DATA MINING - 18IT530

→ Knowledge discovery from data

→ Extraction of interesting patterns or knowledge from huge amount of data.

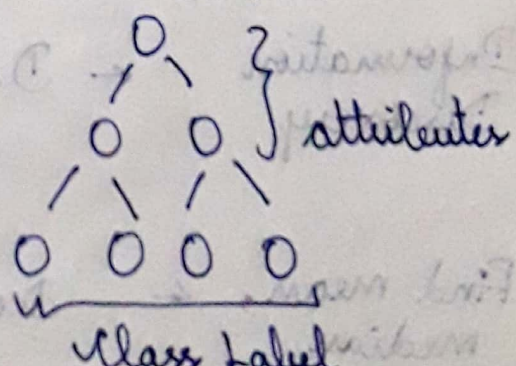
Steps in KDD :-

1. Data cleaning - Remove missing, inconsistent data
2. Data Integration - Combine multiple sources
3. Data Selection - Retrieve relevant data from DB
4. Data transformation - Consolidated into forms

Association rule → frequent itemset

Classification = supervised learning

⇒ Decision tree



Clustering

↓

Unsupervised learning

5. Data Mining - Intelligent methods are used to extract data patterns

6. Pattern Evaluation - Check the model precision

7. Knowledge Evaluation - Visualisation of model (Interestingness measures) is done

Trees, Rules

Bussiness Intelligence (BI)

Process for analyzing data & presenting actionable information to make better business decisions. Helps executives, managers.

Data Mining in BI :-

Decision Making ← End User

Visualisation Techniques ← Data Presentation ← (Business Analyst)

Information Discovery ← Data Mining ← (Data Analyst)

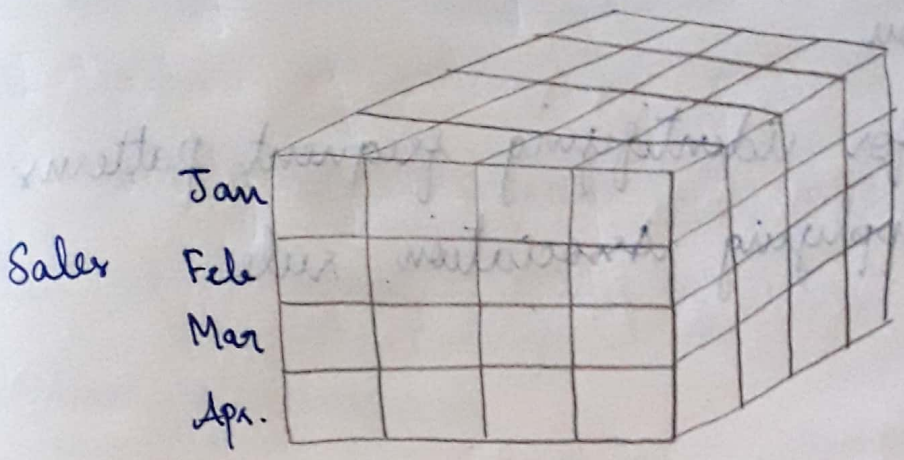
Find mean, Median Identify the relation btw. attributes ← Data Exploration (How data is distributed)

Data Sources (Paper, files) - DBAdministrator

→ BI uses Data Warehouse techniques.

⇓

uses Data Cube → uses Data Cube Operations



Multi Dimensional View of Data Mining :-

* Data to be mined
* Knowledge to be mined (DM functions)
  - Association, Classification, Clustering
* Techniques utilised.
* App. adapted

DM Functionalities / Tasks.

1. Generalisation
2. Association & Correlation Analysis
3. Classification "
4. Clustering "
5. Outlier "

1. Generalisation

- Info. integration & DW construction

2. Association

- Used for identifying frequent patterns
- By Applying Association rule.

Data Warehouse

- Sub. oriented, integrated (multiple), time-variant, non-volatile collection of data in support of management's decision making process.
  - Historical Data Analysis
  - User OLAP

OLTP	OLAP
Online Transaction Processing	Online Analytical Process.
Users - Clerk, IT	Knowledge Workers.
Day to day ops	Decision support
App-oriented	Sub-oriented
Current up-to date	Historical data
Read/Write Ops.	Lots of scans, no read/write
Short, simple Transaction	Complex
Records - More	Less
Users - Thousands	Hundreds.
DB-size - small	large
Transaction throughput	Query throughput

DW Models

* Enterprise warehouse
  - collects all info. about subjects of org.
* Data Mart
  - subset of corporate wide-data that is of value to a specific group of users.
* Virtual Warehouse.
  - set of views over operational DB.

ETL
Extraction - Get Data
Cleaning - Detect errors
Transformation - Convert data
Load
Refresh.

Data Cubes ; DW is constructed using
1. Dimensional table - item or time
2. Fact table - measures & keys.

0-D - apex cuboid
⋮
4-D (base cuboid)

* Star Schema

Single Fact table is connected to many dimension table.

Fact Table - 2 parts

1st part - links to dimension table

2nd part - Measures (Inferences that can be made)

* Snowflake Schema

In fact table, connection is made to Dimension Table. This dimension table within points to another dimension table.

* Fact Constellation

Here, more than one fact table is present.

Drill down operation - OLAP op.

- Expansion of data. (eg :- from time quaters to months)

Roll up

- Consolidation of data (eg :- from cities to countries)

Dice

- Only specific data are retrieved.

Slice

Pivot

DM Function :-

2. Association & Correlation Analysis :-

- Used to find frequent patterns (items bought together) → called Market Basket Analysis
- For this calculation, Association Rule is used.
- eg Bread, Jam (using support, confidence)

3. Classification

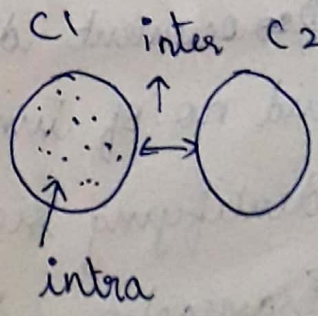
comes under supervised learning. ie they have class labels. Using class labels, a classifier model is constructed. For unknown class labels, the results are predicted.

4. Cluster Analysis (unsupervised learning)

⇓

ie class label is unknown.

- Group data into clusters having similarities



: Principle :-

Maximise intra class similarity &

Minimise inter class similarity

5. Outlier Analysis

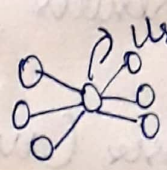
- Where we can get some useful information from data.
- Noisy (unuseful) data can also be useful sometimes.
- Used in fraud detection.

Time & Ordering :-

- Sequential pattern mining

ie first buy a product, then a related product later.

