Lecture 4:- Visualization Model, pipeline & data preparation

- Visual Mapping must be:-
  - Computable (math) visual = f(data)
  - Comprehensible (invertible) data = f⁻¹(visual)
  - Creative
- Visualization pipeline → • Data preparation step within Knowledge discovery process.

Collection → Pre-processing & transformation → Mapping → rendering → Visualization output → user interaction.

also known as ~~data preparation~~.

- Pre-Processing is must for high data quality which impacts reliability & effectiveness
  - Enhance accuracy
  - ~~Enhance~~ Completeness (Fill in missing values & ensure all data is present)
  - maintains consistency
  - Improves Timeliness
  - Boosts believability

- ① Pre-Processing → • Data cleaning • Data integration • Data reduction • Data transformation & data discretization

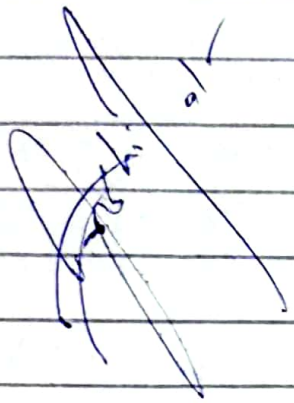
Data cleaning → maybe • missing data • noisy, errors or outliers • inconsistent

- Reason for Missing data → • Equipment malfunction (data collection devices failed, leading to gaps in data) • Inconsistent data (data might be deleted) • misunderstanding • Perceived unimportance.

If not handled then,
- incomplete analysis
- inaccurate models

How to handle?
- Ignore tuple
- Fill missing values (use mean (symmetric) median (skewed))

- Most likely value:-
  use bayesian or decision trees to predict value
- -- Inference-Based → • Bayesian
  methods • decision tree
- -- Identify relationship among Variables → • Linear regression
  • Multiple regression
  • Non-linear regression



- -- Nearest-Neighbour Estimator
  • Fill missing values with most frequent value (categorial data) or average value (numerical data)

- -- Noisy data → • Random error or variance in measured variable.
  • Handle noisy data by Binning, Regression, Clustering
  • Binning → • Sort the data
  • partition into bins
  • Smooth bins by → mean, median, boundary.

Example:-

data = (8, 4, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34)

① Sort → 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

② Bin 1 → [4, 8, 9, 15] $\frac{8+9}{2} = 8.5$

Bin 2 → [21, 21, 24, 25]

Bin 3 → [26, 28, 29, 34]

Bin Means →
[9, 9, 9, 9], [23, 23, 23, 23],
[29, 29, 29, 29]

Bin Medians → [8.5, 8.5, 8.5, 8.5], [23, 23, 23, 23], [28, 28, 28, 28]
(Adding middle number found divided by 2)

Bin Boundaries → (identify lowest and highest value)
[4, 4, 4, 15]

metadata → name, meaning, datatype, range, null values.

Predicts one attribute based on other.

- Regression → • used to smooth data by fitting into regression function
  - Linear regression → • Finding the best line that fits two attributes goal is to minimize distance b/w data points and the line.
- Clustering → detect & remove outliers.
  - values outside set of clusters.

(2)
- Data integration (combining data from multiple sources into one) → - EIP (Entity identification problem)
  - while integrating, EIP is a great challenge.
  - Schema integration & object matching
    - combining data from different sources which use different formats.
  - Metadata can help avoid errors in schema integration and transform data.
  - When matching attributes from two databases, structure of data should be checked.
- Handle redundant data → • by detecting through correlation analysis and covariance analysis.

(3)
- Data reduction → • reduced representation of data set (smaller in volume but produced same analytical results)
  - ↓ use
  - b/c move data & complex data analysis take very long time to run.

Strategies
- Dimensionality reduction
- Wavelet transforms
- PCA
- Feature selection, creation.

•– Numerosity reduction → • Parametric (Regression, log linear model)
• Non-Parametric (Histogram, clustering sampling)
• data cube aggregation.

•– Data compression → lossless (reconstruction, no loss of information)
lossy (reconstruct only approx of orignal data)

i) Dimensionality reduction (① Wavelet transforms ② PCA ③ Supervised & non-linear)

•– curse of dimensionality → • no. of features in data ↑, data becomes harder to analyze
• difficult to measure distances between points, imp for clustering ev outliers.

a) •– Dimensionality reduction → • removes noise in data and unnecessary features.
• less memory used & data processing faster.
• easier to visualize data. by showing in fewer dimension (3D to 2D)

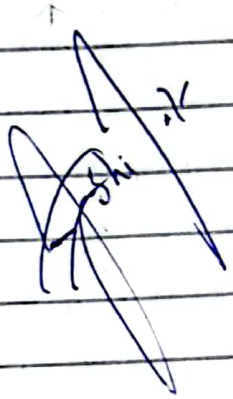
b) •– Principle Component analysis → (PCA)
(• reduce no. of features in data while keeping most important information)
• Works for numerical data.
• PCA handles sparse data better than wavelet transforms.

steps.
• Normalize input data.
Min-Max → $X' = \frac{X - X_{min}}{X_{max} - X_{min}}$
(Final for every entry in the table)
• Covarience matrix / Z-score standardization
• Eigen value, Eigen vector
• Select principle components
• Transform data.
• Sort Components by significance
• reduce data dimensionality.

c) :- Numerosity reduction →

- Parametric methods → • data follows a specific model.
  • estimate and store parameters and discard data (except possible outliers)
  (eg: Regression, non-linear model) → Consider each tuple as a point in an n-dimensional space.
- Non-parametric → • do not assume models.
  (e.g., histograms, clustering, sampling).

• Regression Analysis (way to understand relationship b/w dependent & independent variables)
  • Parameters are estimated to give best fit.
  • Best fit is evaluated by using least squares method.
  • Used for prediction, hypothesis testing & modeling for casual relationships.



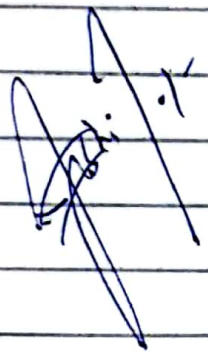
• Histogram Analysis (Divide data into buckets and store avg sum of each bucket)
  Rules → Equal width & frequency.

- Clustering (store cluster representation)
- Sampling (obtain small data to represent whole dataset).
  ↓
  may not reduce database I/O

d) :- Data Compression → String Compression (lossless)
  Audio/Video Compression (lossy)
  ↓
  Dimensionality & numerosity may be forms of data compression.

(4) Data Transformation →

- **Descriptive Statistics**
  - Range, Min / Max
  - Average (one measure of central location in data set)
  - Median
  - Mode (Another measure of central location in data set).
- **Distribution Statistics**
  - How meaningful is chosen average?
    - Variance
    - Std (√Variance)
    - Histogram / Normal distribution



- When data skewed, Mean & SD can be misleading.
- Skewness
  - $Sk > |1|$ (non-symmetrical)
  - Negatively Skewed (Mean < Median)
  - Positively Skewed. Mean > Median
- Histogram shows (center, spread, skewness, outliers, multiple modes in data).
  - ↓ for small dataset Histogram can be misleading.
  - Histogram tells nothing about relation among variables
- Normalization → normalize data b/w [1, -1] or [0, 1]

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

Discretization (Divide range of continuous attribute into intervals) →

- Binning (unsupervised)
- Histogram (un) analysis
- Clustering (un) analysis
- Decision tree (supervised) (uses entropy)
- Correlation (un)