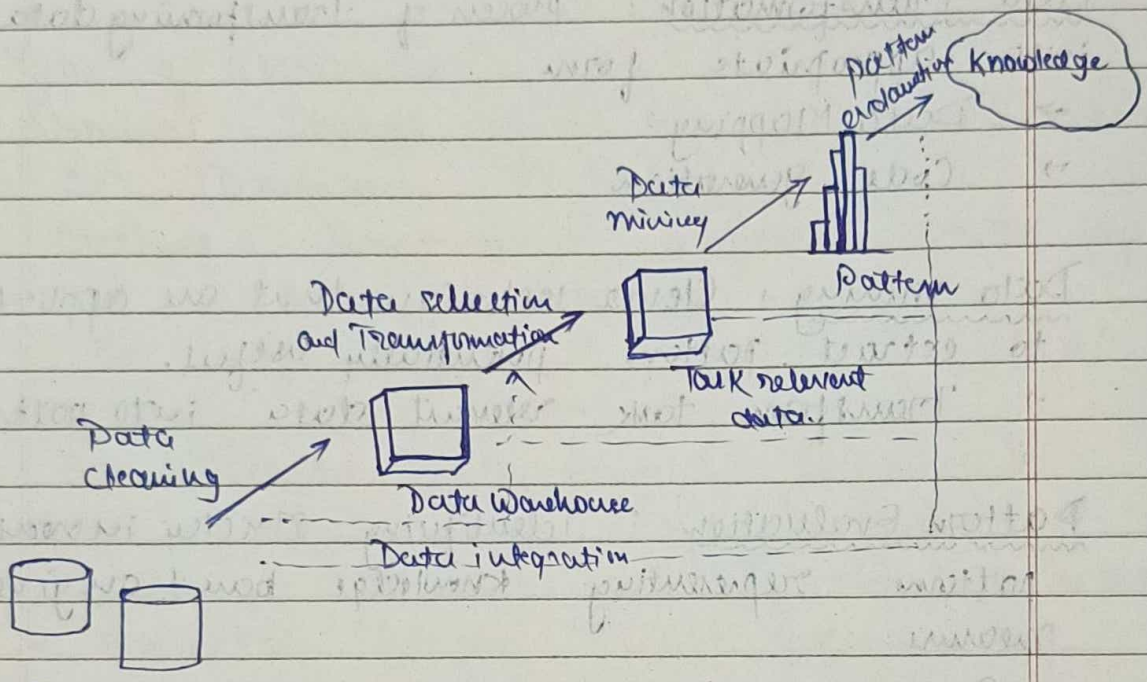
uses → * Automatic summarization of data
↑ * Extracting essence of information
Data Mining * Discovering patterns in raw data

Extraction of interesting (non-trivial, implicit, previously unknown) pattern of knowledge from huge amount of data.

Alternative Name: Knowledge discovery, knowledge extraction, data pattern analysis.

KDD process (Knowledge Discovery)



* Data cleaning: Removal of noisy and irrelevant data from collection.
  - Missing values
  - Noisy data
    ↳ Noise refers to modification of original value.
* Data integration: Heterogenous data are combined from multiple sources.
  - uses Data Migration tools.
  - uses Data Synchronization tools.

* Data Selection: Data relevant to the analysis is decided and retrived from data collection.
  -> Data selection using Neural Network
  -> " " Decision Trees
  -> " " Naïve Bayes
  -> " " Clustering, Regression

* Data Transformation: Procen of transforming data into appropriate form.
  -> Data Mapping:
  -> Code - generation

* Data mining: Clever techniques that are applied to extract patterns potentially useful.
  -> Transform task relevant data into pattern
  ->

* Pattern Evaluation: Identifying strictly increasing patterns representing knowledge based on given measures:
  -> Find interestingness score of each pattern.
  -> Uses summarization and visualization to make data understood by the user.

* Knowledge representation: Techniques which utilizes visualization tools.
  -> Reports
  -> Tables

* It is an iterative process where evaluation measures can be enhanced, mining can be refined.

