Data Preprocessing

Agenda

- What and Why data preprocessing?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

What is Data Preprocessing?

It is data mining technique that involves transforming raw data into an understandable format.

Why Data Preprocessing?

- Data in the real world is dirty
  - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - Incorrect/Error: Collection instrument may be faulty, mandatory field of personal information may contain wrong data
  - Noisy: containing errors or outliers
  - Inconsistent: containing discrepancies in codes or names
- No quality data, no quality mining results!
  - Quality decisions must be based on quality data
  - Data warehouse needs consistent integration of quality data

Multi-Dimensional Measure of Data Quality

- A well-accepted multidimensional view:
  - Accuracy: How well does a piece of information reflect reality?
  - Completeness Does it fulfill your expectations of what’s comprehensive?
  - Consistency Does information stored in one place match relevant data stored elsewhere?
  - Timeliness Is your information available when you need it?
  - Believability Howmuch data are trusted by user
  - Interpretability How easily the data are understood
  - Accessibility where data resides and how to retrieve it.
  - Value added Is the stored data adding value in the mining process

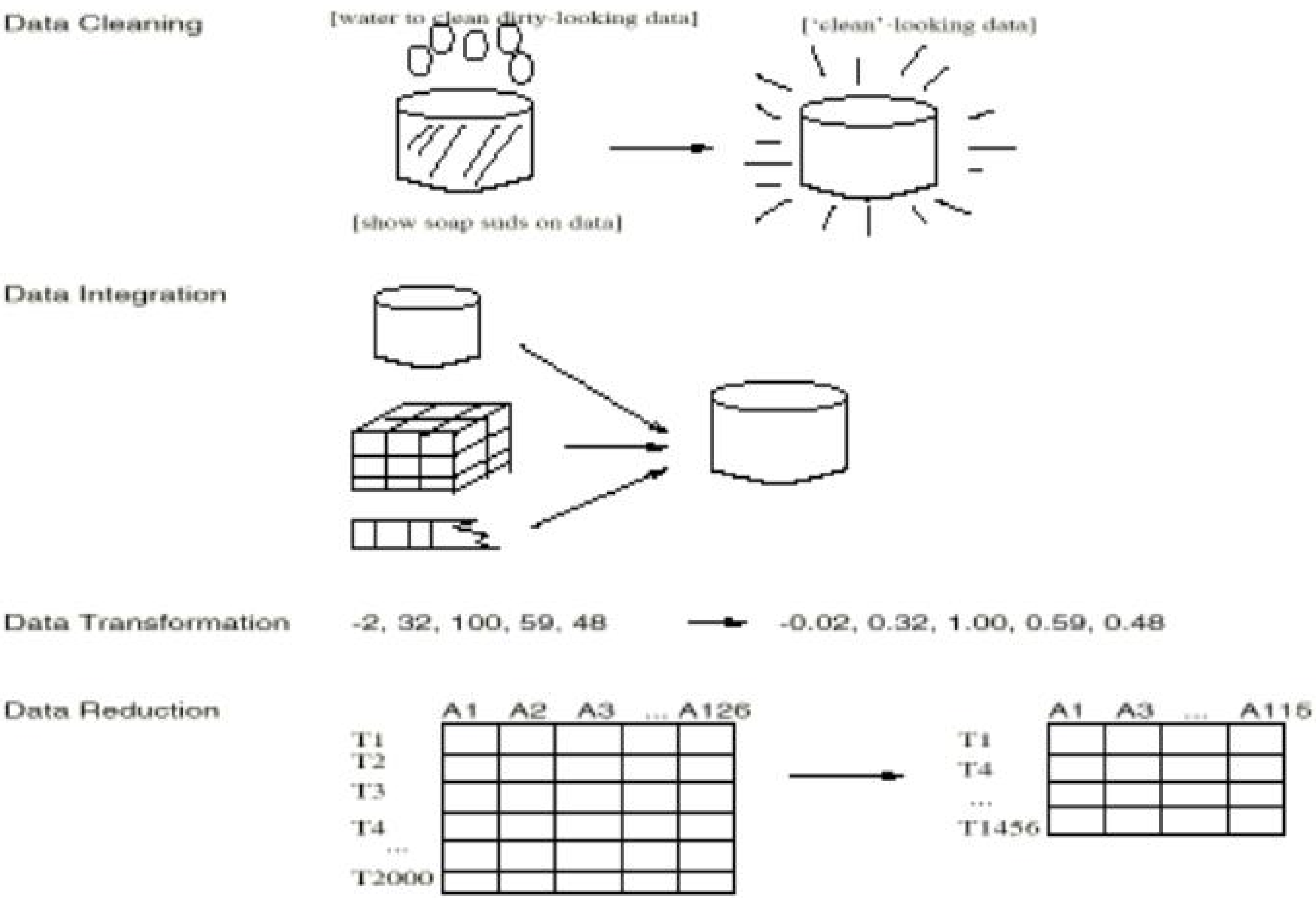
Broad categories:

- contextual, representational, and accessibility.

Major Tasks in Data Preprocessing

1. Data cleaning
   - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies (redundancy)
2. Data integration
   - Integration of multiple databases, data cubes, files, or notes
3. Data reduction
   - Obtains reduced representation in volume but produces the same or similar analytical results
   - Data aggregation, dimensionality reduction, data compression, generalization
4. Data transformation and Data discretization
   - Normalization (scaling to a specific range)
   - Aggregation
   - Data discretization: with particular importance, especially for numerical data

Forms of data preprocessing



Agenda

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

1. Data Cleaning

- Data Cleaning Tasks
  - ➢1.1- Fill in missing values
  - ➢1.2- Identify outliers and smooth out noisy data
  - ➢1.3- Correct inconsistent data

1.1 Missing Data

- **Data is not available/ Missing data may be due to**
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
  - Information is not collected :(e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- **Missing data need to be inferred**

How to Handle Missing Data?

- **Ignore the tuple**: usually done when class label is missing (assuming the task is classification—not effective in certain cases)
- Fill in the missing value **manually**: tedious + infeasible?
- Use a **global constant** to fill in the missing value: e.g., “unknown”, a new class?! simple but not foolproof.
- Use the **central tendency (mean/median)** to fill in the missing value. Normal->Mean, Skewed->Median
- Use the **most probable value** to fill in the missing value: inference-based such as regression, Bayesian formula, decision tree

Bias the data

Types of Missing Values

Why missing data is a problem?
Ans: It creates bias in the data. because we don’t know that the data is missing randomly/missedout/intensionally.
*Bias data: produce lack of prdictivity & trustworthyness

- **Missing completely at random (MCAR)**
- **Missing at Random (MAR)**
- **Missing Not at Random (MNAR)**

Strongest assumptions, easiest to model

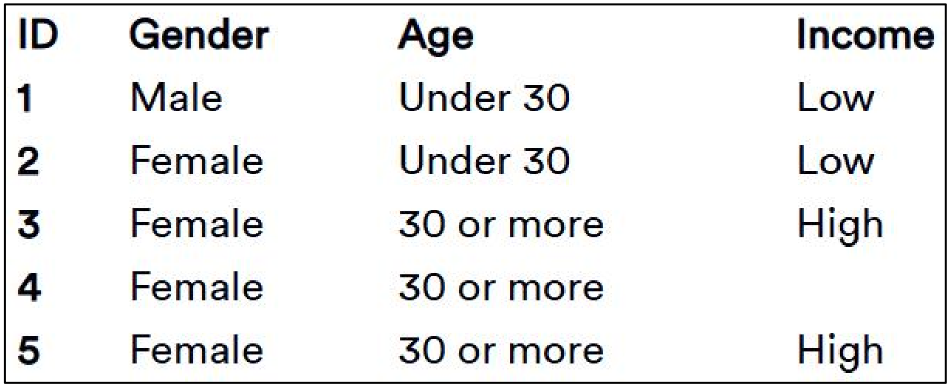
Weakest assumptions, hardest to model

Missing Completely at Random (MCAR) (Types of Missing Values…)

Assumption: If a person has missing data then it is completely unrelated to the other information in the data. The missingness on the variable is completely unsystematic.

- Missingness of a value is independent of attributes
- Fill in values based on the attribute
- Analysis may be unbiased overall

Example when we take a random sample of a population, where each member has the same chance of being included in the sample.



ID	Gender	Age	Income
1	Male	Under 30	Low
2	Female	Under 30	Low
3	Female	30 or more	High
4	Female	30 or more	
5	Female	30 or more	High

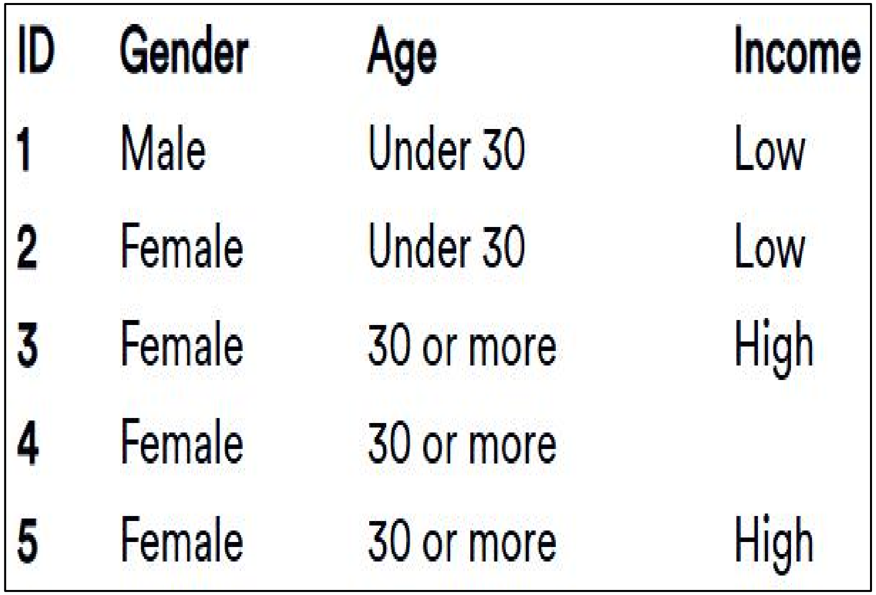
When data is missing completely at random, it means that we can undertake analyses using only observations that have complete data (provided we have enough of such observations).

Missing at Random (MAR) Types of Missing Values…

- Missingness is related to other variables
- Fill in values based other values
- Almost always produces a bias in the analysis

Example of MAR is when we take a sample from a population, where the probability to be included depends on some known property.

A simple predictive model is that income can be predicted based on gender and age. Looking at the table, we note that our missing value is for a Female aged 30 or more, and observations say the other females aged 30 or more have a High income. As a result, we can predict that the missing value should be High.



ID	Gender	Age	Income
1	Male	Under 30	Low
2	Female	Under 30	Low
3	Female	30 or more	High
4	Female	30 or more	
5	Female	30 or more	High

There is a systematic relationship between the inclination of missing values and the observed data. All that is required is a probabilistic relationship

Missing not at Random (MNAR) - Nonignorable

Types of Missing Values…

- Missingness is related to unobserved measurements and they are not random
- The missing values are related to the values of that variable itself, even after controlling for other variables.

MNAR means that the probability of being missing varies for reasons that are unknown to us.

