}{1+y}$

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

(÷ by $e^{\beta_0+\beta_1 x}$ on numerator & denom)

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

Sigmoid func i.e $P(x) = \frac{1}{1+e^{-z}}$

Multiple Logistic Regression :

Consider Dataset

Age (x_1)	BMI (x_2)	Blood Pressure (x_3)	Diabetes (Y)
45	28	140	1
50	25	135	0
40	30	150	1
35	24	125	0

Given,

$\beta_0 = 0.67$ $\beta_2 = 0.47$

$\beta_1 = 0.58$ $\beta_3 = 0.36$

Predict for Age 25 BMI 20 BP 120 Diabetes = ?

Given $x_1 = 25$ $x_2 = 20$, $x_3 = 120$

$$Z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

$\therefore Z = 0.67 + 0.58 \times 25 + 0.47 \times 20 + 0.36 \times 120$

$$Z = 67.77$$

$$\hat{P} = \frac{1}{1+e^{-z}} = \frac{1}{1+0} = 1$$

∴ He is diabetic for given values

Logistic Regression:

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

Use Maximum Likelihood Estimates for estimating the coefficient

Likelihood function:

Product of individual probabilities

$$L(\beta_0,\beta_1) = \prod_{i=0}^{n} P(Y_i=1/X_i)^{Y_i} \times (1-P(Y_i=1/X_i))^{1-Y_i}$$

Y_i – actual outcome

log likelihood

$$l(\beta_0,\beta_1) = \sum_{i=1}^{n} [Y_i \log P(Y_i=1/X_i + (1-Y_i) \log (1-P(Y_i=1/X_i))]$$

start with initial coefficients let $\beta_0 = 0$ $\beta_1 = 0.5$

calculate predicted probability

i.e) for $X = 2$

$$P(Y=1/X=2) = \frac{1}{1+e^{-(0+0.5 \times 2)}} = 0.731$$

$$P(Y=1/X=4) = \frac{1}{1+e^{-(0+0.5\times 4)}} = 0.88$$

$$P(Y=1/X=6) = \frac{1}{1+e^{-(0+0.5\times 6)}} = 0.95$$

$$P(Y=1/X=8) = \frac{1}{1+e^{-(0+0.5\times 8)}} = 0.92$$

$$P(Y=1/X=10) = \frac{1}{1+e^{-(0+0.5\times 10)}} = 0.94$$

compare log likelihood & actual $y = [0\ 0\ 1\ 1\ 1]$

Given Predict whether a student P/F based on no. of hrs studied

Hours x	2	4	6	8	10
P/F Y	0	0	1	1	1

For $x = 2$, $y = 0$

$l(\beta_0, \beta_1) = \log(1 - 0.73) =$

For $x = 4$, $y = 0$

$l(\beta_0, \beta_1) = \log(1 - 0.88) =$

For $x = 6$, $y = 1$ → (For $x = 1$ (bond take as 1 − value))

$l(\beta_0, \beta_1) = \log(0.95)$

update the coefficient:

$$\beta_0 \text{ new} = \beta_0^{\text{old}} + \alpha * \frac{\partial l(\beta_0, \beta_1)}{\partial \beta_0}$$

$$\beta_1 \text{ new} = \beta_1^{\text{old}} + \alpha * \frac{\partial l(\beta_0, \beta_1)}{\partial \beta_1}$$

$\alpha \rightarrow$ learning rate $\partial\beta_0, \partial\beta_1 \rightarrow$ partial derivatives of β_0 & β_1

Wald Test:

→ determine the set of independent variables are significant in a logistic regression model.

$$W = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

$\hat{\beta} \rightarrow$ estimated coefficient

$SE \rightarrow$ std error

↳ can be applied for every coeff

LR (Log likelihood Ratio Test)

$$LR = -2 \times (l_0 - l_1)$$

l_0 → null model → without the predictor

l_1 → full model → with predictor

Score test :

$$S = U(\hat{\beta_0})^T \cdot I(\hat{\beta_0})^{-1} \cdot U(\hat{\beta_0})$$

Logistic Regression model to predict whether a customer will purchase a product $Y=1$ and No $Y=0$ on their income X. $\beta_1 = 0.8$ $SE(\hat{\beta}) = 0.2$

Test the null hypothesis

$H_0 : \beta_1 = 0$ $\quad H_1 : \beta_1 \neq 0$

Wald test :

$$W = \frac{\hat{\beta}}{SE(\hat{\beta})} = \frac{0.8}{0.2} = 4$$

Compare with normal distribution under H_0, two tailed test. Calculate ϕ-value using z distribution

$$P = 2 \times (PZ > 4) \approx 2 \times 0.00003 = 0.00006 < 0.5$$

$\therefore$ P value is very small reject H_0 (which implies β_1 significantly affects Y).