- Biological
- Graph Mining
- Info. Network Analysis



Types of Data Sets

1. Record
   - Relational, Matrix
   - Document data : term - frequency vector (to find no. of times a word occurs.) Used for identifying similar docs.
   - Transaction data
2. Graph & Network - www
3. Ordered
   - Video data
   - Time series
4. Spatial, image

Characteristics of Structured Data :-

1. Dimensionality - Rows x Columns.
   eg 600 x 3 - 3 attributes is not enough for observation of 600 data.
2. Sparsity
3. Resolution
4. Distribution

Data Objects :-

- Nothing but data

Attributes :-

Types

1. Nominal (category)
   eg :- Hair color = {..., ..., ...}
2. Binary
   Only 2 states (0, 1)
   Symmetric Binary (outcomes - equally imp.)
   eg :- gender
   Assymmetric Binary (outcomes - not equally imp.)
   eg :- Medical test (+ve, -ve)
3. Ordinal (Ordering)
   Values will have a meaningful order.
   eg :- Size {small, M, L}
4. Quantity (int or real valued)

5. Interval
   - Scale of equal sized units

6. Ratio

Discrete Attributes
- Finite set of values

Binary attributes are also special case of Dis. Attr.

Continuous Attributes

Real values are as attribute values.

eg:- Weight, Height

These are converted to Discrete Attri by grouping

Basic Statistical Descriptions of Data:-

Mean - Average

30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110

Mean = 58

Weighted Arithmetic Mean =

15, 25, 30, 90, 80, 15, 40 - Mean = 42.14

15, 15, 25, 30, 80, 80, 90 - Mean = 47.85 ✓

✓ - In this case, the mean is 47.85. But many students have got marks less than 47.85. So calculating Mean is not always useful for measuring

center of data.

So we go for Trimmed Mean — chop off values at extremes.

Median :-

Mid value in a set of ordered data values.

30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110

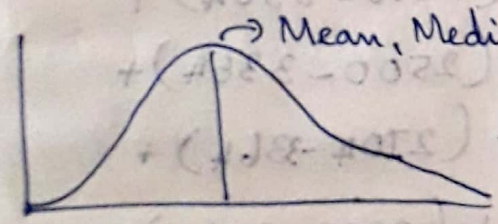
Median = 54

Mode :-

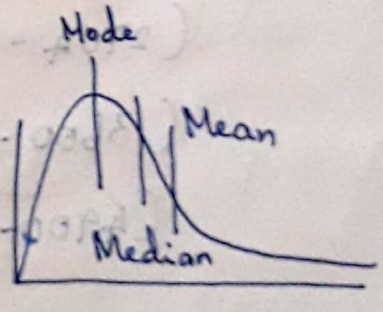
Value that occurs most frequently.

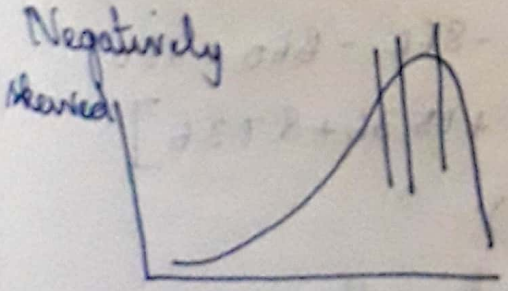
If Mean - Mode = 3 × (Mean - Median), then data is moderately skewed.

Symmetric Data



Mode < Mean, +vely skewed





Refer:- Boxplot Analysis

Variance & Std. Deviation

Low std. deviation - data tends to lee close to mean.

High " " - data are spread out over a large range of values.

30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Find the variance & std. deviation

$$\text{Variance} = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$$

μ - Mean - Avg.
N - No. of obs.

$\mu = 58$

$$V = \frac{1}{12}\Big[(900-3364)+(1296-3364)+(2209-3364)+(2500-3364)+(2704-3364)+(2704-3364)+(3600-3364)+(3969-3364)+(4900-3364)+(4900-3364)+(12100-3364)\Big]$$

$$= \frac{1}{12}\Big[-2464-2068-1155-864-660-660+236+605+1536+1536+8736\Big]$$

$\sigma^2 \cong 379.17$ (large value)

Means data is distributed all over.

Normal Dis. Curve :-

$\mu - \sigma$ to $\mu + \sigma$ - 68% of measurements.

$\mu - 2\sigma$ to $\mu + 2\sigma$ - 95%

$\mu - 3\sigma$ to $\mu + 3\sigma$ - 99.7

Quantile Plot

- For univariate data distribution

Sort the values of the attribute

Unit Price Cost

40 275

43 .

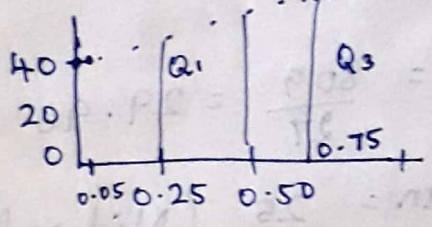
47 .

. .

. .

115

$fi = \frac{i - 0.5}{N} = \frac{1}{40} = \frac{40 - 0.5}{9. \ldots} = 0.05$



i - S.No.

∴ In Quantile Plot, we can't be able to find the Outlier.

Scatter Plot

- Bivariate data.

IQR, Q3 - Q1

Outlier :- $Q_1 - 1.5 \times IQR$

$20 - 1.5 \times (35 - 20)$

$= 17.85$

$Q3 + 1.5 \times IQR$ $= 1.5 \times 1.5 + 35 = 57.5$

Quantile - Quantile (Q-Q) Plot

Quantiles of one variate against another univariate is.

eg :- Same company Sales (attributes) is used for comparing in two different places. Mdu, Chn (Locations - another attribute).

13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

a) Mean = $\frac{809}{27}$ = 29.96 ≃ 30

Median = 25 (Mid. value)

b) Moder - 25 & 35, (highest frequency)

Bimodal - Two have same frequency.

c) Midrange = $\frac{13+70}{2} = \frac{83}{2} = 41.5$

13 - lowest value
70 - highest.

d) 1st Quartile (Q_1) = 25% = $\frac{n+1}{4} = \frac{27+1}{4} = 7$ (7^{th} value = 20)

= 20

3rd Quartile (Q_3) = 75% = $\frac{3(n+1)}{4} = \frac{3(28)}{4} = 21$

= 35

e) Five - number summary

Minimum, 1st Quartile, Median, 3rd Quartile, Maximum.

13, 20, 25, 35, 70

f) Boxplot

Outliers & Five no. summary is to be calculated for Boxplot.