Example: when smoking status is not recorded in patients admitted as an emergency with an intention (not random), then it is more likely to have worse outcomes from surgery.

Strategies to handle MNAR are to find more data about the causes for the missingness, or to perform what-if analyses to see how sensitive the results are under various scenarios.

1.2 Identify outliers and smooth out noisy data

- **Noise**

Random error or variance in a measured variable.
Or simply meaningless data that can’t be interpreted by machines.

- **Incorrect attribute values may be due to**
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in the naming convention

- **Other data problems which require data cleaning**
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

1.2.1 Binning method for data Smoothing:

- first sort data and partition it into (equi-depth) bins
- then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- used also for discretization (discussed later)

1.2.2 Clustering

- detect and remove outliers

1.2.4 Regression

- smooth by fitting the data into regression functions

- Data Cleaning: A Process

1.2.1.Simple Discretization Method: Binning

- Data smoothing refers to a statistical approach of eliminating noise and outliers from datasets to make the patterns more noticeable
- Binning or bucketing is used to smoothing the data. It smooth a sorted data value by consulting its “neighborhood” i.e. the values around it.
- The sorted values are distributed into a number of “buckets,” or bins of equal width or equal frequency
- Bin Size : No. bins or buckets = square root of the no of data points
- Bin Width/Depth: No of objects/elements in a single bin.
- There are 2 methods of dividing data into bins
  - Equal Width Binning -Equal Frequency Binning

1.2.1 Binning Methods

For data set: 0, 5, 14, 15, 17, 18, 22, 25, 27 (sorted)

No. of Bins/ Bin size = 3 [3*3=9]

Equal Width Binning

- **Bins have equal width with a range of each bin are defined as**

[min + w], [min + 2w] …. [min + nw]

where w = (max – min) / (no of bins)

= (27-0)/3=9

- **How do I use that 9 to make the bins?**

1. 0 + 9 = 9 (from 0 to 9)
= Bin 1: 0, 5

2. 9 + 9 = 18 (from 9+ to 18)
= Bin 2 : 14, 15, 17, 18

3. 18 + 9 = 27 (from 18+ to 27)
= Bin 3 : 22, 25, 27

Equal Frequency Binning

Make bins according to bin size with **equal frequency/depth i.e. equal elements i.e. 9/3=3**

- Bin size=3
  - Bin 1: 0, 5, 14
  - Bin 2: 15, 17, 18
  - Bin 3: 22, 25, 27

Types of Smoothing in Equal Frequency Bins and Equal Width Bins

- Smoothing by Mean
- Smoothing by Median
- Smoothing by Boundaries

Smoothing the data by **Equal Frequency Bins**

Step-1 Sort the data in **ascending order:**

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

Step-2: Number of Bins = $\sqrt{No.\,of\ elements} = \sqrt{12} = 3.4 \approx 3$

Step-3 Partition into equal frequency (equi-depth) of 3 bins:

- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

Smoothing the data by Equal Frequency Bins contd..

1. Smoothing by <u>BIN MEANS</u>: Find the mean values of each bin and Replace all with mean values

- Bin 1: 9, 9, 9, 9
- Bin 2: 23, 23, 23, 23
- Bin 3: 29, 29, 29, 29

2. Smoothing by <u>BIN MEDIANS</u>: Find the median values of each bin and Replace all with the median

- Bin 1: 8.5, 8.5, 8.5, 8.5
- Bin 2: 22.5, 22.5, 22.5, 22.5
- Bin 3: 28.5, 28.5, 28.5, 28.5

3. Smoothing by <u>BIN BOUNDARIES</u>: Min and Max will be the Bin boundary, and middle element will be replaced by the closet boundary value

- Bin 1: 4, 4, 4, 15
- Bin 2: 21, 21, 25, 25
- Bin 3: 26, 26, 26, 34

Smoothing the data by **Equal Width Bins**

Step-1 Sort the data in **ascending order:**

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

Step-2: Number of Bins = $\sqrt{No. of\ elements} = \sqrt{12} = 3.4 \approx 3$

Step-3: **Partition into equal width (equi-width) bins:**

[min + w], [w + 2w] …. [(n-1)w + nw]

where w = (max – min) / (no of bins)

= (34-4)/3=10

Partition into equal width of 3 bins:

- Bin 1: 4, 8, 9 (4 to14)
- Bin 2: 15, 21, 21, 24 (14+ to 24)
- Bin 3: 25, 26, 28, 29, 34 (24+ to 34)

Smoothing the data by **Equal Width Bins** contd..

1. Smoothing by <u>BIN MEANS</u>: Find the mean values of each bin and Replace all with mean values

- Bin 1: 7, 7, 7
- Bin 2: 20.25, 20.25, 20.25, 20.25
- Bin 3: 28.4, 28.4, 28.4, 28.4

2. Smoothing by <u>BIN MEDIANS</u>: Find the median values of each bin and Replace all with the median

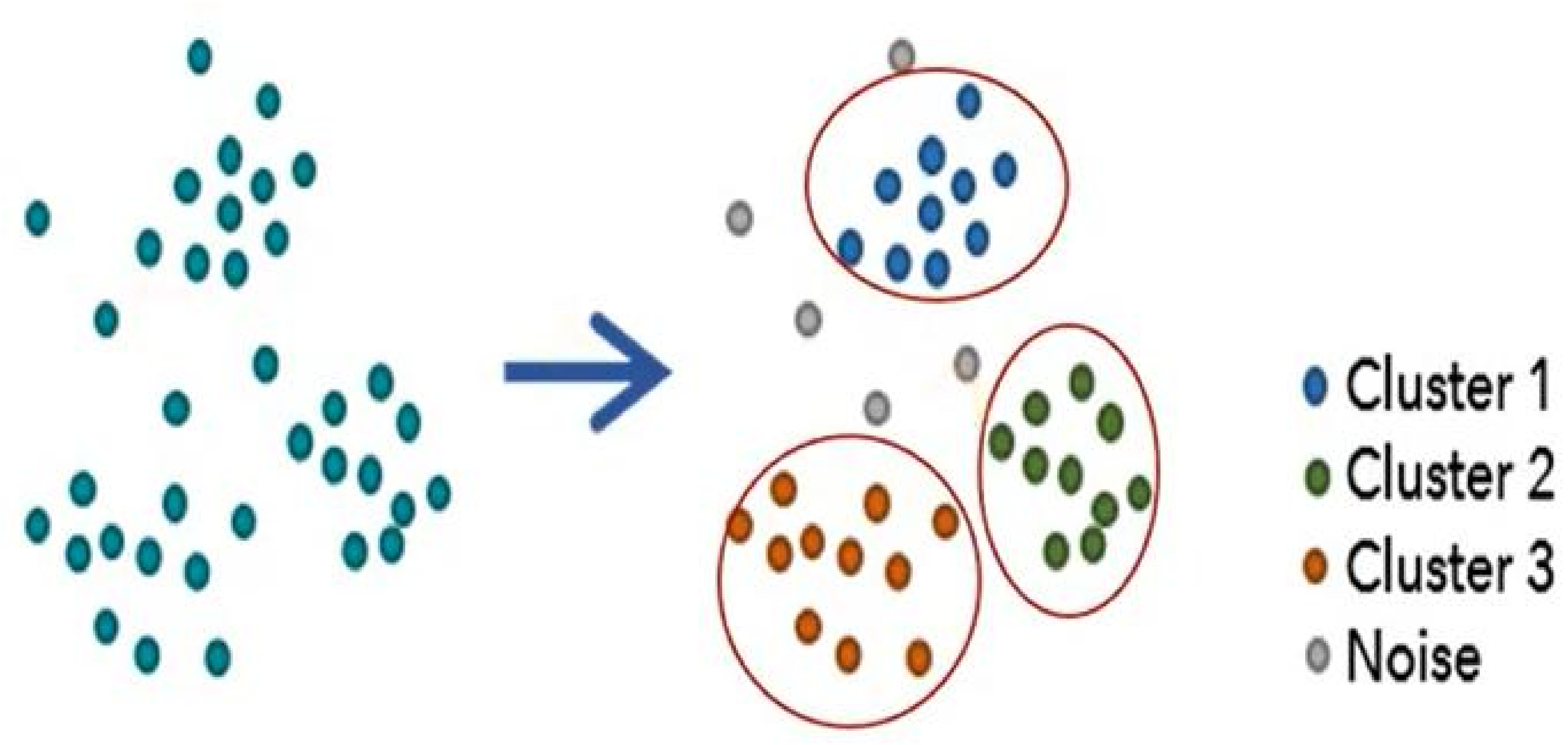
- Bin 1: 8, 8, 8, 8
- Bin 2: 21, 21, 21, 21
- Bin 3: 28, 28, 28, 28

3. Smoothing by <u>BIN BOUNDARIES</u>: Min and Max will be the Bin boundary, and middle element will be replaced by the closet boundary value

- Bin 1: 4, 9, 9
- Bin 2: 15, 24, 24, 24 (14+ to 24)
- Bin 3: 25, 25, 25, 25, 34 (24+ to 34)

1.2.2 Cluster Analysis

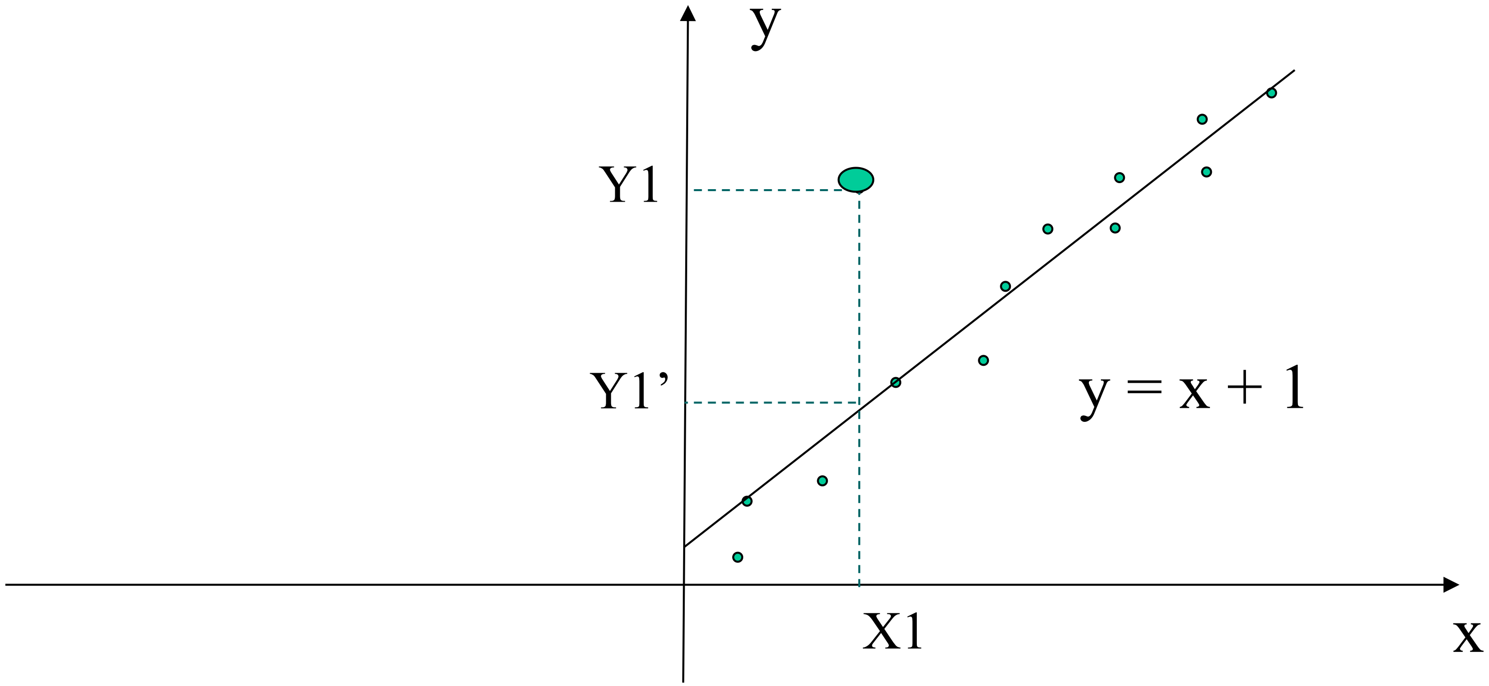
Clustering is the task of dividing the population or data points into a number of groups (without prior knowledge of class labels) such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.