LR test :

Predict whether a patient will develop disease based on age

Full model : $\log \frac{P(Y=1)}{P(Y=0)} = \beta_0 + \beta_1 x$

Null model : $\log \frac{P(Y=1)}{P(Y=0)} = \beta_0$

Consider full & null models are -120 & -140

$\therefore LR = -2 \times (l_0 - l_1) = -2 * (-140 - (-120))$

$= -2 \times -20$

$= 40$

test statistics follows a chisquare distribution with degree of freedom equal to the difference in the no. of parameters between two models.

p-value calculated using chisquare distribution with ~~diff~~ df = 1

$$p = P(\chi_1^2 > 40) \approx 0.0001$$

∴ p-value is very small. Reject H_0 age significantly contributes to the model.

Score test :

→ X_1 → education level, X_2 → income of owning a car (Y)

→ Score statistics for adding x_1 is 5.8, for adding x_2 is 12.1.

→ Follow chisquare distribution critical value of χ_1^2 at a 5% significance is 3.84

$5.8 > 3.84 \Rightarrow$ reject H_0

$12.1 > 3.84 \Rightarrow$ reject H_0 ∴ Both X_1 & X_2 shd be included in the model.

Problem :

Age	Monthly Premium (MP)	Year as customer	Renewed Y/N
30	150	2	Yes
45	200	5	NO
50	180	4	Yes
60	220	6	NO

B_0 (Intercept) = 3.5 $\beta_1 = -0.05$ $\beta_2 = -0.02$ $\beta_3 = 0.3$

i) Fit the logistic regr ii) Evaluate the model performance.

iii) Predict whether renewed for Age = 35, MP = 160, Years = 3

i) Logistic Regr Model fit:

$X_1 \rightarrow$ Age
$X_2 \rightarrow$ Appp
$X_3 \rightarrow$ Year
$Y \rightarrow$ Renewed

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$Y = 3.5 + 0.05 X_1 + 0.02 X_2 + 0.3 X_3$$

ii) For evaluating model performance:
Calculating predicting probability for each $X_1 X_2 X_3$ to compare with already given Renewed (Yes/No).

X_1	X_2	X_3	Given Y	$\hat{p}$
30	150	2	Yes	$3.5 - 0.05 \times 30 - 0.02 \times 150 + 0.3 \times 2 = 0.4 = z$ $\Rightarrow \frac{1}{1+e^{-0.4}} = 0.4$
45	200	5	No	0.22
50	180	4	Yes	0.198
60	220	6	No	0.109

Prediction Probability, $\hat{p} = \frac{1}{1+e^{-z}}$ where $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

> Here every $\hat{p}$ is < 0.5 so No but in given data Yes is there

iii) For $X_1 = 35$, $X_2 = 160$, Year $X_3 = 3$

$$\hat{p} = \frac{1}{1+e^{-(3.5 - 0.05 \times 35 - 0.02 \times 160 + 0.3 \times 3)}} = 0.366$$

∴ Prediction for given employee Not Renewed.

Multiple Logistic Regression:

Forward selection & Backward selection Methods:

Forward selection:

→ Start with null model i.e only intercept
→ add variables based on their significance
→ at each step variable improve the model
→ continue adding until no more significant variable is found.

Backward Selection (Elimination) :

→ start with all possible predictor variables in the model
→ gradually remove predictor that do not contribute for prediction.
→ continue removing until all remaining variables are statistically significant.

Example :
Predict whether student will pass / fail based on hours of study, attendance, GPA
where hours of study, attendance are more significant than GPA.

Interpretation of Parameters :

Log odds & Odds Ratio :

coefficients $\beta_0, \beta_1, \cdots \beta_k$ change outcome of one unit change as predictors by holding other variables constant.

odds ratio is calculated as e^{β_i}

eg: $\beta = 0.5$ for hrs of study

$\therefore$ odds ratio $\Rightarrow e^{0.5} = 1.65$

$\Rightarrow$ each additional hrs of study, odds of passing the exams are 1.65 times higher, while while holding others constant.

Multi nominal Logistic Regr : where the category is more than 2 like for eg : High, Medium, low. This model holds 1 category for reference & performs LR for other 2

Significance of coefficients :

> P-value is used to access the significance and uncertainty of the coefficient estimates.