OUTLIERS :-

$Q1 - 1.5 \times (IQR)$

$Q1 - 1.5 \times (Q3 - Q1)$

$20 - 1.5 \times (35 - 20)$

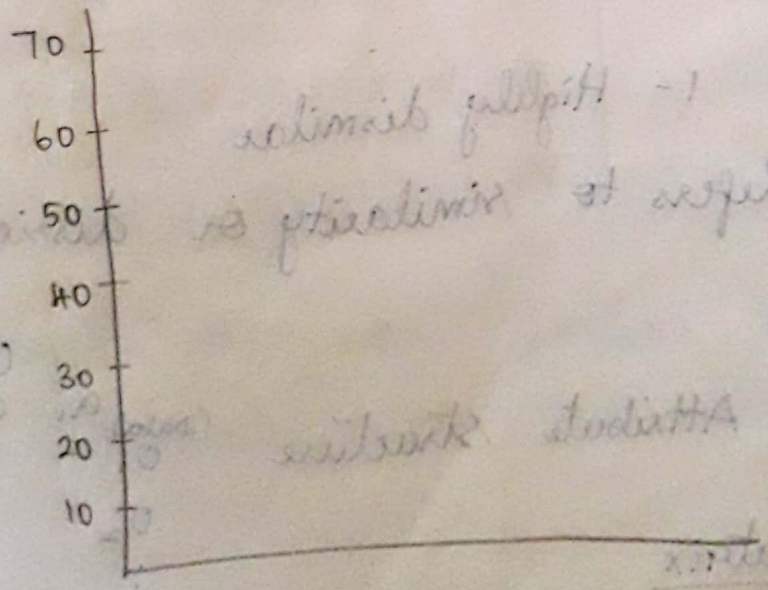
$20 - 1.5 \times 15$

$= 17.85$

$Q3 + 1.5 \times (IQR)$

$35 + (1.5 \times 1.5)$

57.5



DATA VISUALIZATION

Why?

Gain Insight

Find patterns

Provide qualitative overview

* **Icon - Based Technique :-**
  1. Chernoff Face
  2. Stick figure

* Pixel
* Circle segment
* Parallel lines

Similarity

Range [0,1] 1 - similar

Measure of how alike two obj. are.

Dissimilarity

0 to 1 - Range :- 1 - Highly dissimilar

Proximity - Refers to similarity or dissimilarity

Data matrix

Object by Attribute structure

(Data)	(Attributes) a_1	a_2	a_3 ...
O_1			
O_2			
⋮			

Dissimilarity matrix

Object by Object structure

	O_1	O_2	O_3	O_n
O_1	0			
O_2	$d(2,1)$	0		
O_3	$d(3,1)$	$d(3,2)$	0	

- Δ^{le} Matrix

Diagonal - always - 0 (as same objs)

Proximity Measures :- (For Nominal Attributes)

- Takes 2 or more states.

Method 1 Simple Matching

$$d(i,j) = \frac{p-m}{p}$$

m - matches
p - Total no. of vars.

$$Sim(i,j) = 1 - d(i,j)$$

eg:-

Name	Test (Nominal Attri)
X	0 +ve
Y	0 -ve
Z	A +ve
A	0 +ve

Dissimilar Matrix

	X	Y	Z	A
X	0			
Y	d(Y,X)	0		
Z	d(Z,X)	d(Z,Y)	0	
A	d(A,X)	d(A,Y)	d(A,Z)	0

$$d(Y,X) = \frac{1-0}{1} = 1$$

$$d(Z,X) = \frac{1-0}{1} = 1$$

$$d(A,X) = \frac{1-1}{1} = 0$$

	X	Y	Z	A
X	0			
Y	1	0		
Z	1	1	0	
A	0	1	1	0

Here p = no. of var = 1 i.e Test

∴ Objects A, X are similar. Other pairs are dissimilar.

Name	Test 1	Test 2	Test 3
X	O +ve	Code A	X
Y	O -ve	Code B	X
Z	A +ve	Code C	Z
A	O +ve	Code A	X
B	O +ve	Code B	Y

Here Test 1, Test 2, Test 3 are Nominal Attributes.

P = 3

$d(Y,X) = \frac{p-m}{p}$

$= \frac{3-0}{3} = 3/3 = 0.66$

$d(Z,X) = \frac{3-0}{3} = 1$

$d(Z,Y) = \frac{3-0}{3} = 1$

$d(A,X) = \frac{3-3}{3} = 0$

$d(A,Y) = \frac{3-1}{3} = \frac{2}{3} = 0.66$

$d(A,Z) = \frac{3-0}{3} = 1$

$d(B,A) = \frac{3-1}{3} = 0.66$

$d(B,X) = \frac{3-1}{3} = 0.66$

$d(B,Y) = \frac{3-1}{3} = 0.66$

$d(B,Z) = \frac{3-0}{3} = 1$

	X	Y	Z	A	B
X	0				
Y	0.66	0			
Z	1	1	0		
A	0	0.66	1	0	
B	0.66	0.66	1	0.66	0

Dissimilarity between Binary Variables

Contigency table for Binary Objects (For 2 objects based on 1 attribute comparison)

Object i \ Object j	1	0	
1	q (11)	r (10)	q+r
0	s (01)	t (00)	s+t
	q+s	r+t	p

Symmetric variables :- (0-1 - Gender - eg)

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

Assymmetric

$$d(i,j) = \frac{r+s}{q+r+s}$$

$$Sim(i,j) = \frac{q}{q+r+s}$$

↙ Jaccard coefficient

Eg

Name	Gender	Fever	Cough	T-1	T-2	T-3	T-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

Symmetric var - Gender

Assymmetric var - Fever, Cough, T-1, T-2, T-3, T-4

Y-1 ; N=0 ; P=1

	Fever	Cough	T-1	T-2	T-3	T-4
Jack(1)	Y(1)	N(0)	P(1)	N(0)	N(0)	N(0)
Mary(2)	Y(1)	N(0)	P(1)	N(0)	P(1)	N(0)
Jim(3)	Y(1)	P(1)	N(0)	N(0)	N(0)	N(0)

Dissimilarity Matrix

$$d(i,j) = \frac{r+s}{q+r+s}$$

(2 1)
r = 1 0
s = 0 1
q = 1 1

$$d(2,1) = \frac{r+s}{q+r+s} = \frac{1+0}{2+1+0} = \frac{1}{3} = 0.33$$

$$d(3,1) = \frac{r+s}{q+r+s} = \frac{1+1}{1+1+1} = \frac{2}{3} = 0.67$$

$$d(3,2) = \frac{r+s}{q+r+s} = \frac{1+2}{1+1+2} = \frac{3}{4} = 0.75$$

	Jack	Mary	Jim
Jack	0		
Mary	0.33	0	
Jim	0.67	0.75	0

∴ Jim & Mary - Highly dissimilar

Jack & Mary - Highly similar.