1.2.3 Regression

Regression is a method to determine the relationship between a dependent variable and one or more independent variables. It smoothes by fitting the data into regression functions

- Linear regression (best line to fit two variables)
- Multiple linear regression (more than two variables, fit to a multidimensional surface



1.3 How to Handle Inconsistent Data?

- Manual correction using external references
- Semi-automatic using various tools
  - To detect violations of known functional dependencies and data constraints
  - To correct redundant data

Data Cleaning: A Process

It says how one tackles **missing values, noise, and inconsistency**

1. **Discrepancy Detection**

   Causes: Poorly designed data entry forms

   Human error in data entry

   Deliberate errors and Data Decay (outdated address)

   Error in instruments

   Solution: Gather Knowledge i.e. metadata [Mean, Median, Mode, domain of attribute, data symmetric or skewed, values concerning standard deviation and mean] and find the discrepancy.

2. As a data analyst **Look out for the inconsistent use of codes**

   Date format (2010/02/18) and (18/02/2010)

3. **Examine data regarding**

- Unique rule: Each value of one attribute should be unique
- Consecutive rule: Lowest to highest, all the attribute's values should be continuous.
- Null rule: How the null values will be handled [NULL, NA, $, unknown]

Tools for Discrepancy Detection

- **Data scrubbing tools** (Uses simple domain knowledge)
  Ex. Knowledge of Postal addresses, spell checking to detect errors and make corrections
- **Data auditing tools** (employ statistical analysis to find correlations)
  Ex. To find discrepancies by analyzing the relationships and rules.
- **Data migration tools** (Allows simple transformations)
  Ex. Replace string “Gender” with “Sex”

Agenda

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

2. Data Integration and Transformation

Data Integration

– It is a preprocessing method that combines data from multiple sources into a coherent store i.e. data warehouse

- Issues to be Considered

a. Schema integration and object matching (Entity Identification Problem)

Ex. customer_id in one database and cust_number in another

Solution: Metadata can be helpful to avoid such errors in schema

b. Redundancy (unwanted attribute)

Ex. Age and DoB in the same schema,

Solution: Remove the unnecessary attribute if it can be derived from another.

c. Detecting and resolving data value conflicts

For the same real-world entity, attribute values from different sources are different.

Possible reasons (Ex.): different representations, different scales, e.g., metric vs. British units, different currency

Solution: Correctly modify the values

2. Handling Redundant Data in Data Integration

- Redundant data occur often when integrating multiple DBs
  - The same attribute may have different names in different databases or one attribute may be a "derived" attribute in another table, e.g., annual revenue, age
- Redundant data may be detected by **correlational analysis**
  - Given two attributes, correlation analysis can measure how strongly one attribute implies another, based on available data.
    - **Correlation coefficient for numerical data values:** Can be computed by Pearson's product moment coefficient named after Karl Pearson

$$\mathbf{r_{A,B} = \frac{\Sigma(A - \overline{A})(B - \overline{B})}{(n - 1)\sigma_A \sigma_B} = \frac{Cov(A, B)}{\sigma_A \sigma_B}}$$

    - **Chi-square test for categorical or discrete data values:** A chi-square test is a statistical test used to compare observed results with expected results

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \text{Where} \quad E_{ij} = \sum \frac{\text{Row Total} \times \text{Col Total}}{\text{Total population}}$$

2. Handling Redundant Data in Data Integration contd..

Correlation Analysis (Nominal Data)

χ^2 **(chi-square) test**

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o_{ij}-e_{ij})^2}{e_{ij}}, \qquad \text{Eq.(1)}$$

where o_{ij} is the observed frequency (i.e., actual count) of the joint event (A_i,B_j) and e_{ij} is the expected frequency of (A_i,B_j), which can be computed as

$$e_{ij} = \frac{count(A=a_i)\times count(B=b_j)}{n} \qquad \text{Eq.(2)}$$

The larger the χ^2 value, the more likely the variables are correlated.

2. Handling Redundant Data in Data Integration contd..

Chi-Square Calculation: An Example

- Suppose that a group of 1500 people was surveyed. The gender of each person was noted. Each person was polled as to whether his or her preferred type of reading material was fiction or nonfiction.
- Thus, we have two attributes, gender and preferred reading. The observed frequency (or count) of each possible joint event is summarized in the contingency table.
- The expected frequencies are calculated based on the data distribution for both attributes using Eq. (2).
- The expected frequency for the cell (male, fiction) is

$$e_{11} = \frac{count(male) \times count(fiction)}{n} = \frac{300 \times 450}{1500} = 90$$

- Similarly, calculate the expected frequency for each cell of the contingency table.

2. Handling Redundant Data in Data Integration contd..

Chi-Square Calculation: An Example

	Male Observed	Male Expected	Female Observed	Male Expected	Sum (row)
Like science fiction	250	90	200	360	450
Not like science fiction	50	210	1000	840	1050
Sum(col.)	300		1200		1500

- χ^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- For this 2 × 2 table, the degrees of freedom are $(2-1)(2-1) = 1$. For 1 degree of freedom, the χ^2 value needed to reject the hypothesis at the 0.001 significance level is 10.828 (taken from the table of upper percentage points of the χ^2 distribution, typically available from any textbook on statistics).
- As computed value (507.93) is more than 10.828. Hence, preferred reading and gender are correlated.

2. Handling Redundant Data in Data Integration contd..

Correlation Analysis (Numeric Data)

- For numeric attributes, we can evaluate the correlation between two attributes, A and B, by computing the correlation coefficient (also known as Pearson’s product moment coefficient, named after its inventer, Karl Pearson).

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B},$$

- where n is the number of tuples, ai $\bar{A}$ and bi $\bar{B}$ are the respective values of A and B in tuple i, and are the respective mean values of A and B, σ_A and σ_B are the respective standard deviations of A and B.

2. Handling Redundant Data in Data Integration contd..

Correlation Analysis (Numeric Data) contd...

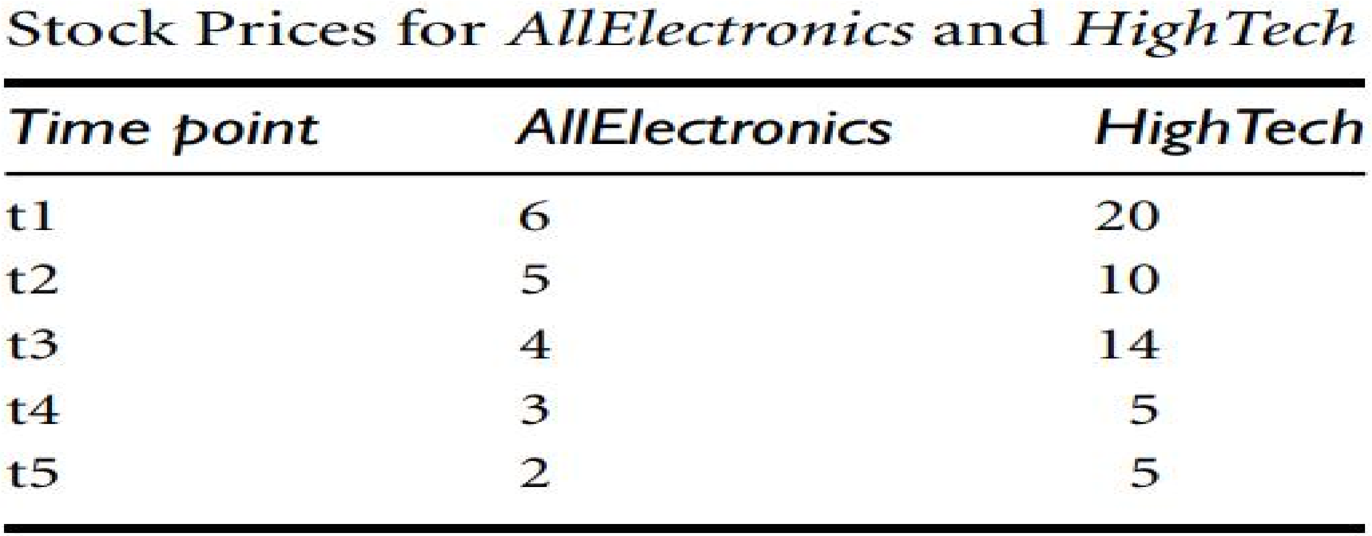
- Note that $-1 \leq r_{A,B} \leq +1$. If $r_{A,B}$ is greater than 0, then A and B are positively correlated, meaning that the values of A increase as the values of B increase.
- The higher the value, the stronger the correlation (i.e., the more each attribute implies the other). Hence, a higher value may indicate that A (or B) may be removed as a redundancy.
- If the resulting value is equal to 0, then A and B are independent and there is no correlation between them.
- If the resulting value is less than 0, then A and B are negatively correlated, where the values of one attribute increase as the values of the other attribute decrease.
- Scatter plots can also be used to view correlations between attributes.

2. Handling Redundant Data in Data Integration contd..

Correlation Analysis (Numeric Data)

Example:

Consider the stock prices listed in the below table



Stock Prices for *AllElectronics* and *HighTech*

Time point	*AllElectronics*	*HighTech*
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

Find out wheather AllElectronics and HighTeck prices are correlated or not?

2. Handling Redundant Data in Data Integration contd..

Correlation Analysis (Numeric Data)

Step-1: Calculate the mean of the attribute AllElectronics (A) and HighTech (B).

Mean(A)=(6+5+4+3+2)/5=4

Mean(B)=(20+10+14+5+5)/5=10.8

Step-2: calculate the standard deviation of both attributes A and B i.e. 1.414 and 5.706 respectively

Step-3: calculate the correlation value using the given equation

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_ib_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B},$$

Step-4: calculated value is greater than 0, hence both are correlated.

Correlation coefficient

A measure of the strength of linear association between two variables. Correlation will always between -1.0 and +1.0.

Correlation can be:

a. Positive Correlation

b. Negative Relation

c. Null Correlation

a. Positive Correlation:

The correlation in the same direction is called positive correlation. If one variable increase other is also increase and one variable decrease other is also decrease. For example, the length of an iron bar will increase as the temperature increases.

