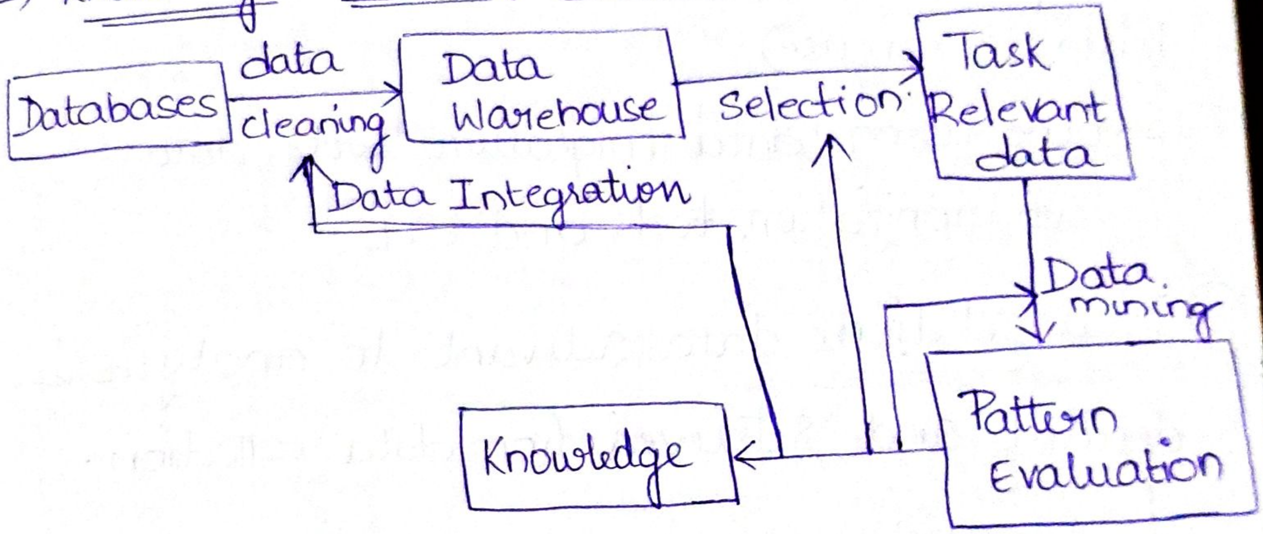
27/09/2024

UNIT-1

→ Data mining is the extraction of interesting patterns or knowledge from huge amount of data.

→ Data mining is also called knowledge discovery in databases, knowledge extraction, business intelligence etc.

→ Knowledge Discovery Process: (KDD)



→ Ex: Web Mining Steps:

↳ Data cleaning, data integration, warehousing the data, data cube construction, data selection, data mining, presentation of results, knowledge.

* Steps included in KDD process:

① Data cleaning: removal of noisy and irrelevant data from collection.

- ↳ missing values
- ↳ noisy data
- ↳ data discrepancy detection & data transformation tools.

② Data integration: heterogeneous data from multiple sources combined in a common source (data warehouse).

↳ done using data migration tools, data synchronization tools and ETL

③ Data Selection: data relevant to analysis is decided and retrieved from data collection.

↳ done using neural network, decision trees, Naive Bayes clustering & regression methods.

④ Data transformation: transforming data into appropriate form.

↳ 2 steps:

- ↳ Assigning elements from source to destination : Data Mapping
- ↳ Code generation: transformation program.

⑤ Data Mining: extract patterns potentially

↳ decided using classification con characterization

⑥ Pattern Evaluation: identifying strictly ↑sing patterns representing knowledge based on given measures.

⑦ Knowledge Representation: score of each pattern and uses summarization & visualization.

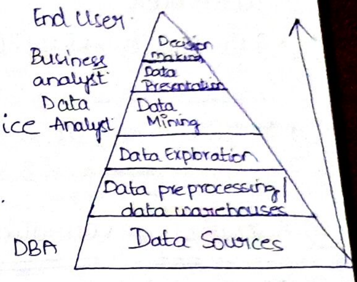
* Advantages of KDD:

① ↑sed efficiency.

② better customer service

③ fraud detection

④ Predictive modelling.



* Multi-dimensional view of data mining:

→ Data to be mined : database data.

→ Knowledge to be mined: characterization, discrimination, association, classification etc.

→ Techniques used: Data warehouse (OLAP), ML, statistics, pattern recognition etc.

→ Applications adapted: telecommunication, banking, fraud analysis etc.

