mod 2

# Data Preprocessing – its a technique that involves transforming raw data into a useful and efficient format.

↳ Need –

Real world data are generally –

→ Incomplete – missing certain attributes of importance

→ Noisy – containing errors or outliers

→ Inconsistent – containing discrepancies in codes or names

which may lead to false analysis, so preprocessing is required.

Data preprocessing provides – accuracy – completeness – consistency – timeliness & – Interpretability to data.

# Major Tasks in Data Preprocessing –

1. Data Cleaning – filling in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
2. Data Integration – Integration of multiple databases, data cubes & files.
3. Data Reduction – dimensionality red$^n$, numerosity red$^n$, & data compression
4. Data Transformation & Discretization – normalization & concept hierarchy generation.

# DATA CLEANING –

Data in real world is dirty, there is potentials of incorrect data, eg. its

→ incomplete : lacking of attribute or lacking of value in attribute.

- it may be due to –
  - → equipment malfunction
  - → inconsistent with other recorded data
  - → certain data may not be considered imp at time of entry.
  - → data not entered due to misunderstanding.
- how to handle such data –
  - → replace with global const
  - → fill in it with attribute mean.
  - → fill with attribute mean for all samples belonging to same class.
  - → fill it with most probable value using decision tree.

→ Noisy – random error or variance in a measured variable

- it may be due to –
  - → data entry problems
  - → faulty data collection instruments
  - → data transmission problems
  - → technology limitation
- how to handle noisy data –
  - → binning : sort the data and partition into bins & smoothen it
  - → regression : smooth by fitting data into regression fn.
  - → clustering : detect and remove outliers
  - → combined computer & human inspection – detect suspicious values and check by human.

# DATA INTEGRATION –

Combines data from multiple sources into a coherent store. It includes detecting and resolving data value conflicts that may arise due to different scales or representations.

↳ Problem – Data Redundancy i.e

- same object or attribute may have diff names in diff DB.
- one attribute may be a derived attribute in another table

↳ Solution – Correlation & covariance analysis

- Correlation –

1) Numerical Data –

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_x \sigma_y} = \frac{\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}}{(n-1)\sigma_x \sigma_y}$$

Where $n$ = no. of tuples

$\bar{x}, \bar{y}$ = respective means

$\sigma_x, \sigma_y$ = respective standard deviations

NOTE – if $r > 0 \Rightarrow$ (+)ve relation
$r < 0 \Rightarrow$ (–)ve relation
$r = 0 \Rightarrow$ no relation (independent)

2) Nominal Data – $\chi^2$ test.

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

larger the value of $\chi^2$ more likely they are related.

• Covariance –

$$Cov(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$r = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

NOTE → if $Cov > 0 \Rightarrow$ x & y tend to be larger than expected values

if $Cov(x,y) < 0 \Rightarrow$ if x is more than expected value then y will be less than expected value & vice-versa

~~if Cov(x,y) = 0~~

# DATA REDUCTION –

its a process of obtaining a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost same) analytical results.

→ Strategies :

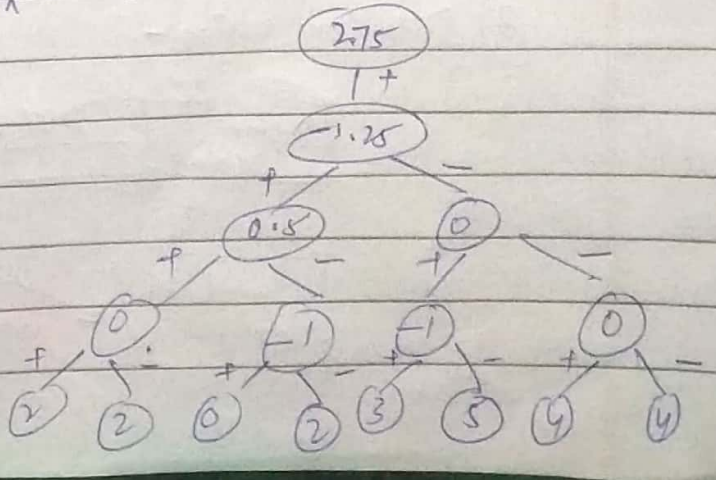
1) Dimensionality Reduction – removes unimportant attributes.

⇒ Wavelet Transformation – decomposes a signal into diff frequency subbands. Data is transformed to preserve relative distance b/w objects at diff levels of resolution.

S = [2, 2, 0, 2, 3, 5, 4, 4]  $(\frac{a-b}{2})$

| Resolution | Averages $(\frac{a+b}{2})$ | Detail co-efficients |
|---|---|---|
| 4 | [2, 1, 4, 4] | [0, −1, −1, 0] |
| 2 | [1.5, 4] | [+0.5, 0] |
| 1 | [2.75] | [−1.25] |

$S_\lambda$ = [2.75, −1.25, 0.5, 0, 0, −1, −1, 0]



⇒ PCA (Principal Component Analysis) – Find a projection that captures the largest amount of variation in data. the original data is projected onto a much smaller space, resulting in dimensionality red".

2) Numerosity Reduction – Reduce data volume by choosing alternative, smaller forms of data representation.

⇒ Parametric methods – [Regression]

- Linear – data modeled to fit a straight line
- multiple – allows a response variable Y to be modeled as a linear fn of multi-dimensional feature vector
- Log-linear – approximates discrete multidimensional probability distributions.

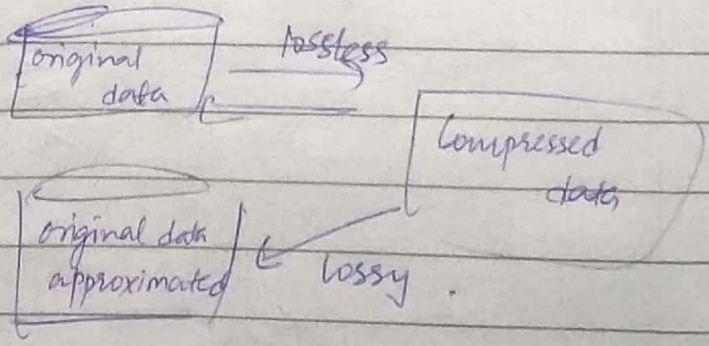
⇒ Non-Parametric –

- Histogram Analysis – divide data into buckets & store the sum/average for each bucket.
- Clustering – partition data set into clusters based on similarity, and store cluster representation.
- Sampling – obtaining a small sample S to represent the whole data set N.

↳ Types –

1. Simple random sampling
2. Sampling without replacement
3. Sampling with replacement
4. Stratified sampling – partition dataset, and draw samples from each data propostionally.

3) Data Compression –



# DATA TRANSFORMATION — A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new ones.

⇒ Methods –

1. Smoothing – [Binning] [partitioning the data]
   - equi-depth/frequency: each bin has same no. of samples
   - equi-width – W = (highest value – lowest value)/no. of intervals
   - bin means – fill value with means of value
2. Aggregation – [cube Lattice]
3. Normalization –
   - min max –

$$v' = \frac{v - min_i}{max_i - min_i} (new\ max - new\ min) + new\ min$$

   - Z score –

$$v' = \frac{v - \mu}{\sigma}$$ ($\mu$ → mean, $\sigma$ → SD)

   - Decimal scaling –

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

where $j$ is smallest integer such that $Max(|v|) < 1$

4. Discretization –
   - Correlation
   - Covariance
   - Concept Hierarchy Generation: organizes concepts (i.e. attribute values) hierarchically and is usually associated with each dimension in a data warehouse. It facilitates drilling & rolling in data warehouses to view data in multiple granularity.