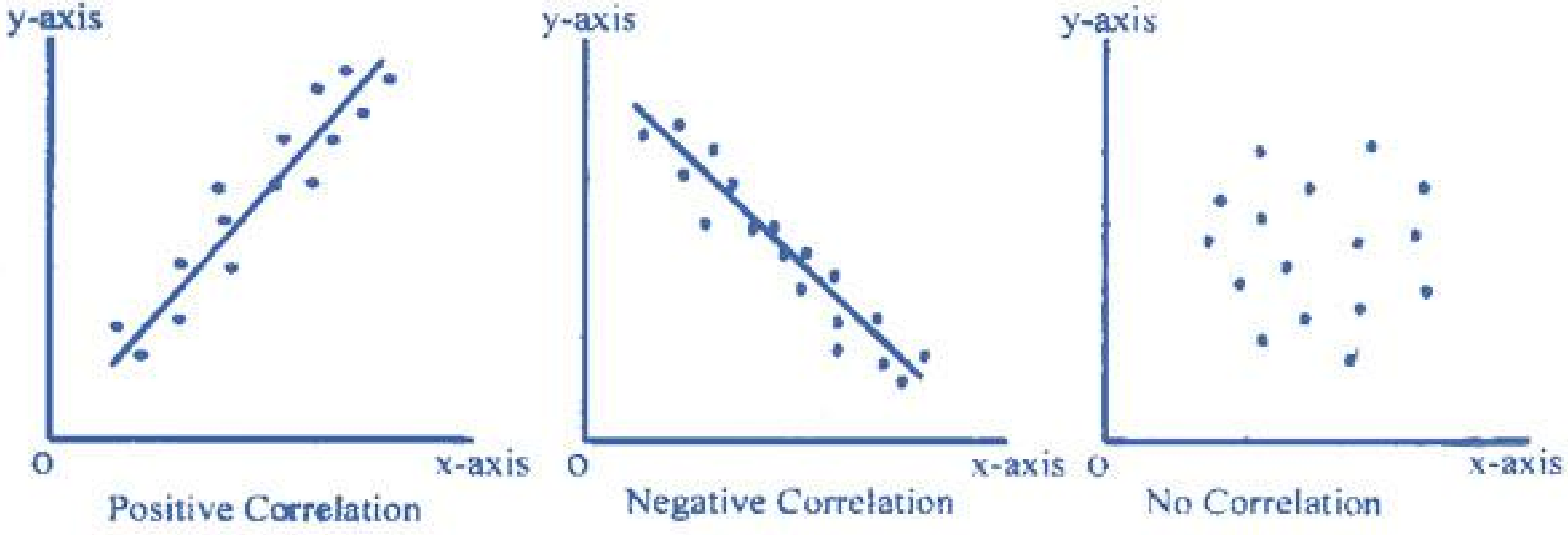
b. Negative Correlation:

The correlation in opposite direction is called negative correlation, if with the increase in one variable the other decreases and vice versa. Example, the volume of gas will decrease as the pressure increase or the demand of a particular commodity is increase as price of such commodity is decrease.

2. Handling Redundant Data in Data Integration contd..

Correlation coefficient contd..

c. No Correlation or Zero Correlation:
If there is no relationship between the two variables such that the value of one variable change and the other variable remain constant is called no or zero correlation.



Positive Correlation

Negative Correlation

No Correlation

2. Data Integration and Transformation

Data Transformation

It maps the values of one attribute to another new set of replacement values so that each old value can be identified with one of the new values

Strategies for Data Normalization are:

- **Smoothing:** remove noise from data (binning, clustering, regression)
- **Aggregation:** summarization process is applied, data cube construction
- **Generalization/Concept hierarchy climbing:** Attributes can be generalized to higher-level concept
- **Normalization:** scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- **Attribute/feature construction:** New attributes are constructed from the given ones to help the mining process
- **Discretization:** Raw values will be replaced by interval labels/ conceptual labels

2. Data Transformation: Data Aggregation

- Data aggregation is any process in which data is brought together and conveyed in a summary form. It is typically used prior to the performance of a statistical analysis.
- **Combining two or more attributes (or objects) into a single attribute (or object).**
- Data aggregation generally works on the big data or data marts that do not provide enough information value as a whole.

Aggregation with mathematical functions:

- **Sum** -Adds together all the specified data to get a total.
- **Average** -Computes the average value of the specific data.
- **Max** -Displays the highest value for each category.
- **Min** -Displays the lowest value for each category.
- **Count** -Counts the total number of data entries for each category.

Data can also be aggregated by date, allowing trends to be shown over a period of years, quarters, months, etc.

2. Data Transformation: Data Normalization

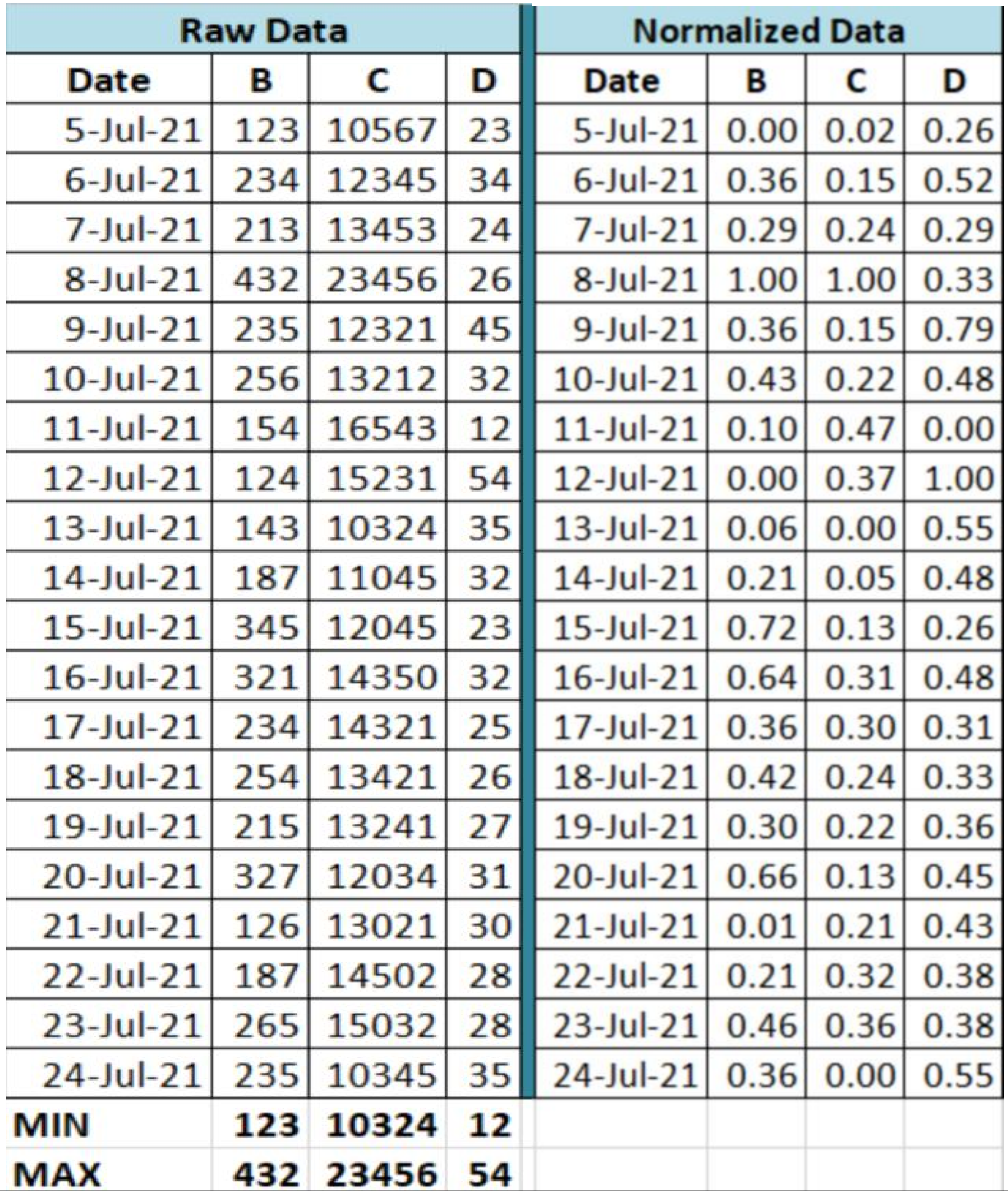
Data normalization makes data easier to classify and understand. It is used to scale the data of an attribute so that it falls in a smaller range

Need of Normalization?

- Normalization is generally required when multiple attributes are there but attributes have values on different scales, this may lead to poor data models while performing data mining operations.
- Otherwise, it may lead to a dilution in effectiveness of an important equally important attribute(on lower scale) because of other attribute having values on larger scale.
- Heterogenous data with different units usually needs to be normalized. Otherwise, data has the same unit and same order of magnitude it might not be necessary with normalization.
- Unless normalized at pre-processing, variables with disparate ranges or varying precision acquire different driving values.

2. Data Transformation: Data Normalization contd..

Example



Raw Data				Normalized Data			
Date	B	C	D	Date	B	C	D
5-Jul-21	123	10567	23	5-Jul-21	0.00	0.02	0.26
6-Jul-21	234	12345	34	6-Jul-21	0.36	0.15	0.52
7-Jul-21	213	13453	24	7-Jul-21	0.29	0.24	0.29
8-Jul-21	432	23456	26	8-Jul-21	1.00	1.00	0.33
9-Jul-21	235	12321	45	9-Jul-21	0.36	0.15	0.79
10-Jul-21	256	13212	32	10-Jul-21	0.43	0.22	0.48
11-Jul-21	154	16543	12	11-Jul-21	0.10	0.47	0.00
12-Jul-21	124	15231	54	12-Jul-21	0.00	0.37	1.00
13-Jul-21	143	10324	35	13-Jul-21	0.06	0.00	0.55
14-Jul-21	187	11045	32	14-Jul-21	0.21	0.05	0.48
15-Jul-21	345	12045	23	15-Jul-21	0.72	0.13	0.26
16-Jul-21	321	14350	32	16-Jul-21	0.64	0.31	0.48
17-Jul-21	234	14321	25	17-Jul-21	0.36	0.30	0.31
18-Jul-21	254	13421	26	18-Jul-21	0.42	0.24	0.33
19-Jul-21	215	13241	27	19-Jul-21	0.30	0.22	0.36
20-Jul-21	327	12034	31	20-Jul-21	0.66	0.13	0.45
21-Jul-21	126	13021	30	21-Jul-21	0.01	0.21	0.43
22-Jul-21	187	14502	28	22-Jul-21	0.21	0.32	0.38
23-Jul-21	265	15032	28	23-Jul-21	0.46	0.36	0.38
24-Jul-21	235	10345	35	24-Jul-21	0.36	0.00	0.55
MIN	123	10324	12				
MAX	432	23456	54				