* **Kinds of data:** data streams, sensors data time-series data, temporal data sequence data etc. Multimedia, Text, Object-relational databases etc.

* **Functions of data mining:**

① **Generalization:**

* Information integration and data warehouse construction.
* Data cube technology: Online Analytical processing (OLAP).
* Multi-dimensional concept description: Characterization & discrimination.

② **Association & Correlation analysis:**

* Frequent patterns.
* association, correlation v/s causality.

③ **Classification:**

* Classification & label prediction.
  - construct models.
  - describe & distinguish classes (or) concepts
  - predict some unknown class labels.

* Typical methods:
  - decision trees, support vector machines
  - Neural networks, pattern based classification

* **Applications:**
  - Fraud detection, classifying stars

④ **Cluster Analysis:**
- unsupervised learning
- group of data form new categories
- maximizing intra-class similarity & minimizing inter-class similarity

⑤ **Outlier Analysis:**
- Outlier: a data object that does not comply with general behaviour of the data.
- methods: product of clustering, regression analysis.

* **Evaluation parameters:**
  → descriptive v/s predictive
  → coverage → Accuracy → Timeliness
  → Typicality v/s novelty.

* **Confluence of Multiple Disciplines:**

→ ML, Pattern recognition, statistics, applications, algorithm, database technology, HPC, visualization.

→ used due to:
① Tremendous amount of data
② High dimensionality of data.
③ High complexity of data.
④ New & sophisticated appons.

* Applications of DM:

→ Web page analysis

→ Collaborative analysis

→ Basket data analysis to targetted marketing

→ Biological & medical data analysis

* Issues in data mining:

1. Mining Methodology - handling noise, uncertainity
2. User interaction - visualization & user interaction
3. Efficiency & scalability - parallel, distributed, stream
4. Diversity of data types - complex data types
5. Data mining and society - impacts of DM, privacy.

* Types of data sets:

→ Record: Relational records, data matrix.. etc.

→ Graph and network: WWW, molecular structures.

→ Ordered: video data, genetic sequence data.

→ Spatial, image and multimedia: image data, video data.

* Important characteristics of Structured data:

1. Dimensionality
2. Sparsity
3. Resolution
4. Distribution.

→ Data object represents an entity.

→ Data objects are described by attributes

→ Ex: Sales database, medical database

→ Attribute: a data field, representing a characteristic or feature of a data object

Types:

1. Nominal: categories, states, names of things

   Ex: marital status, occupation, zip codes.

2. Binary: 2 states = 0/1

   ↳ Symmetric: both becomes equally important.

   Ex: gender.

   ↳ Asymmetric: outcomes not equally important

   Ex: medical test (+ve v/s -ve)

3. Ordinal: values have a meaningful order but magnitude b/w successive values is not known.

→ Numeric Attribute Types:

(Integer (or) real-valued)

* Interval: measured on a scale of equal-sized units.

Ex: Temp in °C or °F, dates

* Ratio: inherent zero-point.

Ex: Temp in K, length, counts.

→ Discrete v/s continuous attributes:

1. Discrete: Has only a finite (or) countably infinite

set of values (Integer variables)

Ex: zip codes, profession.

(2) Continuous attributes: Has real numbers as attribute values. (float-pointing

Ex: temperature, height (or) weight.

* Basic Statistical descriptions of data:

→ central Tendency, variation and spread.

→ median, max, min, quantiles, outliers, variance etc.

→ Boxplot (or) quantile analysis on sorted intervals. as well as transformed cube.

*→ Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\mu = \frac{\sum x}{N}$

Weighted arithmetic mean: $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

*→ Median:

$$= L_1 + \left(\frac{n/2 - (\sum freq)l}{freq_{median}}\right) \times width$$

*→ mode = 3median − 2 mean.

*→ variance ⇒ $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

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

* Properties of normal distribution

→ from $\mu - \sigma$ to $\mu + \sigma$: about 68%.

→ from $\mu - 2\sigma$ to $\mu + 2\sigma$: contains about 95%.

→ from $\mu - 3\sigma$ to $\mu + 3\sigma$: contains about 99.7%.

* Graphic displays of Basic Statistical descriptions:

→ Boxplot: graphic display of 5-number summary.

→ Histogram: x-axis: values, y-axis: representation frequencies.

→ Quantile plot: each x_i is paired with f_i

→ Quantile-quantile (q-q) plot: graphs the quantiles of one univariant distribution against the corresponding of another.

→ Scatter plot: plotted as points in the plane.

