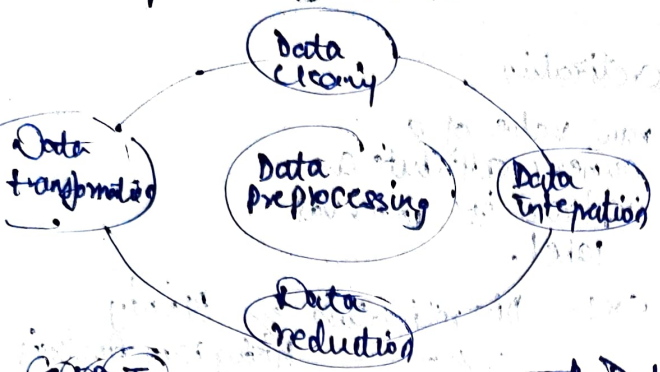
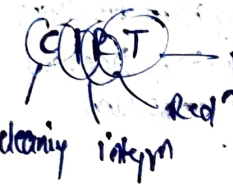
Data Pre-processing

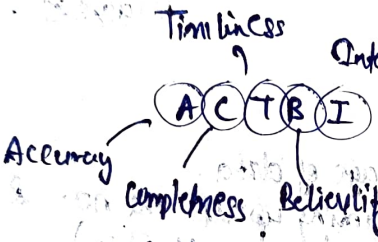
↳ is a data mining technique which converts raw data into understable format. Real-world data is often incomplete, inconsistent, or lacking in certain trends and contain error. Data pre-processing is proven method to resolve such issues.



→ Data processing have 4 major tasks.
- Data cleaning
- Data integration
- Data Reduction
- Data transformation



→ Data pre processing is important to achieve accuracy (whether data entered is correct or not).



- Completness (—//— data is available or not recorded)
- Timeliness (data should be updated correctly)
- Believlibly (—//— be trustable)
- Intrepetablity (understablity of data)

→ Data cleaning

↳ is the process to remove incorrect data, inaccurate data from incomplete data dataset.

↳ it also replaces the missing values.

↳ it also smooth out the noise while identifying outliers.

↳ correct the inconsistency.

→ Handling Missing values:

↓

Ignore the row (tuple)

→ used when class label or primary key is missing

→ not very effective unless tuple contain several attribute with missing value.

→ 'Not available' or 'NA' can be used to replaced the missing values.

→ filling manually.

→ time consuming
→ not feasible

→ use global constant

→ replace all missing attribute by same constant like 'Unknown' or '∞'.

→ simple but not foolproof.

→ use a measure of central tendency (1)

→ use of mean or median to fill the missing value

→ for normal symmetric data distribution mean is used.

→ for skewed →/- dist median is used.

→ use the most probable value to fill in the missy value.

→ determined with regression, inference base tools using Bayesian formalism or decesion tree.

→ Noisy Data : means random error or containing unecessary data points

(1) Binning → This method is to smooth or handle noisy data. first data is sorted and then sorted value are seperated and stored in the form of bins.

→ 3 method of smoothing data.

Smoothing by bin mean method

→ value in bin replaced by mean.

Smoothing by bin medianmethod

→ value in bin replaced by median.

Smoothing by bin boundary

→ max & min value are taken out of bin & replaced by closest boundary value.

(2) Regression → used to smooth data & handle data when unnecessary data is present.

→ for analysis purpose regression helps to decide the variable.

(3) Clustering → used for finding the outliers & also in grouping the data.

→ used in unsupervised learning.

→ Data Reduction → helps to reduction of volume of the data which make analysis easier and produces almost the same result.

→ helps to reduce storage space.

→ 3 techniques.

- Dimensionality red^n
- Numerosity red^n.
- Data comp^n.

① Dimensionality redu^n.

↳ process of reducing the no. of variables (random) or attributes under consideration.

↳ Some method.

- Wavelet transform. (sine wave to discrete fun^n)
- Principal Component Analysis (vector red^n)
- Attribute Subset Collec^n (removing irrelevant or redundant attributes)

② Data red^n Strategies Numerosity reduction :-

↳ process of removing original data by alternative or smaller form of data repr^n.

→ 2 type

Parametric method

↳ only data parameter are stored

↳ ex. regression or log linear model.

non-Parametric method.

↳ stores reduce repres^n of data

↳ ex. histograms, clustering & sampling.

③ Data Compression

↳ applied to obtaine a reduced or compressed representation of the original data.

↳ reconstructing original data from compressed data without any loss of info^n is known as lossless.

↳ —//— with some loss of info^n. —//— lossy.

→ Data Transformation.

↳ Change made in the format of structure of data is known as Data Transformation.

↳ Data is transformed so that mining process may be more efficient & patterns found may be easier to understand.

Strategies.

Smoothing, Attribute Construction, Aggregation, Normalization, Discretization

Smoothing
- Used to remove noise from data
- technique include binning, regression, & clustering.

(2)

Attribute Construction
- also known as feature Construction
- where new attrib are constructed & added from given set of attribute.

Normalization
- attribute data are scaled so as to fall with smaller range such as -1.0 - 1.0

Aggregation
- where summary or aggregation oprn is applied
- used in constructing data cube for analysis of data at multiple level of abstraction.

Discretization
- raw value of a numeric attribute are replaced by interval label.
- ex:- Discretization by Binning
  - // ——— histogram Analysis
  - // ——— by cluster, Decision Tree and Correlation Analysis.

Mean $= \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$ average of data. (summing up all the no. & then dividing up by no. of obs.)

$= \frac{10+20+30+40+50}{5} = 30$ mean.

Or ~~10~~ 1 -10, 20, 30, 40, ~~200~~

Median

$\frac{n}{2}$ (even) $\frac{n+1}{2}$ (odd) ∴ n - no. of obs?

10, 20, 30, 40, 50

$\frac{n+1}{2} = \frac{5+1}{2} = 3 \Rightarrow$ Median = 30

10, 20, 30, 40, 50, 60

$\frac{n}{2} = 3 \rightarrow$ 30, 40 $\rightarrow \frac{30+40}{2}$

Median = 35

Mode

1, 3, 4, 6, 7, 3, 3, 5, 10, 3

Mode = 3 {highest frequency}

IQR $Q_3 - Q_1$

(second half) (first half)

Or 48, 52, 57, 61, 64, 72, 76, 77 81, 85, 88

48, 52, 57, 61, 64, 72 — Q_1

Q_1 = median of 1st half = $\frac{57+61}{2}$

72, 76, 77, 81, 85, 88 — Q_3

Q_3 = median of second half = $\frac{77+81}{2}$

Range

= Maximum value − Minimum value

32, 41, 28, 54, 35, 26, 23, 33, 38, 40

54 − 23 = (31) Range

Standar deviation

↳ deviation from mean value.

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

n :- no. of popul?

μ = mean = $\frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$

Variance

:. squared deviated

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{n}$$ Popl^n variance ∴ μ = mean

$$\sigma^2 = \frac{\Sigma (x - \bar{x})^2}{n}$$ Sample varian $\bar{x}$ = average.

→ Box Plot

- standardized way of displaying the distribution of data based on five no. of summary ("minimum" first quartile (Q_1) median third quartile (Q_3) "maximum")
- It can tell you about your outliers & what are their value.
- also tell about your data is symmetrical, skewed, grouped.