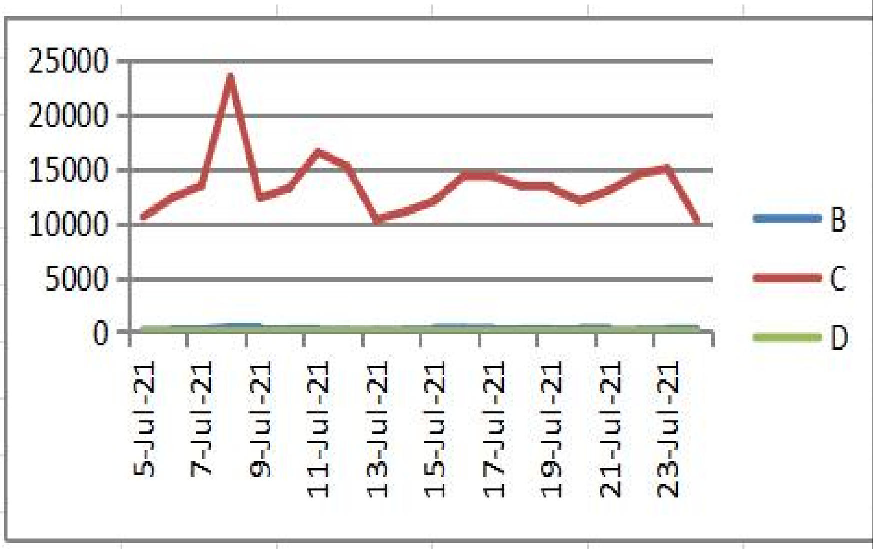


Chart for Raw Data

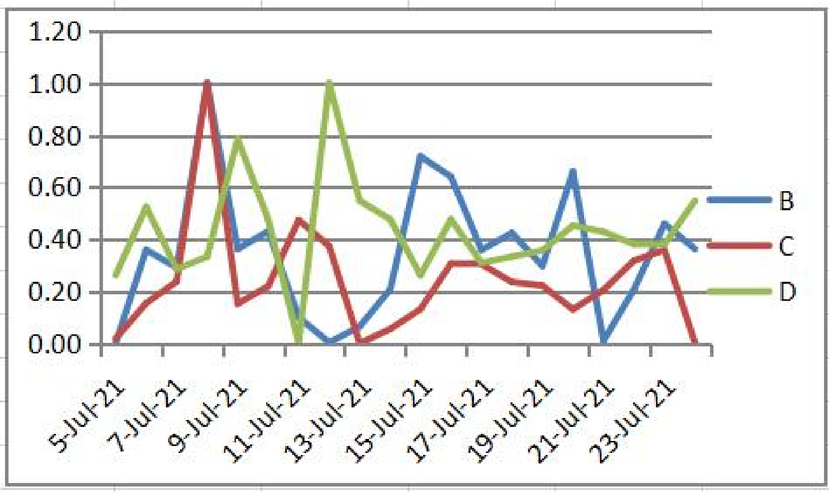


Chart for Normalized Data

2. Data Transformation: Data Normalization contd..

Methods of Data Normalization:

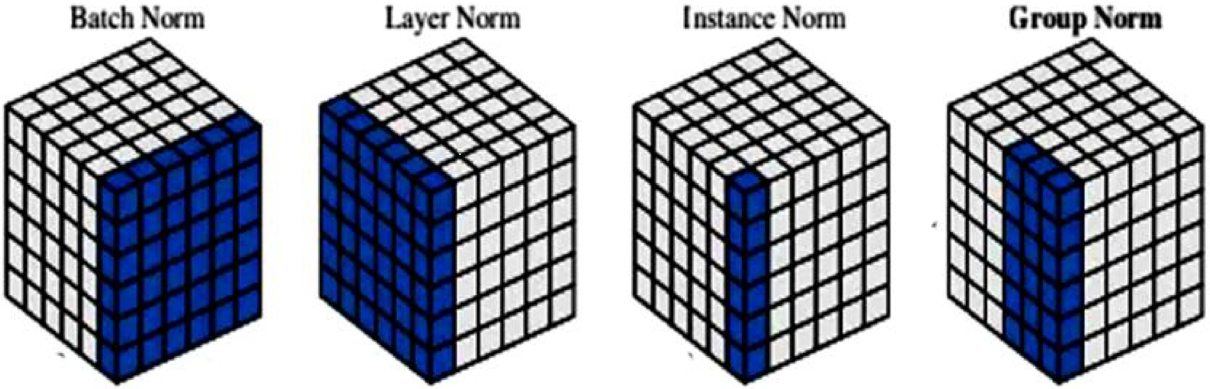
a. **Decimal Scaling**

b. **Min-Max Normalization**

c. **z-Score Normalization(zero-mean Normalization)**

There are several approaches in normalisation which can be used in deep learning models.

- **Batch Normalization**
- **Layer Normalization**
- **Group Normalization**
- **Instance Normalization**
- **Weight Normalization**



2. Data Transformation: Data Normalization contd..

a. Decimal Scaling Normalization

- It normalizes by moving the decimal point of values of the data.
- To normalize the data by this technique, we divide each data value by the maximum absolute value of data set.
- The data value, $\boldsymbol{v_i}$, of data is normalized to $\boldsymbol{v'_i}$ by using the formula

$$\boxed{v_i' = \frac{v_i}{10^j}}$$

[where j is the smallest integer such that max(|$\boldsymbol{v'_i}$|)<1.]

In this technique, the computation is generally scaled in terms of decimals. It means that the result is generally scaled by multiplying or dividing it with pow(10,k).

Example:

- Normalize the input data is: - 15, 121, 201, 421, 561, 601, 850
- **Step 1:** Maximum value in given data(m): **850** and hence maximum absolute value is 1**000**
- **Step 2:** Divide the given data by 1000 (i.e j=3)
- Result: The normalized data is: - 0.015, 0.121, 0.201, 0.421, 0.561, 0.601, 0.85

2. Data Transformation: Data Normalization contd..

b. Min-Max Normalization (Linear Transformation)

- Minimum and maximum value from data is fetched and each value is replaced according to the following formula.

$$v' = \frac{v - \min(A)}{\max(A) - \min(A)}(\text{new\_max}(A) - \text{new\_min}(A)) + \text{new\_min}(A)$$

Where - A is the attribute data(col)

- ***v** and **v'*** is the old and new value of each entry in data
- ***min***(A), m***ax***(A) are the minimum and maximum of A
- new_max(A), new_min(A) is the max and min value of the required range(i.e boundary value) respectively.

Example

Input:- **10, 15, 50, 60**

Normalized to range 0 to 1.

Here min=10, max= 60, new_min=0, new_max=1

Output:- **0, 0.1, 0.8, 1**

2. Data Transformation: Data Normalization contd..

c. z-Score Normalization (zero-mean Normalization)

- Values are normalized based on mean and standard deviation of the data A.
- It is also called **Standard Deviation method.**
- Unstructured data can be normalized using z-score parameter,

$$\boxed{v' = \frac{v - \bar{x}}{s}}$$

where - $\bar{x}$: mean

- S is the standard deviation.
- ***v** and **v'*** is the old and new value of each data

Input:- **10, 15, 50, 60**

$$mean = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 33.75$$

$$SD = \bar{x} = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n-1}}$$

Output:- **0.9515, 0.7512, 0.6510, 1.0517**

Agenda

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

3. Data Reduction

- Problem:

Data Warehouse may store terabytes of data: Complex data analysis/mining may take a very long time to run on the complete data set

- Solution?
  - Data reduction…

3. Data Reduction contd...

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

- Integrity of the original data should be maintained even a after reduction in data volume.
- It should produce the same analytic result as on original data

Data reduction strategies

A. Data cube aggregation (applied to data cube)

B. Attribute Subset Selection(irrelevant attributes detected & removed)

C. Dimensionality reduction

D. Data compression

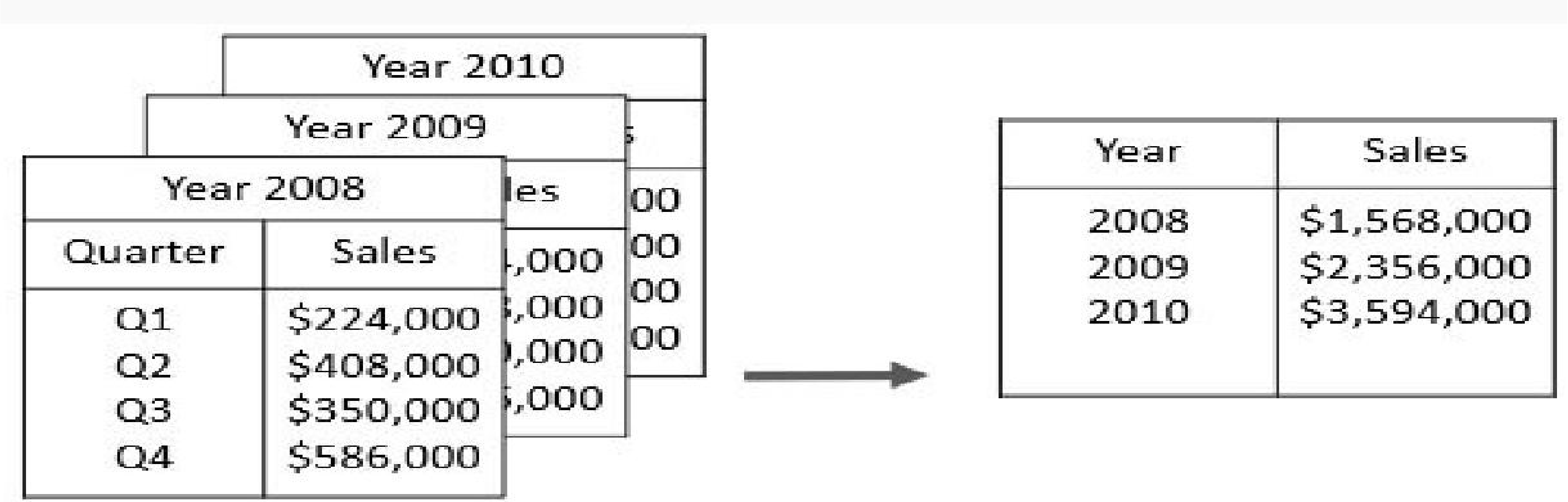
E. Numerosity reduction

F. Discretization and concept hierarchy generation

3. Data Reduction contd...

A. Data Cube Aggregation

It is a process in which information is gathered and expressed in a summary form



Year 2010

Year 2009

Year 2008

Quarter	Sales
Q1	$224,000
Q2	$408,000
Q3	$350,000
Q4	$586,000

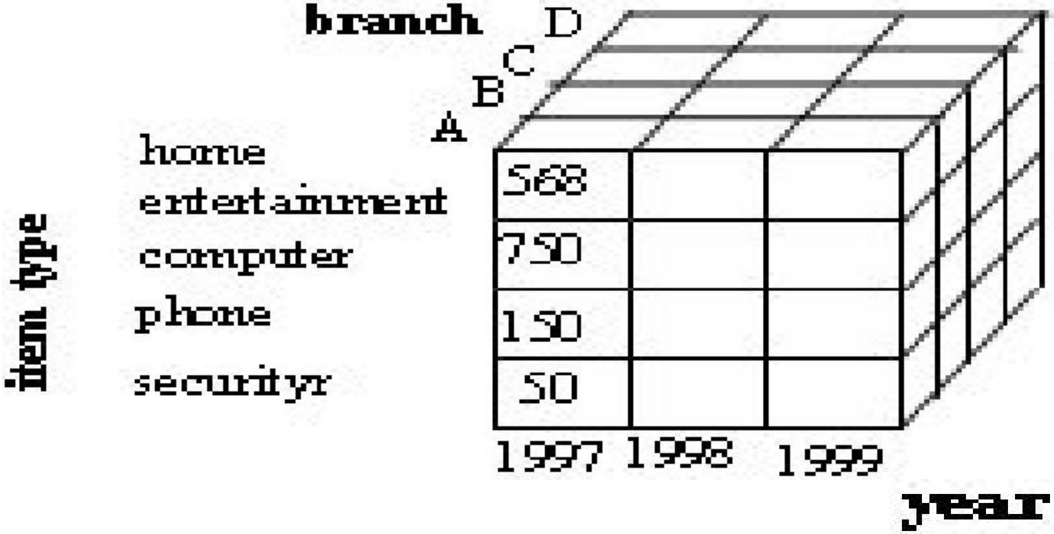
Year	Sales
2008	$1,568,000
2009	$2,356,000
2010	$3,594,000

For example, above is the data of one company's sales per quarter for the year 2018 to the year 2022. If the problem is to get the annual sale per year, then it is required to aggregate the sales per quarter for each year. In this way, aggregation provides you with the required data, which is much smaller in size, and thereby we achieve data reduction even without losing any data.