Similarity matrix :-

$$\begin{bmatrix} 1 & & \\ 0.67 & 1 & \\ 0.33 & 0.25 & 1 \end{bmatrix}$$

Dissimilarity of Numeric Data :-

- Using distance measures

1. Euclidean Distance

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

where $i = x_{i1}, x_{i2}, \dots$ $j = x_{j1}, x_{j2}, \dots$

2. Manhattan (or city block) Distance

$$d_{ij} = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots + |x_{ip} - x_{jp}|$$

Properties satisfied by Euclidean & Manhattan

1. Non-negative :- $d(i,j) \geq 0$
2. $d(i,i) = 0$ (Identity of Indiscernibles)
3. $d(i,j) = d(j,i)$ (symmetry)
4. $d(i,j) \leq d(i,k) + d(k,j)$ (Triangle Inequality)

3. Minkowski Distance (Gen. of Eucld & Manh.)

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

where $h \geq 1$; h - real no.

$h = 1$; Manhattan distance

$h = 2$; Euclidean Distance.

Supremum Distance

$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} \right)$$

1. Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8).

a) Compute Euc, Man, Min. ($q = 3$) & Sup. distance bet. two objs.

	A_1	A_2	A_3	A_4
Obj. 1	22	1	42	10
Obj. 2	20	0	36	8

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

$$= \sqrt{(22-20)^2 + (1-0)^2 + (42-36)^2 + (10-8)^2}$$

$$= \sqrt{4+1+36+4}$$

$$= \sqrt{45} = 6.7$$

Manhattan Distance $= |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \dots$

$$= |22-20| + |1-0| + |42-36| + |10-8|$$

$$= 2+1+6+2$$

$$= 11$$

Minkowski Distance $= \sqrt[h]{|x_{i1}-x_{j1}|^h + |x_{i2}-x_{j2}|^h + \dots}$

$= \sqrt[3]{|22-20|^3 + |1-0|^3 + |42-36|^3 + |10-8|^3}$

$= \sqrt[3]{8+1+216+8} = \sqrt[3]{233} = 6.153$

Supremum Distance :-

$$d(i,j) = \lim_{h \to \infty} \left(\sum_{f=1}^{p} \max_{f}^{p} |x_{if} - x_{jf}| \right)$$

$= \max(|22-20|, |1-0|, |42-36|, |10-8|)$

$= \max(2, 1, 6, 2)$

$= 6$

Proximity Measures for Ordinal Attributes :-

↳ meaningful ranking or ordering.

eg :- S, A, B, C, D, U (Grades)

By discretisation (grouping) of numeric attributes into finite categories.

$$Z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

Ordinal Attri → Give rank → Normalise → Numeric attributes →

Calculate Distance.

eg

Name ID	Test-1
1	Excellent
2	Fair
3	Good
4	Excellent

3 states ⇒ Fair, Good, Excellent

↳ $M_f = 3$ 1 2 3

Step 1 Replace by Rank

Name ID	Test-1
1	3
2	1
3	2
4	3

Step 2 Normalisation $Z_{if} = \frac{r_{if}-1}{M_f-1}$

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

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

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

Fair, Good, Excellent

0 , 0.5 , 1

Step 3 Use Distance Measure

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix} \end{array}$$

Name ID	Test
1	1
2	0
3	0.5
4	1

$$d(2,1) = \sqrt{(0-1)^2} = 1$$

$$d(3,1) = \sqrt{(0.5-1)^2} = \sqrt{0.25} = 0.5$$

$$d(4,1) = \sqrt{(1-1)^2} = 0$$

$$d(4,2) = 1$$

$$d(4,3) = 0.5$$

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{bmatrix} \end{array}$$

Obj. 1 & 2 - Dissimilar, Ulfy for 4 & 2.

Obj. 1 & 4 - similar.

Dissimilarity for Attributes of Mixed Type

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$$

1.) If x_{if} or x_{jf} is missing
2.) $x_{if} = x_{jf} = 0$
} then $\delta_{ij}^{(f)} = 0$

If f is numeric

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$$

eg

Obj. Identifier	Test-1 (Nominal)	Test-2 (Ordinal)	Test-3 (Numeric)
1	Code A	Excellent	45
2	Code B	Fair	22
3	Code C	Good	64
4	Code A	Excellent	28

Age	Name	T-1	T-2	T-3	Gender	Fever	Cough	T-A	T-B	T-C	T-D
52	X	O+ve	Code A	X	M	Y	N	P			
73	Y	O-ve	Code B	X	F	Y	N	P			
65	Z	A+ve	Code C	Z	M	Y	P	N			
48	A	O+ve	Code A	X	F	N	P	P			
37	B	O+ve	Code B	Y	M	N	P	N			

Cosine Similarity for Document :-

$$Sim(x,y) = \frac{x \cdot y}{\|x\| \|y\|}$$

↳ thousands of attributes each recording the frequency of a particular word.

↓

Keywords or attributes

Eg

Doc.	Team	Coach	Hockey
Doc 1	5	0	
Doc 2	2	3	
Doc 3	7	8	

Term frequency matrix.

$\|x\| = \sqrt{x^2}$

$x_{(1)} = \{5,0,3,0,2,0,0,2,0,0\}$

$z_{(3)} = \{0,7,0,2,1,0,0,3,0,0\}$

$x \cdot z = 0+0+0+0+2+0+0+6+0+0 = 8$

$|x| = \sqrt{25+9+4+4+0} = \sqrt{42} = 6.48$

$|z| = \sqrt{4+36} = \sqrt{40} = 6.3 \quad \sqrt{49+4+1+9} = \sqrt{63} = 7.93$

$$Sim(1,3) = \frac{8}{6.48 \times 6.3} = \frac{8}{40.32} = 0.198$$
$$7.93 \quad 51.38 \quad 0.155$$