Date: / /22

Outlier: A data object that does not comply with the general behaviour of the data.

Application of Data mining: Healthcare, Fraud detection, Education, Education, Lie Detection, Financial Banking, Manufacturing Engineering.

Data: Collection of data objects and attributes
(collection of attributes)

property of characteristic of an object.

* Types of attributes:
→ Nominal: Things which are fixed or universal. (unique)
Ex: ID number, Zip code.
→ Ordinal: Ranking of Interval needs by oneself accordingly.
Ex: weight category (small, medium, tall)
→ Interval: Fixed universal intervals.
Ex: Calender dates, temp. in Celcius or F.
→ Ratio: length, time, temp in kelvin
→ Binary: Attribute with only two states (0 and 1)
* Symmetric Binary: Both outcomes are equally important.
* Asymmetric Binary: Outcomes not equally important

⇒ Distinctness : = ≠
Order : < >
Addition : + -
Multiplication : * /

⇒ Nominal: Distinctness
Ordinal: Distinctness, Order
Interval: " , " addition
Ratio : " , " , " Multiplication

Page No.
Date: / /22

* Continuous attributes: comes from an infinite set (i.e. real numbers, you can make them as large or small). (floating point variables)
Discrete attributes comes from a finite set or countably infinite sets (i.e integers). (binary attribs)

Types of Missing Data
* MCAR: (Missing completely at random)
The missingness of the variable will be empirically tested based on analysis of other data from dataset.
* MAR: (Missing At Random)
Missingness to related to other variable
* MNAR: (Missing not at Random)
To find cause of missingness of data.

Data Preprocessing:
* Aggregation
* Sampling
* Dimensionality Reduction
* Feature Subset selection

Euclidean Distance
$(x_1, y_1), (x_2, y_2)$ dist $= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

Manhattan Distance.
$= |x_1 - x_2| + |y_2 - y_1|$

Minkowski Distance: $\Rightarrow$ dist $= \left(\sum_{k=1}^{n} |p_k - q_k|^r \right)^{1/r}$
$r = 1$ (Manhattan Dist.)
$r = 2$ (Euclidean Distance)
$r = 3$ ("Supremum") This is the max diff. bet^n any component of vectors.

Page No.
Date: / /22

* SMC = $M_{11} + M_{00} / (M_{01} + M_{10} + M_{11} + M_{00})$
* Jaccard (J) = $M_{11} / (M_{01} + M_{10} + M_{11})$

* Cosine Similarity:

$$\cos(d_1, d_2) = (d_1 \cdot d_2) / \|d_1\| \|d_2\|$$

Ex: d_1 : 1 2 5 0 2

d_2 : 3 1 2 7 1

$d_1 \cdot d_2$: 1*3 + 2*1 + ... + 2*1

$\|d_1\|$: 1*1 + 2*2 + ... + 2*2

$\|d_2\|$: 3*3 + 1*1 + ... + 1*1

Mean: $\frac{\text{Sum of all elements}}{\text{Total no. of elements}} = \frac{\sum_{i=1}^{n} x_i}{n}$

* Trimmed Mean:

Ex: Trimmed 20% mean for 60, 81, 83, 91, 99

$\Rightarrow \frac{81 + 83 + 91}{3}$

Median: Middle most element.

Mode: Most repeated : Unimodal
Bimodal
Multimodal

Grouped Data:

Estimated

* Mean: (Interval midpoint) = m and frequency $f_1, f_2 \ldots$

$$\therefore \text{mean} = \frac{m_1 * f_1 + m_2 * f_2 + \ldots m_n * f_n}{f_1 + f_2 + \ldots + f_n}$$

$\Rightarrow$ Estimated median

$$= L + \frac{(n/2) - B}{G} * w$$

L: lower class boundary of the group containing median

n: Total no. of values (sum of frequencies)

B: Commulative frequency of groups before median group

G: Frequency of median group.

w: group width.