P-value < 0.05 reject hypothesis
> 0.05 accept "

Categorical Data Analysis :
Multinomial Logistic Regression for more than 2 categories.
eg : Education level
UG (1) PG (2) HS (0) → keeping as reference

Sum : Perform forward & Backward selection : → Find odds ratio → Some write ups (theory)

Age	Monthly income	customer satisfaction	Retained (y)
25	3000	7	0
35	4000	9	1
45	5000	8	1
30	3500	6	0
40	4500	7	1

Given $\beta_0 = -5.5$, $\beta_1 = 0.002$, $\beta_2 = 0.8$, $\beta_3 = 0.03$

(Based on odds ratio give some explanation & eqns in each step)

General Linear Models :
① Distribution & Link function
→ What is GLM
→ Expo family (distributions & eqns
→ basic deseri for each member
→ Link functions
→ Derivation for regd (not for exam)

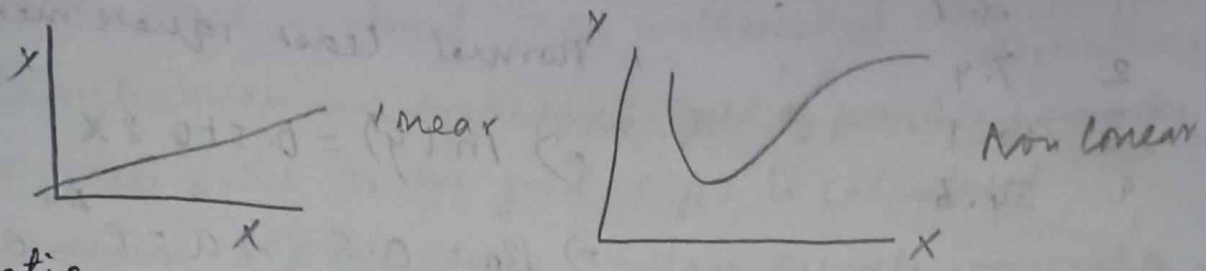
Theory
GLM
Log Reg, Fwd, Bwd selection
(PPT) Mult Log Regr

CIA - II

Sum & Algo
Decision Tree
KNN (PPT)
Naive Bays (PPT)
Linear Regression
→ Derivation of sigmoid function

4 × 10 = 40
other 5
1 × 10 = 10
50

Non Linear Regression :



Equation :

$$y = f(x, \beta) + \epsilon$$

$f(x, \beta)$ → Regression function x → Independent variable

β → parameter that model aims to estimate

ϵ → Error term

i) Polynomial Regression :

Uses higher order polynomial functions.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_n x^n$$

ii) Logistic Regression :

> Binary classification > Probability estimation > Non linear relationships

iii) Exponential Regression :

follow exponential growth (or) delay pattern

$$y = \alpha e^{(\beta x)}$$

Linearization (Techniques) Transform : → to transform non-linear eqn to linear

i) log :

eg : $y = ae^{bx}$

$$\ln(y) = \ln(a) + bx$$

$$\because Y = \beta_0 + \beta_1 x$$

$$\Rightarrow \beta_0 = \ln(a), \quad \beta_1 = b$$

x	y	ln(y)
1	2.7	
2	7.4	
3	20.1	
4	54.6	

Find the β_0 & β_1 using normal least square method

$\Rightarrow \ln(y) = 0.5 + 0.8x$

$\Rightarrow \beta_0 = 0.5 \quad \therefore a = e^{\beta_0} = e^{0.5}$

$= 1.648$

$\beta_1 = 0.8 \Rightarrow b = 0.8$

$\therefore$ Final Non linear eqn $\boxed{y = 1.648\, e^{0.8x}}$

II) Reciprocal :

$$\frac{1}{y} = \frac{a}{b}\frac{1}{x} + \frac{1}{b}$$

III) Polynomial :

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n$$

(Theory)

Iterative Procedure for NLS (Non Linear Least Square)

> It is used to find best fit parameters.

> Aims to minimize sum of squared residuals (RSS)

i.e) $S(\beta)$

$$\boxed{S(\beta) = \sum_{i=1}^{n} (y_i - f(x_i, \beta))^2}$$

$y_i \rightarrow$ observed value

$f(x_i, \beta) \rightarrow$ model predicted value using β

$S(\beta) \rightarrow$ sum of squared residuals

Common Iterative Procedures :

i) Grid search

ii) Newton Raphson

iii) Steepest Descent

iv) Marquardt's Method

i) Grid search :

> brute force method > No mathematical calculation of derivatives
> Search over a predefined grid of possible parameter values
> At each point in the grid $S(\beta)$ is calculated.
> Minimum $S(\beta)$ is chose > less accurate compared to others.

ii) Newton Raphson :

> Iterative Procedure > Involves mathematical derivatives
> Use second order info (Hessian Matrix of second derivatives) for parameter estimation
> fast & converges rapidly near the solution.

