25 Friday
11 Ramazan

Data Mining and Warehousing

Storing Procedure

Store data extract into useful information Data mining
↙
Knowledge Data Discovery

⇒ Clean (Redundancy Removed) ~~data~~
↓
Duplication

Same data in Same table Or multiple table

26 Saturday 12 Ramazan / 27 Sunday 13 Ramazan

⇒ Data Integration
↓ data
Irrelevant paragraph / Ambuguidy

2) Selection (Normalization)
Or Transformation step by step
(i) Normalization Several Techniques
(ii) selection کو اس سے Data relevant ہو.
chor dain گے جو نا adap

Applications
Data

Data Pre Processing

Pre - Before task ہے پہلے اس سے ایسا کرنا، understand کو ڈیٹا
post - After flawful data اور ... کرے / Perform
Processing error data

Pre - Before Post after
fee before Salary after month

⇒ Cleaning
⇒ Integration DataBase File meaningful form
Different Resources سے ڈیٹا
کو لے گا

⇒ Transformation (RepAutan Removal)
Normalization
Dont Data repeat

⇒ Reduction (Volume reduce but meaning Not changed).
→ Discritization
→ Clarify
alphabet / Digit / Numeric

30 Wednesday
16 Ramazan

Age of Data
Bulk of data
Large Collection of Knowledge
Future Prediction through datamining

1960 Earlier
↓
1970 → Database

1970-80
Advance
database

Advance
Database
Analysis

Present
and Future

31 Thursday
17 Ramazan

Customer and
Consumer

Previous Data ⅓ future data سے Predicti[

Which Can be mined?
→ Characterization → Classes | entri
entri
2) Association and Correlation
→ Clarification and Regression entity → T
→ Cluster Analysis
→ Outlier Analysis

Table entries
↑
① Characterization — Data Characterization K
— Data Discriminization K
↓
comparison

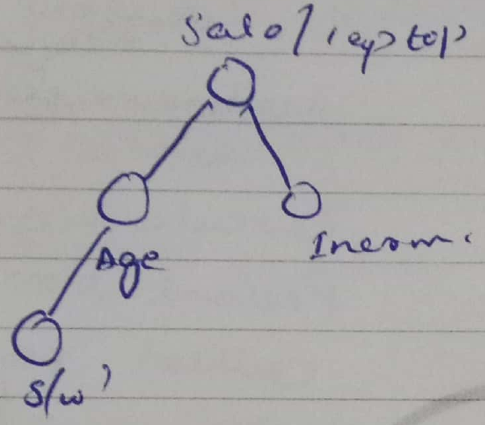
Data کو elaborate اور explain کرنا ڈیٹا کو کس طرح

ایک کلاس کے feature کو دوسرے کلاس کے feature کو compare کرنا

Classification on the basis of
↙ functi

Process of modeling
function

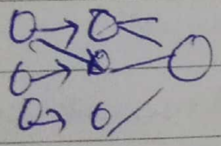
ایک ٹاسک ی



Neural Network

① middle layer Network input

Deep layer Network output
layer



feed forward

→ Regression.
↓
Numeric Analysis
Identify data

Taxonomy
↓
→ cluster → Same group of data Observation
No dataset boundaries

→ outlier → Noise / fraud
chance of error existence

S	S					
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

4 Monday
21 Ramazan

Data can be mined?

→ Class / concept Discrimination

→ mining frequent Pattern → Single or multi dimensional

Classification / Regression

↓

Supervised learning

↓

label data

→ Cluster Analysis

↓

unlabeled data

5 Tuesday
22 Ramazan

→ outlier Analysis

→ outlier Analysis

↓

Fraud Detection

/ Unusual activity

→ Purchasing

→ Location

↓ pattern frequency

Mathematically

↓

Calculation

Presentation is

called regression

Data & Regression is

Internet

↓

minimize similar

win

Intranet → group

↓

maximize similar

① Information retrieval

Reword or not Complex

unstructured data

adhoc

9 Surpervised Learning
Unsupervised Learning
Semisupervised learning
Active learning

→ Business Intelligence → manual task Perform

→ web Search engine
↓
→ Directories

Issues

→ Mining Methodology → Interdiscipilary
→ User Interaction
→ Efficiency and Scalability
→ Diversity of Data type
→ Data mining and Society

uncertainity
Pattern evaluation
Detection

handling of complex data type

User Interaction
↓
Friendly user interface
or usable

efficiency relate to performance
Do accurate work

Analysis Daily Life

8 Friday
25 Ramazan

Getting to know your data   combination of features

1. Attribute → Characteristics of data
↓
Major Two Types
① Numeric → digits
② categorical → Qualitative data

Numeric
Discrete (finite) boundary 0 – 9
continuous → Age, Time continuous

9 Saturday
26 Ramazan
/ 10 Sunday
27 Ramazan

Feature vector

First N } Bivariance
Last N

low intensity High

Binary A(0–1)
↓ off ↓ on

Categorial
Nominal
↓
Names
colors

ordinal
↓
Size of drinks } dependent
drink size

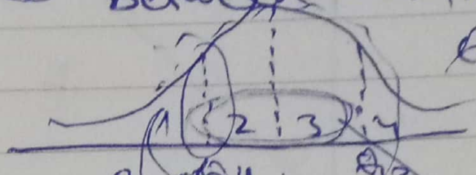
Center of Tendency
Multiple observation
center } middle Point