(1) $x = \{5, 0, 3, 0, 2, 0, 0, 2, 0, 0\}$

(4) $W = \{0, 1, 0, 0, 1, 2, 2, 0, 3, 0\}$

$|x| = 6.48$

$|W| = \sqrt{1+1+4+4+9} = \sqrt{19} = 4.35$

$x \cdot W = 0+0+0+0+2+0+0+0 = 2.$

$$Sim(1,4) = \frac{2}{6.48 \times 4.35} = \frac{2}{28.188} = 0.070$$

$2 = \{3, 0, 2, 0, 1, 1, 0, 1, 0, 1\}$

$4 = \{0, 1, 0, 0, 1, 2, 2, 0, 3, 0\}$

$|2| = \sqrt{9+4+1+1+1+1} = \sqrt{17} = 4.12$

$|4| = 4.35$

$2 \cdot 4 = 1+2 = 3$

$$Sim(2,4) = \frac{3}{4.12 \times 4.35} = \frac{3}{17.922} = 0.167$$

$2 = \{3, 0, 2, 0, 1, 1, 0, 1, 0, 1\}$

$3 = \{0, 7, 0, 2, 1, 0, 0, 3, 0, 0\}$

$|2| = 4.12$

$|3| = 7.93$

$2 \cdot 3 = 1+3 = 4$

$$Sim(2,3) = \frac{4}{4.12 \times 7.93} = \frac{4}{32.671} = 0.122$$

$3 = \{0, 7, 0, 2, 1, 0, 0, 3, 0, 0\}$

$4 = \{0, 1, 0, 0, 1, 2, 2, 0, 3, 0\}$

$|3| = 7.93$

$|4| = 4.35$

$3.4 = 7 + 1 + = 8$

$$\text{Sim}(3,4) = \frac{8}{7.93 \times 4.35} = \frac{8}{34.49} = 0.23$$

$1 = \{5, 0, 3, 0, 2, 0, 0, 2, 0, 0\}$

$2 = \{3, 0, 2, 0, 1, 1, 0, 1, 0, 1\}$

$|1| = 6.48$

$|2| = 4.12$

$1.2 = 15 + 6 + 2 + 2 = 25$

$$\text{Sim}(1,2) = \frac{25}{6.48 \times 4.12} = \frac{25}{26.69} = 0.936$$

∴ Doc 1 & 2 are similar.

Doc 1 & 3 are dissimilar.

REAL - WORLD DATABASES

Data preprocessing :-

1. Data cleaning

   Remove noisy, inconsistency

2. Data Integration

   Merge multiple sourced data

3. Data Reduction

   Reduce Data size by clustering, eg Age+ DOB.

4. Data Tranformation & Data Discretisation

   Have values within range 0 - 1

Age	Name	T-1	T-2	T-3	Gender	Fever	Cough	T-A	T-B	T-C	T-D
52	X	O+ve	Code A	X	M	Y	N	P	N	N	N
73	Y	O-ve	Code B	X	F	Y	N	P	N	P	N
65	Z	A+ve	Code C	Z	M	Y	P	N	N	N	N
48	A	O+ve	Code A	X	F	N	P	P	P	N	P
37	B	O+ve	Code B	Y	M	N	P	N	P	P	N

Age → Numeric

T-1, T-2, T-3 } Nominal

Fever, Cough, T-A, T-B, T-C, T-D } Binary Assymmetric

Gender - Binary Symmetric

Nominal Attributes :-

$d(i,j) = p-m/p$

$d(Y,X) = \frac{3-1}{3} = \frac{2}{3} = 0.66$

$d(Z,X) = \frac{3-0}{3} = \frac{3}{3} = 1$

$d(Z,Y) = \frac{3-0}{3} = \frac{3}{3} = 1$

$d(A,X) = \frac{3-3}{3} = 0$

$d(A,Y) = \frac{3-1}{3} = \frac{2}{3} = 0.66$

$d(A,Z) = \frac{3-0}{3} = 1$

$d(B,X) = \frac{3-1}{3} = 0.66$

$d(B,Y) = \frac{3-1}{3} = 0.66$

	X	Y	Z	A	B
X	0				
Y	0.66	0			
Z	1	1	0		
A	0	0.66	1	0	
B	0.66	0.66	1	0.66	0

$d(B,Z) = \frac{3-0}{3} = 1$

$d(B,A) = \frac{3-1}{3} = 0.66$

Binary Asymmetric :- $d(i,j) = \frac{r+s}{q+r+s}$

Name	Fever	Cough	T-A	T-B	T-C	T-D
X	Y(1)	N(0)	P(1)	N(0)	N(0)	N(0)
Y	Y(1)	N(0)	P(1)	N(0)	P(1)	N(0)
Z	Y(1)	P(1)	N(0)	N(0)	N(0)	N(0)
A	N(0)	P(1)	P(1)	P(1)	N(0)	P(1)
B	N(0)	P(1)	N(0)	P(1)	P(1)	N(0)

$d(Y,X) = \frac{1+0}{2+1+0} = \frac{1}{3} = 0.33$

$d(Z,X) = \frac{1+1}{1+1+1} = \frac{2}{3} = 0.67$

$d(Z,Y) = \frac{1+2}{1+1+2} = \frac{3}{4} = 0.75$

$d(A,X) = \frac{3+1}{1+3+1} = \frac{4}{5} = 0.8$

$d(A,Y) = \frac{3+2}{1+3+2} = \frac{5}{6} = 0.83$

$d(A,Z) = \frac{3+1}{1+3+1} = \frac{4}{5} = 0.8$

$d(B,X) = \frac{3+2}{0+3+2} = 1$

$d(B,Y) = \frac{2+2}{1+2+2} = \frac{4}{5} = 0.8$

$d(B,Z) = \frac{2+1}{1+2+1} = \frac{3}{4} = 0.75$

$d(B,A) = \frac{1+2}{2+1+2} = \frac{3}{5} = 0.6$

	X	Y	Z	A	B
X	0				
Y	0.33	0			
Z	0.67	0.75	0		
A	0.8	0.83	0.8	0	
B	1	0.8	0.75	0.6	0

Binary Symmetric :- $d(i,j) = \frac{r+s}{q+r+s+t}$

Name	Gender
X	M(1)
Y	F(0)
Z	M(1)
A	F(0)
B	M(1)

	X	Y	Z	A	B
X	0				
Y	1	0			
Z	0	1	0		
A	1	0	1	0	
B	0	1	0	1	0

$d(Y,X) = \frac{0+1}{0+0+1+0} = 1$

$d(Z,X) = \frac{0+0}{1+0+0+0} = 0$

$d(Z,Y) = \frac{1+0}{0+1+0+0} = 1$

$d(A,X) = \frac{0+1}{0+0+1+0} = 1$

$d(A,Y) = \frac{0+0}{0+0+0+1} = 0$

$d(A,Z) = \frac{0+1}{0+0+1+0} = 1$

$d(B,X) = \frac{0+0}{1+0+0+0} = 0$

$d(B,Y) = \frac{1+0}{0+1+0+0} = 1$

$d(B,Z) = \frac{0+0}{1+0+0+0} = 0$

$d(B,A) = \frac{1+0}{0+1+0+0} = 1$

Numeric Attributes :-

Name	Age
X	52 (3)
Y	73 (5)
Z	65 (4)
A	48 (2)
B	37 (1)