Date: / /22

Page No.

Estimate Mode $= L + \frac{(f_m - f_{m-1})}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \times w$

* Symmetric vs. Skewed Data:

⇒ Symmetric: Mean, Median, Mode lies on same line

⇒ +ve Skewed Data: Mode < Median

⇒ -ve Skewed Data: Mode > median

* Standard Deviation

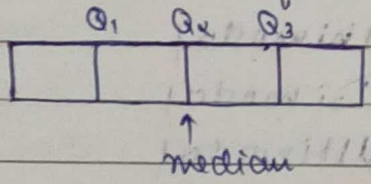
$$sd = \sqrt{\frac{\sum_i (d_i)^2}{N-1}} \quad \text{where } d_i = x_i - \bar{x}$$

($\bar{x}$ ← mean; $N-1$ ← Bessel's correction)

* Variance:

$$\sigma = (sd)^2$$

⇒ Quartiles: It divides an ordered data set into four equal parts.



Even Sample Size

median = 71 = Q2

64 63 64 64 70 72 76 77 81 81

Q1 = 64 Q3 = 77

Odd Sample Size

median = 72

63 64 64 70 72 76 77 81 81

$\frac{64+64}{2} = 64 = Q_1$ $\frac{77+81}{2} = 79 = Q_3$

Interquartile Range (IQR)

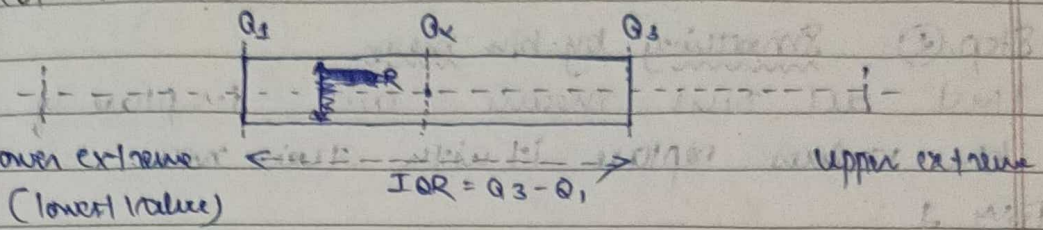
IQR = $(Q_3 - Q_1)$

Page No.

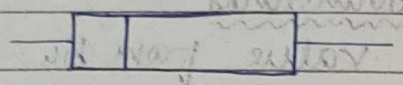
Date: / /22

Outliers: Below $Q_1 - 1.5(IQR)$
Above $Q_3 + 1.5(IQR)$ } This boundary is called Turkey fences

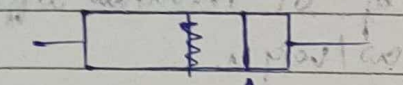
Boxplot :



Boxplot Analysis :



Right / positive skewed
$(Q_3 - Q_2) > (Q_2 - Q_1)$



Left / Negative skewed
$(Q_3 - Q_2) < (Q_2 - Q_1)$

Normalization

① Decimal Normalization.

$v_i' = \frac{v_i}{10^j}$ where j = no. of digits in greatest element of data set.

Note: Arrange given data into ascending order.

② Max-Min Normalization

$v_i' = \frac{v_i - \max(D)}{\max(D) - \min(D)}$ D: data set.

Noisy Data Handling

(I)

① Binning Method: (Applied to sorted data)

→ Step ① Partition into (equi-depth) bins
Bin 1 : // Equal beans
Bin 2 :
:

→ Step ② Smoothing by bin mean.
Find all bins means and in place of bin values replace it with their means.
Bin 1 :
Bin 2 :
:

→ Step ③ Smoothing by bin boundaries
Take lowest and highest value from bin then write the bin value as lowest / highest according to nearer element or minimum diff. betⁿ ith element and low / high.