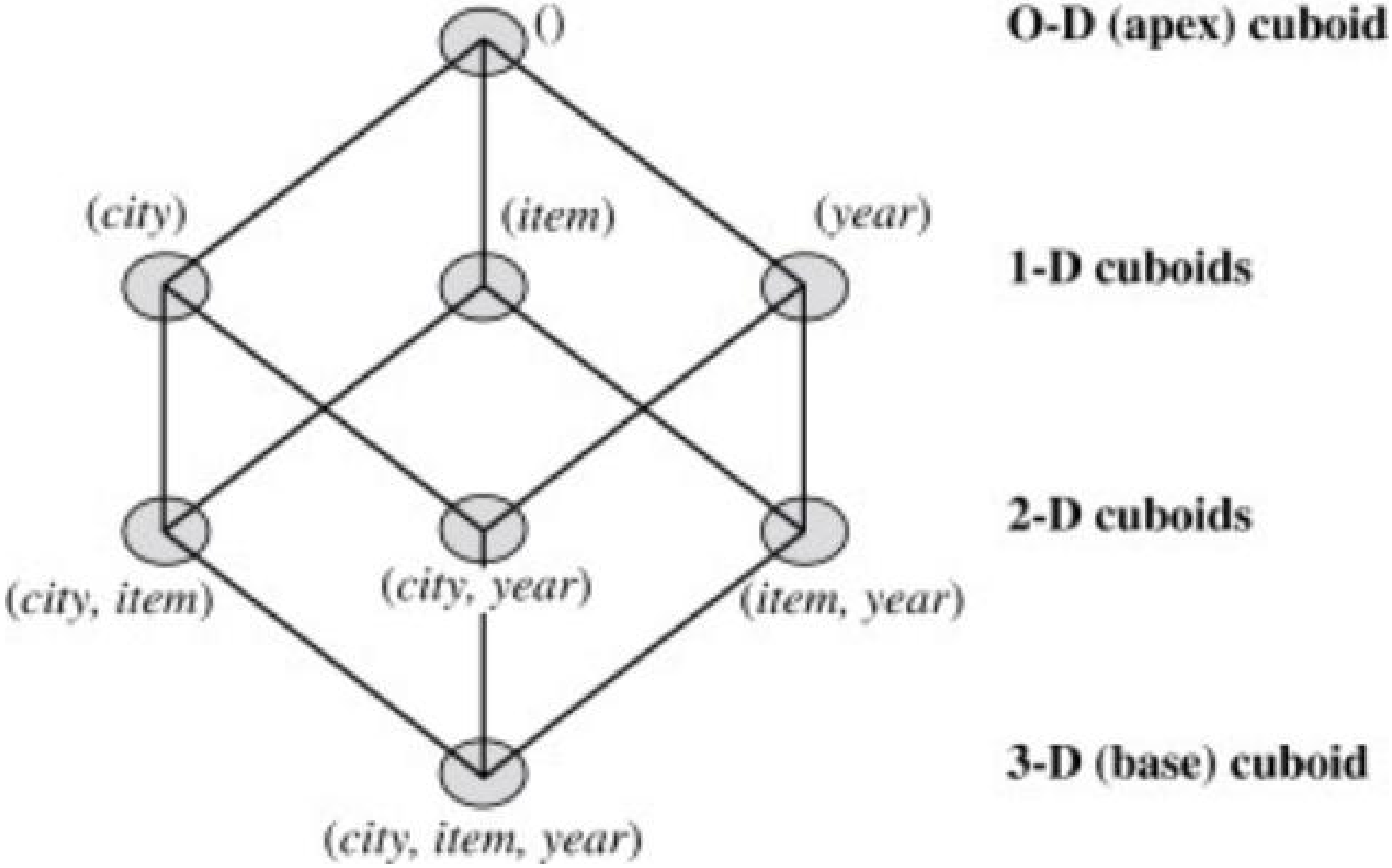
A. Data Cube Aggregation contd..

- Data cubes store multidimensional aggregated information.
- Data cubes provide fast access to precomputed, summarized data, thereby benefiting online analytical processing as well as data mining.
- **Base cuboid:** – The cube created at the lowest level of abstraction is referred to as the base cuboid. Example: The base cuboid should correspond to an individual entity of interest, such as sales and customers.
- **Apex cuboid:** A cube at the highest level of abstraction is the apex

cuboid. Example: For the sales data, the apex cuboid would give one total:- the total sales



Base Cuboid Vs Apex Cuboid



3. Data Reduction contd...

B. Attribute subset Selection

- It is the way to reduce the dimensionality of data through the use of **Feature selection.**
- It aims to discover a minimum set of attributes such that the resulting probability distribution of the data classes is as close as applicable to the original distribution using all attributes.
- Attribute subset selection decreases the data set size by eliminating irrelevant or redundant attributes (or dimensions).
- **Redundant attributes**
  - Duplicate much or all of the information contained in one or more other attributes
  - E.g., purchase price of a product and the amount of sales tax paid
- **Irrelevant attributes**
  - Contain no information that is useful for the data mining task at hand
  - E.g. students' ID is irrelevant to the task of predicting students' CGPA

B. Attribute subset Selection contd..

How can we find a 'good' subset of the original attributes?

- For n attributes, there are 2^n possible subsets.
- An exhaustive search for the optimal subset of attributes can be expensive, especially as n increase (Brute force method).
- Thus heuristic methods (Greedy method) that explore a reduced search space are commonly used for attribute subset selection.

Heuristic methods:

i. Step-wise forward selection

ii. Step-wise backward elimination

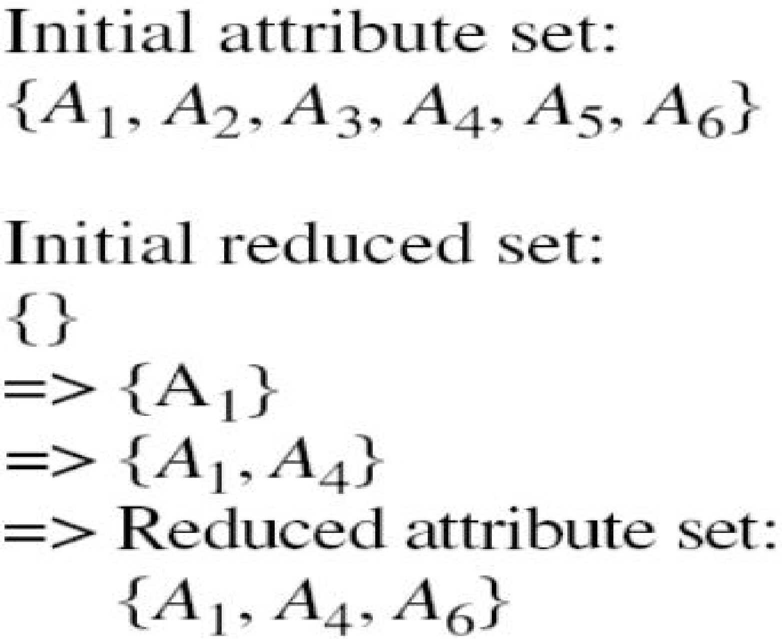
iii. Combining forward selection and backward elimination

iv. Decision-tree induction

B. Attribute subset Selection contd..

i. Stepwise Forward Selection:

- The procedure starts with an empty set of attributes as the reduced set.
- First: The best single-feature is picked.
- Next: At each subsequent iteration or step, the best of the remaining original attributes is added to the set.



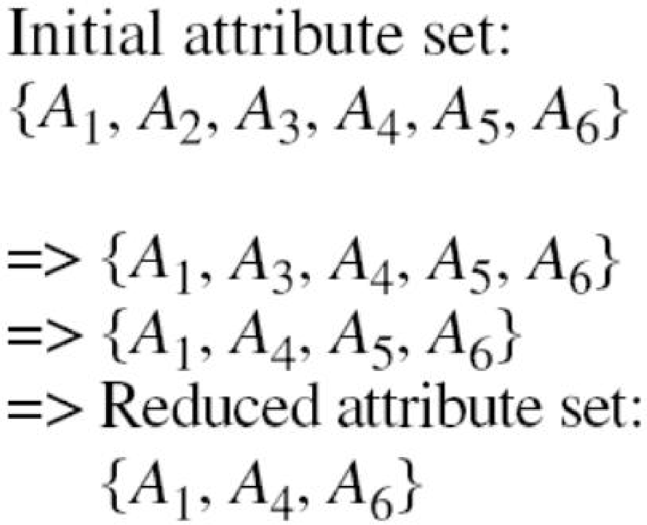
Initial attribute set:
$\{A_1, A_2, A_3, A_4, A_5, A_6\}$

Initial reduced set:
$\{\}$
=> $\{A_1\}$
=> $\{A_1, A_4\}$
=> Reduced attribute set:
$\{A_1, A_4, A_6\}$

B. Attribute subset Selection contd..

ii. Stepwise Backword Selection:

- The procedure starts with the full set of attributes.
- At each step, it removes the worst attribute remaining in the set.