* Data visualization:

→ Importance:

*→ Gain insight

*→ Provide qualitative insight

*→ Search for patterns

*→ Help find suitable regions & parameters.

*→ provide visual proof.

→ Categorization of visualization methods:

1. Pixel-oriented visualization techniques.
2. Geometric projection "
3. Icon-based "
4. Hierarchical "
5. Visualizing complex data & relations.

(1) → Income, Credit Limit, transaction volume, age.

(2) methods: direct visualization, scatterplot & scatterplot matrices, prosection views, hyperslice, parallel coordinates.

(3) methods: Chernoff faces, Stick figures.
general: shape coding, color icons, tile bars.

(4) methods: dimensional stacking, tree map, cone trees, info cube.

* Similarity and dissimilarity:

→ Similarity: Numerical measure of how alike 2 data objects are

range = [0,1]

→ Dissimilarity: Numerical measure of how different 2 data objects are.

- min. dissimilarity = 0 ; upper limit varies.

→ proximity refers to a similarity (or) dissimilarity

* Data Matrix:
  - → n data points with p dimensions.
  - → 2 modes.

* Dissimilarity Matrix:
  - → n data points, but registers only the distance.
  - → triangular matrix
  - → single mode.

* Proximity Measure for nominal attributes:

→ can take 2 (or) more states.

Method-1: Simple Matching

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

Method-2: large no. of binary attributes

→ A contingency table for binary data:

	1	0	sum
1	q	r	q+r
0	s	t	s+t
Sum	q+s	r+t	p

* distance measure for

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

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

→ Jaccard coefficient = $\frac{q}{q+r+s}$.

*** Dissimilarity b/w binary variables:**

Ex:

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4	
Jack	M	Y	N	P	N	N	N	P=1
Mary	F	Y	N	P	N	P	N	Y=1
Jim	M	Y	P	N	N	N	N	N=0.

$$d(\text{Jack}, \text{Mary}) = \frac{r+s}{q+r+s} = \frac{0+1}{2+0+1} = \frac{1}{3}.$$

$$d(\text{Jack}, \text{Jim}) = \frac{r+s}{q+r+s} = \frac{1+1}{1+1+1} = \frac{2}{3}.$$

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

*** Standardizing Numeric Data:**

$$z = \frac{x - \mu}{\sigma}$$

x = raw score to be standardized.

μ = mean of population

σ = standard deviation.

→ minkowski distance: Slide -58

$$d(i,j) = \sqrt[h]{(x_{i1}-x_{j1})^h + (x_{i2}-x_{j2})^2 + \cdots + (x_{ip}-x_{jp})^h}$$

i $(x_{i1}, x_{i2}, \ldots x_{ip})$ } is order

j $(x_{j1}, x_{j2}, \ldots x_{jp})$

properties:

→ $d(i,j) > 0$ & $d(i,i) = 0 \Rightarrow$ +ve definiteness

→ $d(i,j) = d(j,i) \Rightarrow$ Symmetry

→ $d(i,j) \leq d(i,k) + d(k,j) \Rightarrow$ Triangle Inequality

→ Manhattan distance:

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

→ Euclidean distance:

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

*** Cosine Similarity:**

$$\cos(d_1, d_2) = \frac{(d_1 \cdot d_2)}{|d_1| \cdot |d_2|}$$

Ex: $d_1 = (5,0,3,0,2,0,0,2,0,0)$
$d_2 = (3,0,2,0,1,1,0,1,0,1)$ } $\cos(d_1, d_2) = 0.94$

$d_1 \cdot d_2 = 5*3 + 0\times0 + 3\times2 + \cdots + 0\times1$
$= 25.$

$|d1| = \sqrt{(5\times5 + 0^2 + 3^2 + \cdots + 0^2)} = 6.48$

$|d_2| = \sqrt{(3\times3 + 0^2 + 2^2 + \cdots + 1^2)} = 4.12$

→ Z-Score: & Min-Max Normalization:

data = 1000, 2000, 3000, 5000, 9000.

Min = 0 , Max = 1.

min = 1000 max = 9000.

$$v = \frac{x - min}{max - min} \Rightarrow$$

0
0.125
0.25
0.5
1

$$Z = \frac{x - \mu}{\sigma}$$

$$\mu = \frac{20000}{5} = \boxed{4000.}$$

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{n-1}}$$

$$= \sqrt{\frac{(1000-4000)^2 + (2000-4000)^2 + \cdots (9000-40}{4}}$$

$$\boxed{= 2489.97}$$

$$\Downarrow$$

-1.204
-0.803
-0.4016
0.4016
2.008.