$M_f = 5$

37 , 48 , 52 , 65 , 73

(1) (2) (3) (4) (5)

$$Z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

i = 1

$$Z_{if} = \frac{1-1}{5-1} = 0$$

i = 2

$$Z_{if} = \frac{2-1}{5-1} = \frac{1}{4} = 0.25$$

i = 3

$$Z_{if} = \frac{3-1}{5-1} = \frac{2}{4} = 0.50$$

i = 4

$$Z_{if} = \frac{4-1}{5-1} = \frac{3}{4} = 0.75$$

i = 5

$$Z_{if} = \frac{5-1}{5-1} = 1$$

Name	Rank
X	0.50
Y	1
Z	0.75
A	0.25
B	0

	X	Y	Z	A	B
X	0				
Y	0.50	0			
Z	0.25	0.25	0		
A	0.25	0.75	0.50	0	
B	0.50	1	0.75	0.25	0

$d(Y,X) = |1-0.5| = 0.50$

$d(Z,X) = |0.75-0.50| = 0.25$

$d(Z,Y) = |0.75-1| = 0.25$

$d(A,X) = |0.25-0.50| = 0.25$

$d(A,Y) = |0.25-1| = 0.75$

$d(A,Z) = |0.25-0.75| = 0.50$

$d(B,X) = |0-0.5| = 0.5$

$d(B,Y) = |0-1| = 1$

$d(B,Z) = |0-0.75| = 0.75$

$d(B,A) = |0-0.25| = 0.25$

Dissimilarity Matrix for Mixed Types :-

$$d(i,j) = \frac{\sum_{f=1}^{p} \delta_{ij} d_{ij}}{\sum_{f=1}^{p} \delta_{ij}}$$

	X	Y	Z	A	B
X	0				
Y	0.66	0			
Z	1	1	0		
A	0	0.66	1	0	
B	0.66	0.66	1	0.66	0

	X	Y	Z	A	B
X	0				
Y	0.33	0			
Z	0.67	0.75	0		
A	0.80	0.83	0.8	0	
B	1	0.8	0.75	0.6	0

	X	Y	Z	A	B
X	0				
Y	1	0			
Z	0	1	0		
A	1	0	1	0	
B	0	1	0	1	0

	X	Y	Z	A	B
X	0				
Y	0.5	0			
Z	0.25	0.25	0		
A	0.25	0.75	0.5	0	
B	0.5	1	0.75	0.25	0

$d(Y,X) = (1\times0.66)+(1\times0.33)+(1\times1)+(1\times0.5)/1+1+1+1 = \frac{2.49}{4} = 0.6$

$d(Z,X) = (1\times1)+(1\times0.67)+(1\times0)+(1\times0.25)/4 = 1.92/4 = 0.48$

$d(Z,Y) = (1\times1)+(1\times0.75)+(1\times1)+(1\times0.25)/4 = 3/4 = 0.75$

$d(A,X) = (1\times0)+(1\times0.80)+(1\times1)+(1\times0.25)/4 = \frac{2.05}{4} = 0.512$

$d(A,Y) = (1\times0.66)+(1\times0.83)+(1\times0)+(1\times0.75)/4 = \frac{2.24}{4} = 0.56$

$d(A,Z) = (1\times1)+(1\times0.8)+(1\times1)+(1\times0.50)/4 = \frac{3.3}{4} = 0.825$

$d(B,X) = (1\times0.66)+(1\times1)+(1\times0)+(1\times0.5)/4 = \frac{2.16}{4} = 0.54$

$d(B,Y) = (1\times0.66)+(1\times0.8)+(1\times1)+(1\times1)/4 = 3.46/4 = 0.865$

$d(B,Z) = (1\times1)+(1\times0.75)+(1\times0)+(1\times0.75)/4 = 2.5/4 = 0.625$

$d(B,A) = (1\times0.66)+(1\times0.6)+(1\times1)+(1\times0.25)/4 = \frac{2.51}{4} = 0.627$

	X	Y	Z	A	B
X	0				
Y	0.622	0			
Z	0.48	0.75	0		
A	0.512	0.56	0.825	0	
B	0.54	0.865	0.625	0.627	0

Data Cleaning :-

* Incomplete :- Occupation = " "
* Noisy :- Salary = "-10"
* Inconsistent :- Age = 42 ; Bday = "03/07/2010"
* Intentional :- Jan 1 as everyones bday.

Handling Missing Data :-

1. Ignore the tuple :-
   Based on the % of missing values.
2. Fill in the missing values manually.
3. Fill with
   → global constant
   → Mean or Median
   → Smarter way - using same class labels
4. Most probable value
   Using Regression, Bayesian formula.

2. Noisy Data.
   Random error / variance

Handling Noisy Data :-

1. Binning
   first sort & partition into equal frequencies (bins).

a. Smooth by bin means :-

b. " " " median

c. " " " boundaries

4, 8, 15, 21, 21, 24, 25, 28, 34 (sorted)

Bin 1 :- 4, 8, 15

Bin 2 :- 21, 21, 24

Bin 3 :- 25, 28, 34

Smoothing by bin means :-

Bin 1 :- 9, 9, 9

Bin 2 :- 22, 22, 22

Bin 3 :- 29, 29, 29

Smoothing by bin boundaries :-

Bin 1 :- 4, 8, 15

Bin 2 :- 21, 21, 24

Bin 3 :- 25, 25, 34

2. Regression :-

Fit data into Reg. functions.

3. Outlier Analysis

- By clustering

Data that doesn't fit into clusters are called Outliers (Noisy)

4. Combined computer & Human Inspection

Data Cleaning as a Process :-

Data Integration :-

* Schema Integration

  eg:- cust.id & cust.no.

  And conversions need to be done.

  eg :- 160 cm & 5 feet

Handling Redundancy

* Redundant - derived attributes - age, DOB.
* These redundant attributes are detected by Correlation Analysis & Covariance Analysis.

CORRELATION ANALYSIS (Nominal Data)

x^2 (chi square) test :- (for nominal attributes).

$$x^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

Larger x^2

A	B
red	Fiction (2)
(3) blue	Non-Fiction
green	⋮

150 samples

$$X^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

O_{ij} ⇒ observed freq of joint event (A_i, B_j)

e_{ij} - Exp. freq.

$$e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{n}$$

* If A & B are indepent, then there is no correlation between them.

eg:- 1500 people. Gender was noted, & the type of books preferred are noted(F/NF). Check whether gender & preferred readings are correlated.

2 X 2
(G) (PR)
M F F NF

Soln :-

2 × 2 Contigency Table

	Male	Female	Total
Fiction	250 e_{11}	200 e_{12}	450
Non-Fiction	50 e_{21}	1000 e_{22}	1050
	300	1200	1500

$$e_{11} = \frac{300 \times 450}{1500} = 90$$

$$e_{12} = \frac{1200 \times 450}{1500} = 360$$

$$e_{21} = \frac{300 \times 1050}{1500} = 210$$

$$e_{22} = \frac{1200 \times 1050}{1500} = 840$$

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$

$$= 284.44 + 121.90 + 71.11 + 30.47$$

$$\chi^2 = 507.93$$

Degree of freedom :- $(r-1)(c-1) = (2-1) \times (2-1) = 1$

$507.93 > 10.828$ (Table value) sign value = 0.001

∴ Reject the hypothesis that gender & pre. reading are independent.