Mean

Median

Mode → most frequent occur observation

Range → Difference between max/minimum



Quantile
4

Bell Shape graph

each line Q3

Quantized

100 lines Draw
Percentile

Interquantile
IQR

$Q_3 - Q_1$ → Maximum − minimum

Standard Deviation

$6^2 = N^2$ → variance

$S'D = \sqrt{N^2} = N$

Variance $\sigma^2 = N^2$

$\sigma = S \cdot D = N$

variance

$$6^2 = \sum_{x=1}^{N} \left(\frac{x_i - \bar{x}}{n} \right)^2$$

1, 2, 3, 4, 5, 6, 7, 8, ~~9~~

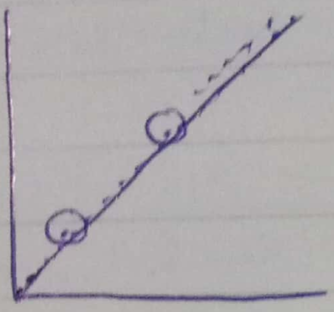
$$= \left(\frac{1 - 4.5}{8} \right)^2 + \left(\frac{2 - 4.5}{8} \right)^2 \cdots\cdots$$

$$= \frac{(9 - 4.5)^2}{8}$$

$$= \left(\frac{(1 - 4.5) + (2 - 4.5)}{8} \right)$$

13 Wednesday
1 Shawal

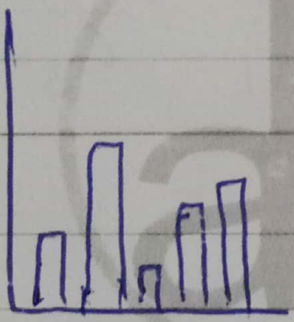
Data Visualization/Representation



Know reading
Univariant
Single, more

Q: Plot 2-dimenshon

Histogram



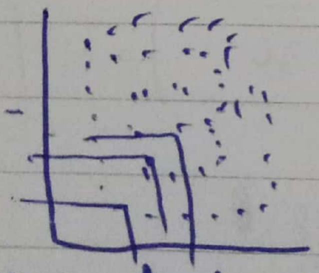
14 Thursday
2 Shawal

Bar graph
(Frequency graph/Chart)

Line of Code LOC

~~Standard~~ Plot Scattered graph

024

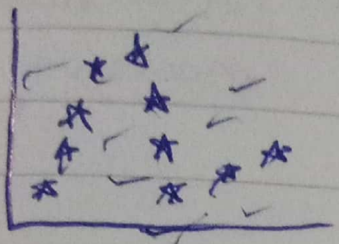


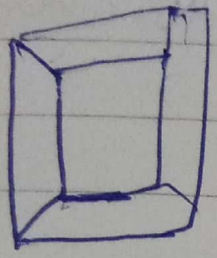
Pixel Oriented data
↓ dot

جہاں color زیادہ ہوں گے
وہاں زیدہ ڈیٹا ہی
0% availability

Pixel are Scattered no data are noise

Geometric Level → Different Shapes
awareness different data representation
↳ Color



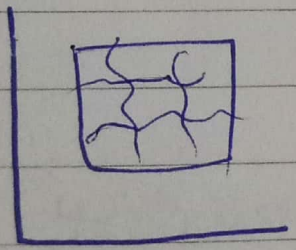


Scattered plot Matrix $\begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$

Icon based visualization
Chernoff

~~Her Hierarchie~~

Hierarical → Step-by-study of data



Measuring of data Similarity and Disimilarties Matrix

0.0....... Disimilar
0.5 ----1 Similar,

Two attributes [illegible]
(x_i, y_i)

one item attributes compare to another item attribute is called Similarity and dissimilarities matrix

Name, Color, Age,

18 Monday
6 Shawal

Proximity Measures Attribute of Nominal attribute.

$$d_{i,j} = \frac{(P-M)}{P}$$

M → Total number of matching attribute

P → Total number of Attributes

Dif
Ordinal Attribute
Dissimilarities

$$d(i,y) = \lim_{h \to \infty} \left(\sum_{f=1}^{P} |x_{if} - x_{i}|^{h} \right.$$

19 Tuesday
7 Shawal

$$L_\infty = \max_{f} |x_{if} - x_{i'f}|$$

Uniform norm

Cosine Similarity

ایک ٹیم کا دوسری ٹیم کے ساتھ کتنے میچ ہوئے ہیں

Total comparison

Wednesday 20
8 Shawal

Chapter #3

Data Processing

ڈیٹا کو accurate بنانے کیلئے

easy to understand and environment or user friendly → e.g windows

Time line

Completeness

Consistency → Flow

→ unusual Data

Data Cleaning → Noise ہے (Unwanted effect)

جو انفارمیشن ہمارے لیے helpful نہ ہو وہ ہمارے لیے Unwanted

Noise ہوتی ہے

Thursday 21
9 Shawal

Data Integration

Dimension Reduction → Features or Attributes

Reduction (volume reduced but meaning not changed)

Missing values

Transformation

Binning

Bin Size of bucket in each type

Recycle Binning

equally distribution by mean

$$\frac{8+4+15}{3} = \frac{27}{3} = 9$$

22 Friday
10 Shawal

Equally
Binning by boundary (Median)

Regression

Linear Regression Two → one to one de
Multiple Regression One to many

Clustering → unlabel Data

Discripeny → Detection user don't response
Data is more important to

23 Saturday 11 Shawal / 24 Sunday 12 Shawal

you but you don't response

e) Resumblance
→ Data clean

Metadata → Data about data
Relevant Data → Metadata are Same

Field Overloading
Unique rule → e.g unique id or attributes
specific person

JULY

M	T	W	T	F	S	S
30	31					1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29

June 2018
Shawal 1439

Monday 25
13 Shawal

→ Consective rule

→ Data discovery Analysis

Data Scubbing rule and
Address

Tuesday
14 Shawal