Initial attribute set:
$\{A_1, A_2, A_3, A_4, A_5, A_6\}$

=> $\{A_1, A_3, A_4, A_5, A_6\}$
=> $\{A_1, A_4, A_5, A_6\}$
=> Reduced attribute set:
$\{A_1, A_4, A_6\}$

iii. Combining forward selection and backward elimination

- The stepwise forward selection and backward elimination methods can be combined
- At each step, the procedure selects the best attribute and removes the worst from among the remaining attributes

B. Attribute subset Selection contd..

iv. Decision Tree Induction

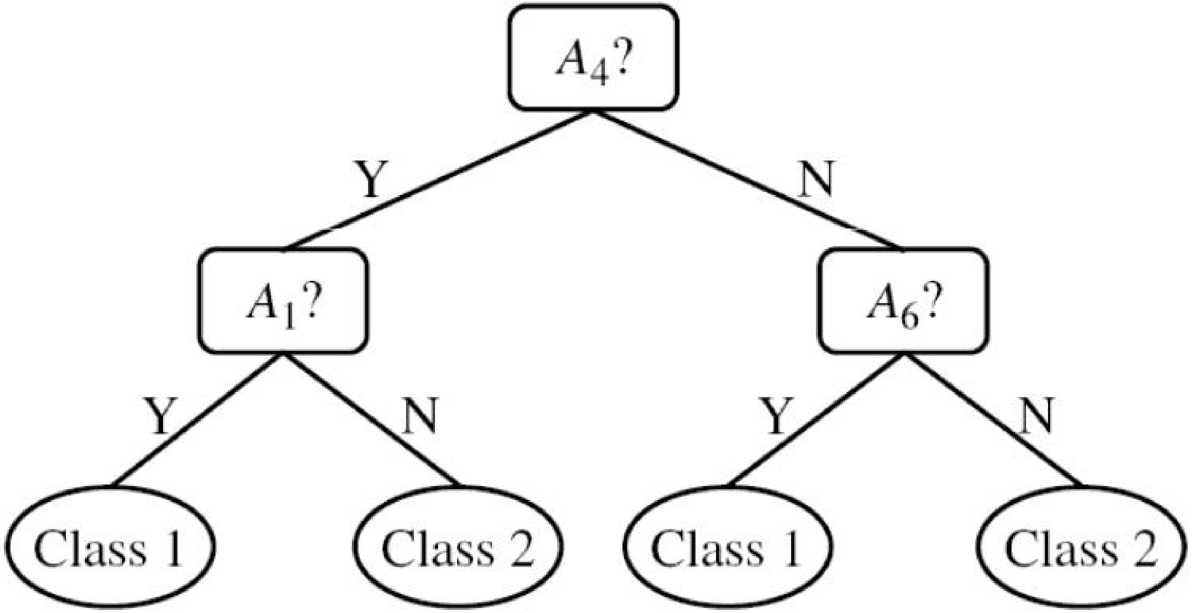
- Decision tree induction (Classification Algorithm) constructs a flowchart-like or tree like structure from given data where each internal node denotes a test on an attribute, each branch corresponds to an outcome of the test and each external node denotes a class prediction.
- At each node, the algorithm chooses the "best" attribute to partition the data into individual classes.
- All attributes that do not appear in the tree are assumed to be irrelevant.

B. Attribute subset Selection contd..

iv. Decision Tree Induction contd...

- Nonleaf nodes: tests
- Branches: outcomes of tests
- Leaf nodes: class prediction

Initial attribute set:
{A1. A2. A3. A4. A5. A6}



------> Reduced attribute set: {A1, A4, A6}

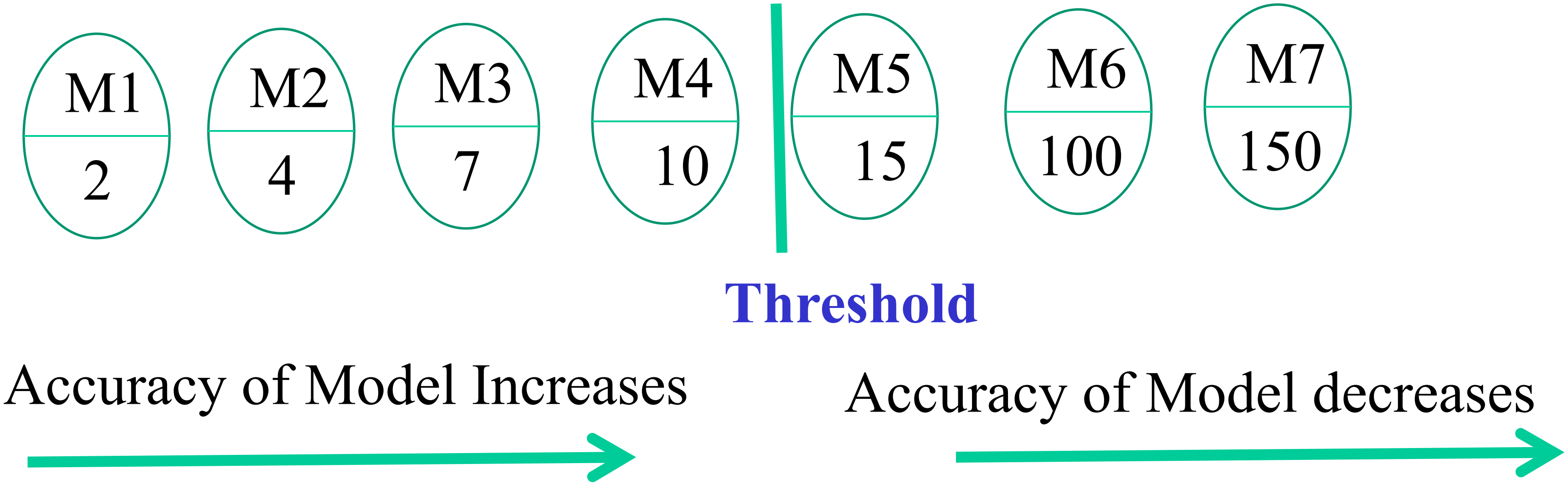
3. Data Reduction contd...

C. Dimensionality Reduction

Curse of Dimensionality:

- Dimensionality: The feature or attribute of the dataset
- Model:- Dataset will be input to the training phase, training phase will study the feature and output the model which can be able to perceive or interpret the similar kind of the object.

Following are differenet models developed from datasets with diffrent features but aiming for the same goal



M1 2

M2 4

M3 7

M4 10

M5 15

M6 100

M7 150

Threshold

Accuracy of Model Increases

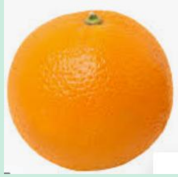
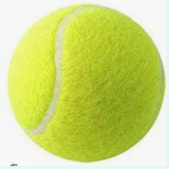
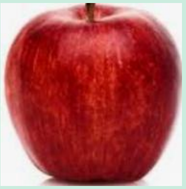
Accuracy of Model decreases

C. Dimensionality Reduction contd..

Curse of Dimensionality:

Example:

Let one cricket ball is given to the training phase and it study the object with 4 feature as Shape (sphere), Eatable (No), Play (yes), Color (red)

Object	Shape	Eatable	Play	Red Color
				
				
				

Here color dimension is the irrelevant dimension and plays the role of the curse of dimension for the model identifying the ball.

The threshold dimension for the above model is 3

C. Dimensionality Reduction contd..

- Dimensionality reduction is a method of converting the high dimensional variables into lower dimensional variables without changing the specific information of the variables.
- It represent the original data in the compressed or reduced form by appling data encoding or transformation.

- In the process of Compression the resultant data can be:

Lossless- If original data can be reconstructed form compressed data without reconstructing the whole.

Lossy- If we can construct only an approximation of original data.

Popular methods of Lossy dimensionality reduction are

i. Discrete Wavelet Transform (DWT) (Sparse matrix created)
ii. Principal component Analysis (PCA) (Combines the essence of attributes by creating an alternative, smaller set of variables)

C. Dimensionality Reduction contd..

i. Discrete Wavelet Transform (DWT)

What is Wavelet?

- Wavelets are mathematical functions

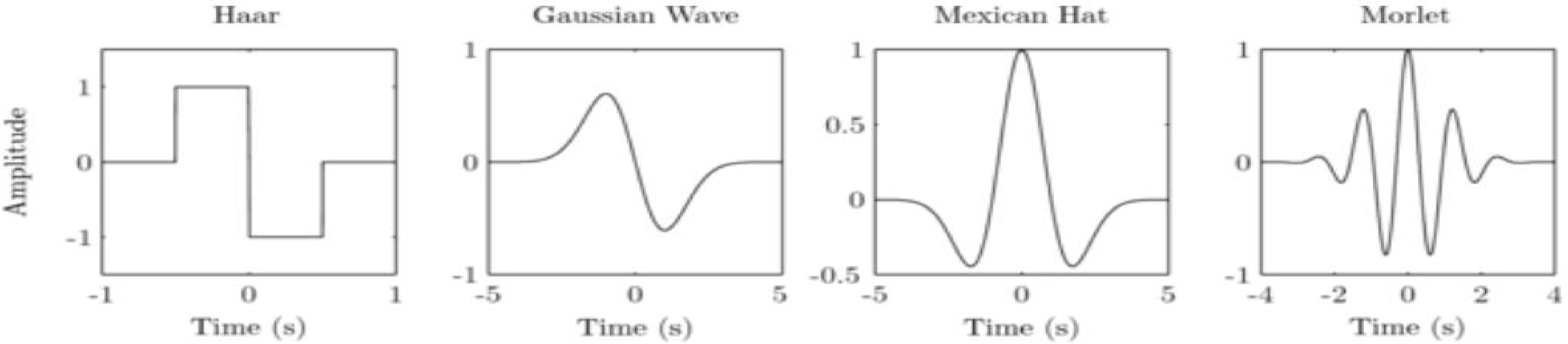
What does it Do?

- Cut off data into different frequency components and then study each component with a resolution matched to its scale.

Why it is needed?

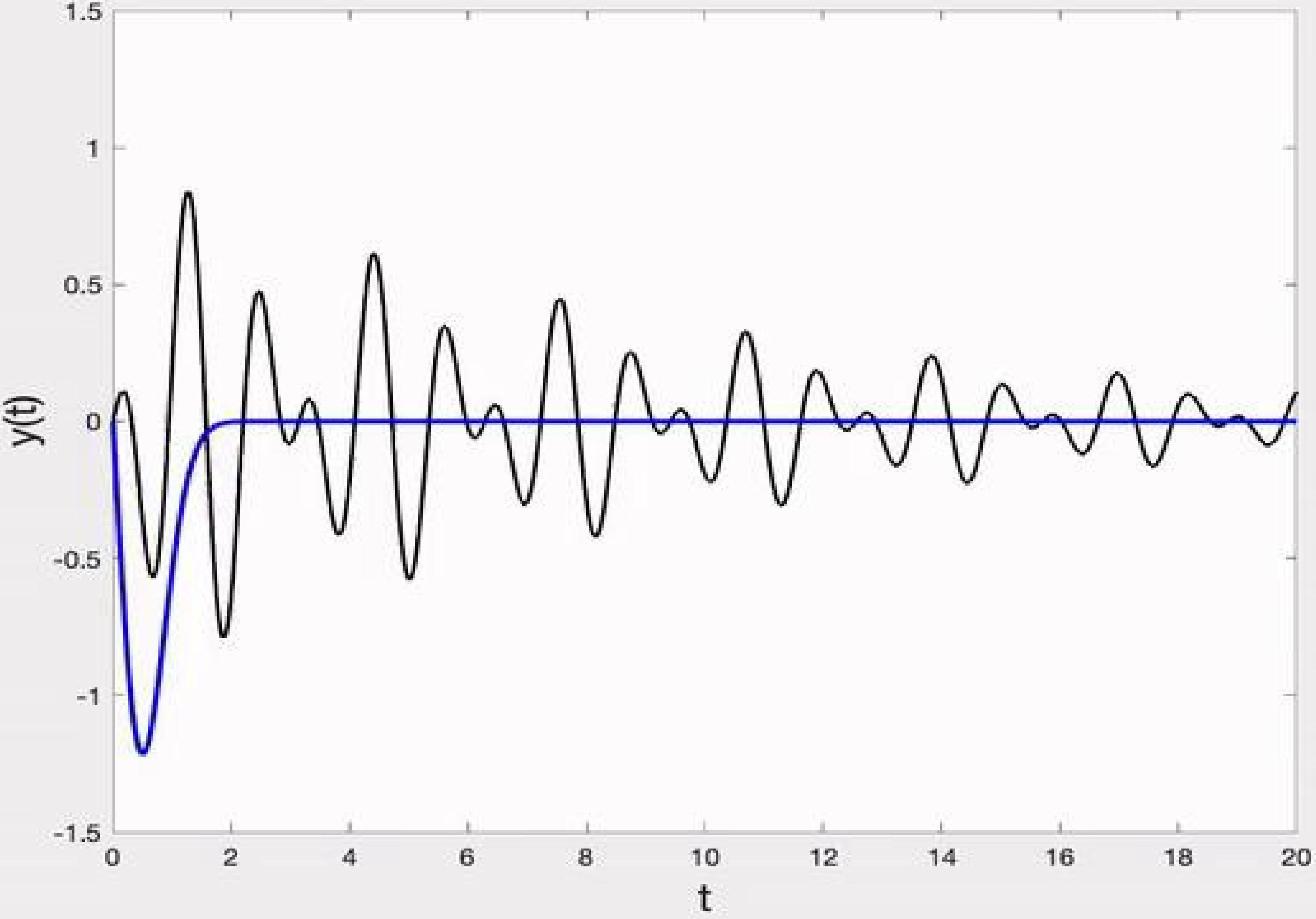
- Analyzing discontinuity and sharp spikes of the signal.
- Applications are image processing through image compression, human vision, earthquake prediction, radar, computer graphics