Ex: 4, 8, 9, 15, 21, 21, 24, 25

Step ① Bin 1: 4, 8, 9, 15 mean = 9
Bin 2: 21 21 24 25 mean = 22.75 ≈ 23

Step ② Bin 1: 9 9 9 9
Bin 2: 23 23 23 23

Step ③ Bin 1: 4 4 4 15 (low = 4, high = 15)
8 and 9 are near to 4
Bin 2: 21 21 25 25
24 is nearer to 25

(II)
Clustering: It is a method of converting a group of abstract objects into classes of similar objects.

* To find centroid: (A – B)
(x1, y1) (x2, y2)

∴ Centroid = (mean(x1, x2), mean(y1, y2))

Page No.
Date: / /22

Ex:

	x	y
A	2	3
B	6	1
C	1	2
D	3	0

A-B (centroid) = $\left(\frac{6+2}{2}, \frac{3+1}{2}\right)$

= (4, 2)

C-D (2, 1)

→ Then Square Euclidean Distance

	A	B	C	D
A-B	5	5	9	5
C-D	4	16	2	2

→ as A is very near to CD then ACD could be a cluster.

Centroid of ACD = $\left(\frac{2+1+3}{3}, \frac{3+2+0}{3}\right)$

= (2, 1.6)

∴

	A	B	C	D
B	20	0	26	10
ACD	1.7	16.44	1.1	3.7

∴ B is very far from ACD ∴ it will not be in cluster with ACD.

(II)

Regression: It is a method to determine the relationship between a dependent variable and one or more independent variable.

① Linear Regression (best line to fit two variables)

$y = mx + b$ but here m we will take a

$\therefore y = ax + b$

$$a = \frac{n(\Sigma xy) - \Sigma x \Sigma y}{n(\Sigma x^2) - (\Sigma x)^2} \qquad b = \frac{\Sigma y \Sigma x^2 - \Sigma x \Sigma xy}{n(\Sigma x^2) - (\Sigma x)^2}$$

Page No.

Date: / /22

Handling Redundant Data In Data Integration

> Correlation Analysis:

$$r_{A,B} = \frac{\Sigma(A-\bar{A})(B-\bar{B})}{(N)\sigma_A\sigma_B} = \frac{\Sigma'(A_iB_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$$

N : no. of instance
$\bar{A}$ and $\bar{B}$: are mean
σ_A and σ_B : standard deviation
$\Sigma'(AB)$: sum of AB cross product.

- $r_{A,B} > 0$: A and B are positively correlated (if A ↑ then B ↑)
- $r_{A,B} < 0$: A and B are negetively correlated (If A ↑ then B ↓ vice versa)
- $r_{A,B} = 0$: independent (No correlation)

> χ^2 (Chi-square Test) :

- Find expected value.
- Find $\chi^2 = \frac{\Sigma(\text{Observed} - \text{Expected})^2}{\text{expected}}$
- Then compare χ^2 value with critical value of Chi-square.
  - If χ^2 < critical value → Null hypothesis (we don't agree)
  - If χ^2 > critical value → Alternate hypothesis.

Note: Significance level → given in question
Degree of freedom (df) : (R-1) * (C-1)

R : No. of Rows , C : no of columns

Example:

	Play chess	Not play chess	Sum
LSF	250	200	450
NLSF	50	1000	1050
Sum	300	1200	1500

Page No.
Date: / /

Expected frequency

$$E(\text{play chess, LSF}) = \frac{\text{count(play chess)} * \text{count(LSF)}}{N} = \frac{300 * 450}{1500} = 90$$

$$E(\text{play chess, NLSF}) = \frac{300 * 1050}{1500} = 210$$

⋮

Expected frequency Table

	Play Chess	Not play Chess
LSF	90	360
NLSF	210	840

$$\text{then } \chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(360-240)^2}{360} + \frac{(840-1000)^2}{840}$$

$$\chi^2 = 507.93$$

Given significance level = 0.001

degree (df) = $(2-1) * (2-1) = 1$

So from Chi-square Table

Critical value is = 10.828

∴ As computed value (507.93) is more than 10.828, hence, ~~prefer~~ them are correlated.