Conclusion :-

Gender & Pre. Reading are strongly correlated.

eg

Gender	Body Image: About Right	Overweight	Underweight	Total
Female	560 e_{11}	163 e_{12}	37 e_{13}	760
Male	295 e_{21}	72 e_{22}	73 e_{23}	440
Total	855	235	110	1200

3 x 2

$$e_{11} = \frac{855 \times 760}{1200} = 541.5$$

$$e_{12} = \frac{235 \times 760}{1200} = 148.83$$

$$e_{13} = \frac{110 \times 760}{1200} = 69.66$$

$$e_{21} = \frac{855 \times 440}{1200} = 313.5$$

$$e_{22} = \frac{235 \times 440}{1200} = 86.2$$

$$e_{23} = \frac{110 \times 440}{1200} = 40.3$$

$$x^2 = \frac{(560 - 541.5)^2}{541.5} + \frac{(163 - 148.83)^2}{148.33} +$$

$$\frac{(37 - 69.66)^2}{69.66} + \frac{(295 - 313.5)^2}{313.5} +$$

$$\frac{(72 - 86.2)^2}{86.2} + \frac{(73 - 40.3)^2}{40.3}$$

$$x^2 = 47.18$$

Table :-

$$(r-1)(c-1) = (2-1) \times (3-1) = 2 \text{ (DOF)}$$

$$X^2 = 13.816$$

$$x^2 :- \quad 47.18 > 13.816$$

Reject the hypothesis

∴ Gender & Body Image are strongly correlated.

CORRELATION COEFFICIENT FOR NUMERIC DATA :-

$$r_{A,B} = \frac{\sum_{i=1}^{n} (a_i b_i) - n \bar{A} \bar{B}}{n \sigma_A \sigma_B} \quad ; \quad -1 \le r_{A,B} \le +1$$

* Also known as pearson's product moment co-efficient. $r_{A,B} > 0 \rightarrow A, B$ are +ve correlated

eg

Time points	All Electronics (A)	Hightech (B)
t_1	6	20
t_2	5	10
t_3	4	14
t_4	3	5
t_5	2	5

$n = 5$

$\Sigma a_i b_i = 6 \times 20 + 5 \times 10 + 4 \times 14 + 3 \times 5 + 2 \times 5$

$= 251$

$\bar{A} = \frac{6+5+4+3+2}{5} = 4$

$\bar{B} = \frac{20+10+14+5+5}{5} = 10.8$

$\sigma_A = \frac{1}{N} \sum_{i=1}^{n} x_i^2 - \mu^2$

$= \frac{1}{5} \left[(6^2 - 4^2) + (5^2 - 4^2) + 0 + (3^2 - 4^2) + (2^2 - 4^2) \right]$

$= \frac{1}{5}[20+9+(-7)+(-12)]$

$= \frac{1}{5}(10) = 2$

$\sigma_B = \frac{1}{5}[(20^2 - 10.8^2) + (10^2 - 10.8^2) + (14^2 - 10.8^2) + (5^2 - 10.8^2) + (5^2 - 10.8^2)]$

$= \frac{1}{5}[$

$=$ S.To

$r_{A,B} = \frac{251 - 5 \times 4 \times 10.8}{5 \times 2 \times 5.70}$

$r_{A,B} = 0.61$

(*) Higher the value, higher the correlation. Hence any one attribute can be eliminated due to redundancy.

COVARIANCE OF NUMERIC DATA :-

$E(A) = \bar{A} = \frac{\sum_{i=1}^{n} a_i}{n}$; $E(B) = \bar{B} = \frac{\sum_{i=1}^{n} b_i}{n}$

$Cov(A,B) = E(A.B) - \bar{A}\bar{B}$

$r_{A,B} = \frac{Cov(A,B)}{}$ (Pearson coeff from Covariance)

Covariance = 0 (independent)

Covariance = ?

$Cov(A,B) = E(A.B) - \bar{A}\bar{B}$

$\bar{A} = 4$

$\bar{B} = 10.8$

$$E(A.B) = \frac{6\times20 + 5\times10 + 3\times5 + 4\times4 + 2\times5}{5}$$

$Cov(A,B) = (50.2) - (4 \times 10.8)$

$= 50.2 - 43.7 = 7$

Positive covariance

→ Implies stock prices for both companies rise together.

* Covariance is calculated to find the trend of two attributes & here no elimination is done.

Tuple Duplication :-

Duplicates should be removed at tuple level.

Data Transformation

1. Binning, Regression.... (Smoothing)
2. Attribute Construction (New attri. added)
3. Aggregation eg:- daily attendance → monthly attenda. (using sum(), avg())
4. Normalisation (0 to 1 range)
5. Discretisation → raw values are converted to
   eg (0-10), (11-20) - interval values (or)
   (youth, adult, senior) - conceptual labels.

6. Concept heirachy generation for Nominal Data

⊛ Bottom up - Top down

Normalisation:- (Give all attri an equal weight)

① Min-Max Normilisation

↳ linear trans. on Original data.

$$V_i' = \frac{v_i - min_A}{max_A - min_A} (newmax_A - newmin_A) + newmin_A$$

↳ Range in which data is to lee fit
eg 0 to 1
-1 to 1

Prob.

The min & max. values for the Attribute Income are $12000 & $98,000. Map in the range [0,1] for $73,600

$$V_i' = \frac{73600 - 12,000}{98000 - 12000} (1-0) + 0$$

$$= \frac{61600}{86000}$$

$$V_i'' = 0.716$$

$$V_i' = \frac{12000 - 12000}{98000 - 12000} (1-0) + 0 = 0$$

$$V_i' = \frac{98000 - 12000}{98000 - 12000} (1-0) + 0 = 1$$

Salary	Normalised value
73,600	0.716
12,000	0
98,000	1

Prob

Age - 10, 15, 17, 21, 23, 45, 70

Range - [-2 to 2]

(10)

$$v_i' = \frac{10-10}{70-10}(2-(-2)) + (-2) = -2$$

(15)

$$v_i' = \frac{15-10}{60}(2+2) + (-2)$$

$$= \frac{5}{60}(4) - 2$$

$$= 0.33 - 2 = -1.67$$

(17)

$$v_i' = \frac{17-10}{60}(4) - 2$$

$$= 0.47 - 2 = -1.53$$

(21)

$$v_i' = \frac{21-10}{60}(4) - 2$$

$$= 0.73 - 2 = -1.27$$

(23)

$$v_i' = \frac{23-10}{60}(4) - 2$$

$$= 0.87 - 2 = -1.13$$

(45) $V_i' = \frac{45-10}{60}(4) - 2$

$= 2.33 - 2$

$= 0.33$

(70) $V_i' = \frac{70-10}{60}(4) - 2$

$= 2$

Z-Score Normalisation (Zero-Mean Normalisation)

$$V_i' = \frac{V_i - \bar{A}}{\sigma_A}$$

$\bar{A}$ - Mean

σ_A - Std. Deviation

Prob

Age, 10, 15, 17, 21, 23, 45, 70

Mean = 28.7

$\sigma_A = 19.77$

(10) $V_i' = \frac{10-28.7}{19.77} = -0.94$

(17) $V_i' = \frac{17-28.7}{19.77} = \frac{-11.7}{19.77} = -0.591$

(21) $V_i' = \frac{21-28.7}{19.77} = \frac{-7.7}{19.77} = -0.389$

(23) $V_i' = \frac{23-28.7}{19.77} = -0.288$

Variation of Z-Score Normalisation :-

$$V_i' = \frac{V_i - \bar{A}}{S_A}$$

S_A - Mean absolute deviation of A.

$$S_A = \frac{1}{n}(|V_1 - \bar{A}| + |V_2 - \bar{A}| + \dots)$$

more robust for outliers.