Steps :

1. Start with initial guess of β_0
2. Update $\boxed{\beta_{new} = \beta_{old} - H^{-1}(\beta_{old}) \cdot \nabla S(\beta_{old})}$

   $H(\beta_{old})$ - Hessian Matrix $\nabla S(\beta_{old})$ - gradient (1st derivatives)
3. Repeat until converges

iii) Steepest Descent (Gradient) :

> First order optimization technique
* Start with initial guess for β_0
* Update parameter by moving in direction opposite to gradient $\boxed{\beta_{new} = \beta_{old} - \alpha \cdot \nabla S(\beta_{old})}$

  $\alpha \rightarrow$ learning rate , $\nabla S(\beta_{old}) \rightarrow$ gradient of objective function $S(\beta)$
* continue updating until the change in the objective function.

N) Marquardt's Method :

> hybrid technique that combines both newton Raphson and steepest Descent.

> start with initial guess β_0

> Update $\beta_{new} = \beta_{old} - (J^T J + \lambda I)^{-1} J^T y$

$J \rightarrow$ Jacobian Matrix $y \rightarrow$ Vector of residuals

$\lambda \rightarrow$ damping parameter $I \rightarrow$ Identity matrix

λ is large $\rightarrow$ steepest descent

λ is small $\rightarrow$ Newton Raphson

> Adjust λ .

Semi Parametric Regression :

> combine both parametric & Non parametric models.

> Predicting income (Y) based on age (x_1) and education level (x_2)

$$Y = \beta_0 + \beta_1 x_1 + f(x_2) + \epsilon$$

Additive Regression :

> Response variable depend on each predictor through an independent smooth function.

$$Y = \beta_0 + f_1(x_1) + f_2(x_2) + \epsilon$$

eg : House Price ← Square footage ← distance from city

Non-parametric :

- → kernel regression – weighted avg of nearby data
- → Smoothing splines – Smooth curve through data
- → Local Regression – Perform Regression locally using nearby data.

eg : bike rental demand vs temperature.

UNIT-4

Covariance :

$$cov(x,y) = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n}$$

Correlation :

$$cor = \frac{cov(x,y)}{\sigma_x \cdot \sigma_y}$$

Problem 1 :

Prod 1	120	125	130	135	140	145	150	$\Sigma x = 823$ $\bar{x} = 117.57$
Prod 2	115	120	125	130	135	140	145	$\Sigma y = 910$ $\bar{y} = 130$

Here ~~$\Sigma(x-\bar{x}) = 0$, $\Sigma(y-\bar{y}) = 0$~~ So, cov = 100 & Cor = 1

2)

	day	1	2	3	4	5		
x	A	320	350	410	490	510	2080	$\bar{x} = 416$
y	B	380	440	520	590	610	2540	$\bar{y} = 508$
$x-\bar{x}$		-96	-66	-6	74	94	0	
$y-\bar{y}$		-128	-68	12	82	102	0	

Here ~~Cov = Cor = 0~~

3)

x	7	8	9	10	11	12	9.5
y	1	2	3	4	5	6	3.5
$x-\bar{x}$	~~-2.5~~	~~-1.5~~	-0.5	0.5	1.5	2.5	
$y-\bar{y}$	~~-2.5~~	-1.5	-0.5	0.5	1.5	2.5	
$(x-\bar{x})(x-\bar{y})$	6.25	2.25	0.25	0.25	2.25	6.25	17.5
$(x-\bar{x})^2$							17.5
$(y-\bar{y})^2$							17.5

$\sigma_x = \frac{17.5}{6}$

$\therefore cov(x,y) = \frac{17.5}{6}$

$= 2.92$

$cor = \frac{2.92}{\sqrt{2.92} \times \sqrt{2.92}}$

$cor = 1$

Features of covariance :

- $cov(x, a) = 0$ (a: constant)
- $cov(x, x) = Var(x)$
- $cov(x, y) = cov(y, x)$
- $cov(x+y, z) = cov(x, z) + cov(y, z)$
- $cov(ax, by) = ab\, cov(x, y)$
- Covariance of independent variable is zero

Auto covariance :

For single variable different times relation,

eg:

2017	2018	2019	2020	2021	2022
7	8	9	10	11	12
X_{t-5}	X_{t-4}	X_{t-3}	X_{t-2}	X_{t-1}	X_t

lag 0, lag 1, lag 2, lag 3

$\gamma_1 = cov(X_t, X_{t-1}) = E(X_t, X_{t-1}) - E(X_t) \cdot E(X_{t-1})$

$\gamma_2 = cov(X_t, X_{t-2})$

$\gamma_0 = cov(X_t, X_t) = Var(X_t)$

Timeseries : → Auto cov, cor, ARMA, ARIMA

components : 1. Trend 2. Seasonality 3. Cyclic 4. Irregular

Stationarity of stochastic process

Stationary → mean, var, autocorr are all constant over time

all constant → Strictly stationary

only mean, var constant → Weakly stationary

Yule Walker Estimation:

Γ -> covariance matrix, λ -> covariance vector, γ_0 - covariance

In order find correlation, 1st have to find covariance

Consider AR eqn of $p=2$

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \varepsilon_t$$

$(\times) y_{t-k}$ on both sides

$$y_t y_{t-k} = \phi_1 y_{t-k} y_{t-1} + \phi_2 y_{t-2} y_{t-k} + \varepsilon_t y_{t-k}$$

Expectation on both sides

$$E(y_t y_{t-k}) = \phi_1 E(y_{t-k} y_{t-1}) + \phi_2 E(y_{t-2} y_{t-k}) + E(\varepsilon_t y_{t-k})$$

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2}$$

(D) B/c only if $k \neq 0$ This whole term can be neglected

÷ by γ_0 (variance) B/c not it is a standard deviation i.e) variance

$$\frac{\gamma_k}{\gamma_0} = \phi_1 \frac{\gamma_{k-1}}{\gamma_0} + \phi_2 \frac{\gamma_{k-2}}{\gamma_0}$$

$$\Rightarrow \boxed{\rho_k = \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2}}$$

ρ -> correlation

$(\rho_{-1}) = (\rho_1)$

ρ_0 (no contribution) $= 1$

Put $k=1$ $\Rightarrow$ $\rho_1 = \phi_1 \rho_0 + \phi_2 \rho_{-1}$

$$= \phi_1 \rho_0 + \phi_2 \rho_1$$

$$\Rightarrow \rho_1 = \phi_1 + \phi_2 \rho_1 \longrightarrow (1)$$

$$\rho_1 - \phi_2 \rho_1 = \phi_1$$

$$\rho_1 (1 - \phi_2) = \phi_1 \Rightarrow \rho_1 = \frac{\phi_1}{1-\phi_2}$$

$k=2$ $\Rightarrow$ $\rho_2 = \phi_1 \rho_1 + \phi \rho_0$

$$\rho_2 = \phi_1 \rho_1 + \phi_2$$

$$\rho_2 = \phi_1 \left(\frac{\phi_1}{1-\phi_2}\right) + \phi_2$$

$$= \frac{\phi_1^2 - \phi_2^2 + \phi_2}{1-\phi_2}$$

Prblms in Unit-2

1. Normalisation

Min Max, Zscore, decimal scaling

> Binning method - Data cleaning

> Dissimilarity (Proximity matrix) for all data

> chi square test

> Wavelet transform - Data preprocessing

Unit-2 Prblms

> Apriori

> FP Growth

> Decision tree

> NB

> KNN

> Linear Regr

> Classification metrics
> Performance evaluation

Unit-3 Prblm

> Testing of hypothesis

> Linear Regr

> (Rare case) Logistic Regr

Unit-4 Prblm

> Covariance & Correlation

UNIT - 1

> OLAP, OLTP - Operation, defaults, diff b/w

> Data Mining Architectures

↳ Techniques.
↳ Task / Functionalities
↳ Applications

> Data Preprocessing

↳ Chisquare, Z-score like Problems → binning
↳ Theory every probms.

> Attribute Oriented Analysis (AOI)

> Dissimilarity matrix, Distance Matrix

UNIT - 2 Problem

Association rule Mining

> Apriori, FP Growth (steps shd be clear)

> Classification (write Procedure by our own)

- Decision Tree
- Naive Bayes
- KNN
- Linear Models

UNIT - 3 (Theory)

> Test of Hypothesis - Linear Regr

> Logistic Regr - Derivation, Diff b/w Linear & Logistic

> Fwd & Backwd > Multi Logistic Regr

> GLM (Based on marks write the no. of distributions)

> Non-Linear Regression (what to what)

> Semiparametric Regress.

> Additive "

> Non Parametric "

} Handwritten notes in GCR

UNIT - 4

> Problem → Covariance & Correlation (For Univariate time series Data)

> Intro of Time Series Data

> Estimation of ARMA → Yule Walker, Max Likelihood, Least Square

→ Prescriptive analysis : Math optimization
Decision & Risk Analysis ↳ Decision Tree.