Normalisation By Decimal Scaling :-

$$V_i' = \frac{V_i}{10^j}$$

$j \rightarrow \max(|V_i'|) < 1$

Prob :-

Range from -986 to 917.

Max. no. of digits (986, 917) = 3

∴ Divide each value by 1000 ie $j = 3$.

-986 :- $V_i' = \frac{-986}{10^3} = -0.986$

-917 $V_i' = \frac{-917}{10^3} = -0.917$

eg

-986 to 1050

Ans :- -0.0986 to 0.1050

Data Reduction :-

- Reduced rep. of Dataset
- Smaller volume
- Must produce results close to that obtained with the whole dataset.

DR Strategies :-

1. Dimensionality Reduction

Forward Selection

$\{A_1, A_2, A_3, A_4, A_5, A_6\}$

↳ Initial Attributes

Steps:-

$\{\} \rightarrow \{A_1\} \rightarrow \{A_1, A_4\} \rightarrow \{A_1, A_4, A_6\}$

↳ Reduced.

Reverse is Backward Elimination.

The attributes are selected/removed based on feature selection Algorithm.

Linear Regression :-

$$y = wx + b$$

response variable ← predictor variable

w, b - regression coefficients.

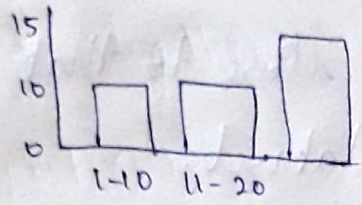
Multiple Linear Regression :-

$$y = P_0 + P_1 x_1 + P_2 x_2 + \dots$$

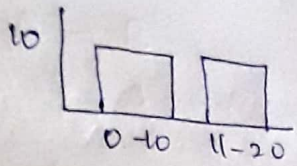
Log Linear Model :-

Histograms.

1. Equal Width

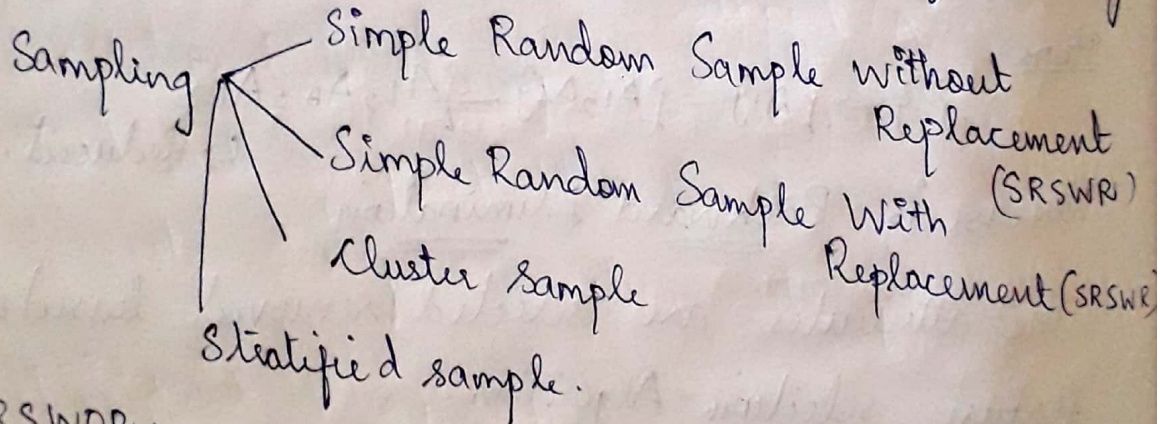


2. Equal Depth (frequency)



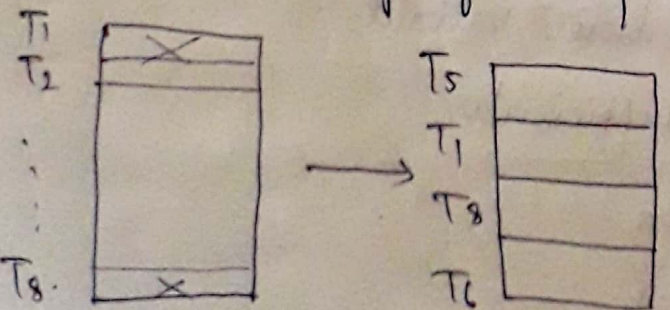
Sampling :-

Taking small amount of data from large data.



SRSWOR :-

* $S < N$
* Probability of 1 tuple $= 1/N$



Original Dataset

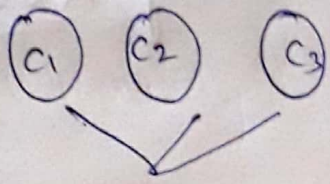
Once these are selected (T_1, T_5, T_6, T_8), these are removed from original dataset

SRSWR

Samples taken for sampling is not removed from original data.

Cluster sample :-

Form clusters & take one sample from each cluster.



Stratified Sample :-

Dataset is divided into parts called strata. Take one attribute & see all the values in the data & take for sampling.

Data Cube Aggregation

Data of Quater year sales are aggregated to Year sales.

Transactional DB

To find which items are bought together.

TID	Items
T_1	I_1, I_2, I_5
T_2	I_1, I_3, I_2

Support (s) :-

$A \Rightarrow B$, s - % of items having $A \cup B$.

eg:- $A \Rightarrow B$

Milk $\Rightarrow$ Jam.

Confidence

$$\text{Confidence}(A \Rightarrow B) = P(B/A) = \frac{\text{support}(A \cup B)}{\text{support}(A)}$$

Association Rule Mining :- (2 step process)

1. Find all frequent itemsets.
2. Generate strong association rule from freq. DS

Freq. Itemset Mining Methods.

1. Apriori → Iterative, level wise
   - ↳ basic algo. → finding one L_k needs one full scan of DB (Disadvant
2. Pattern growth methods

Two step in Apriori

1. Join step

   $L_{k-1} \bowtie L_{k-1}$ is generated.

   No duplicates are generated.