Types of Wavelet:



C. Dimensionality Reduction contd..

i. Discrete Wavelet Transform (DWT)



C. Dimensionality Reduction contd..

i. Discrete Wavelet Transform (DWT)

- The discrete wavelet transform (DWT) is a linear signal processing technique that transform a data vector X to a same length but numerically different vector X' of **wavelet coefficients**.

- When using this technique for data reduction, it can consider each tuple as an n-dimensional data vector, that is, *X* =(x1,x2,…xn) indicating n measurements made on the tuple from n database attributes. The usefulness of reduction lies in the fact that the wavelet transformed data can be **truncated** and the usefulness of the compression lies by **storing only a small fraction of the strongest** of the wavelet coefficient.

- DWT is similar to discrete Fourier Transformation but achieves better Lossy compression, i.e. when reduced data gives more accurate approximation of original data and also localized in space.

C. Dimensionality Reduction contd..

ii. Principal Component Analysis (PCA)

Problem of Overfitting:

To generate a model, when more attibutes/features will be fed at the trainning phase then one confusing/erroneous model will be generated.

Solution is PCA

- It is a technique that is useful for the compression and classification of data. The purpose is to reduce the dimensionality of a data set (sample) by finding a new set of variables, smaller than the original set of variables, that retains most of the sample's information.
- PCA should be used when the variables/features are strongly correlated.

C. Dimensionality Reduction contd..

ii. Principal Component Analysis (PCA) contd..

Problem of Overfitting:

To generate a model, when more attributes/features will be fed at the training phase then one confusing/erroneous model will be generated.

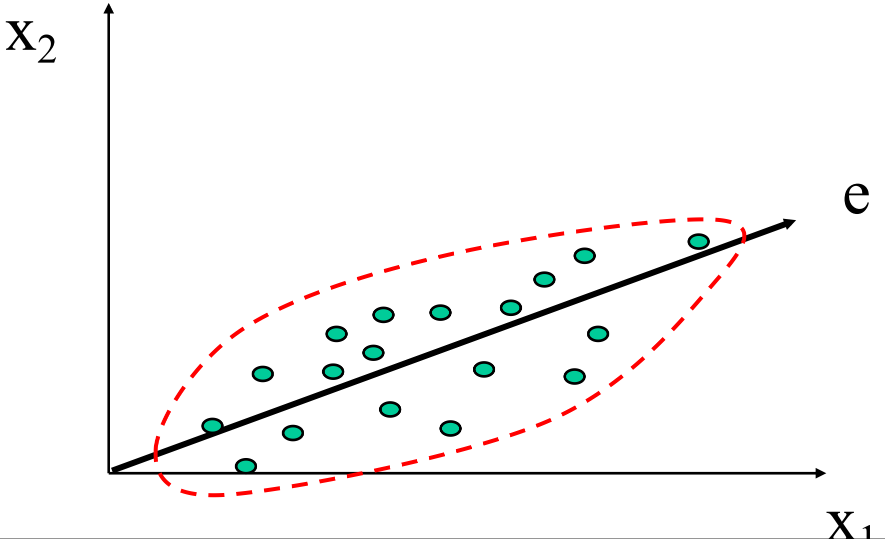
Solution is PCA

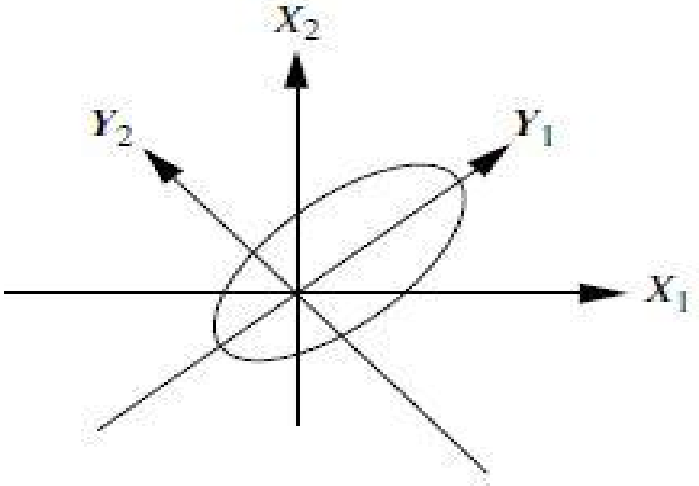
- It is a technique that is useful for the compression and classification of data. The purpose is to reduce the dimensionality of a data set (sample) by finding a new set of variables, smaller than the original set of variables, that retains most of the sample's information.
- PCA should be used when the features are strongly correlated.
- Suppose given is tuples/vectors described by n attributes/dimension then PCA search for ‘k’ n-dimensional orthogonal tuples/vectors that can be best fit to represent the data where k<=n. Thus the original data is projected into a much smaller space

C. Dimensionality Reduction contd..

ii. Principal Component Analysis (PCA)

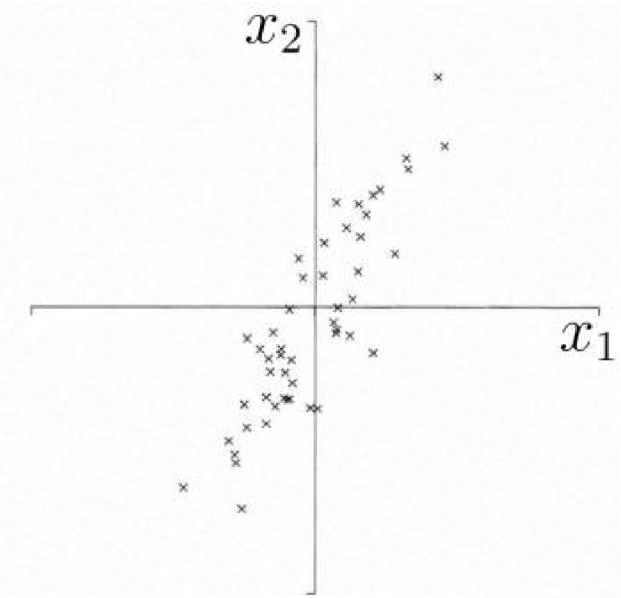
- Find the projections in a step-wise manner that captures the largest amount of variation in data. The projection can be named PC_1,...PC_k where k<=n.
- The original data are projected onto a much smaller space, resulting in dimensionality reduction.
- Attribute subset selection reduces attribute set size by considering a subset of the initial set of attribute where PCA combines the essence of attributes by creating an alternative smaller attribute set.



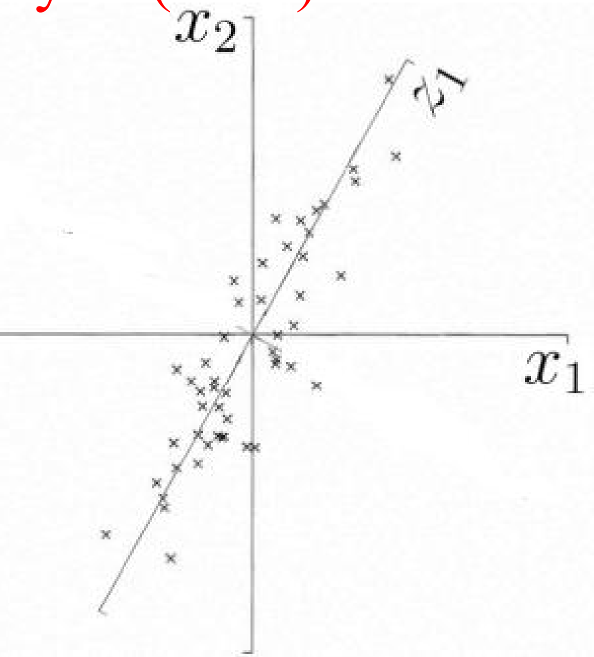


C. Dimensionality Reduction contd..

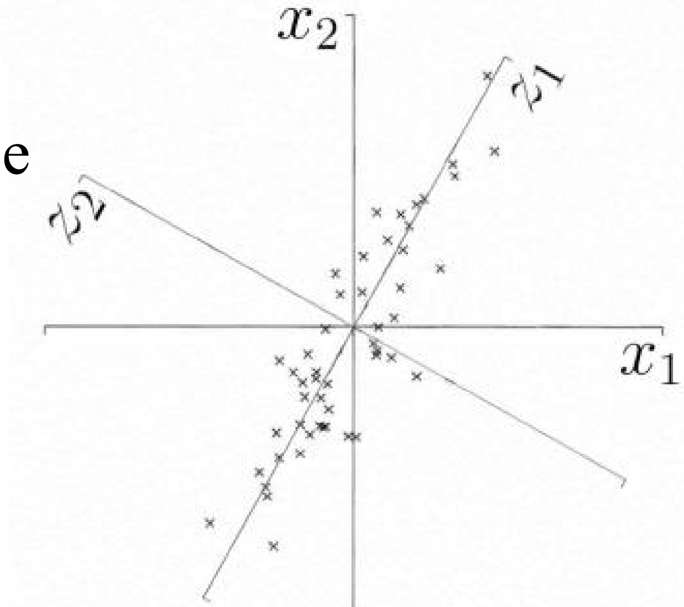
ii. Principal Component Analysis (PCA)



I: Data in 2-D Space



II: Application of PC1 in the direction of large variation of data



III. Application of PC2 after PC1 in the next level of larger variation of data

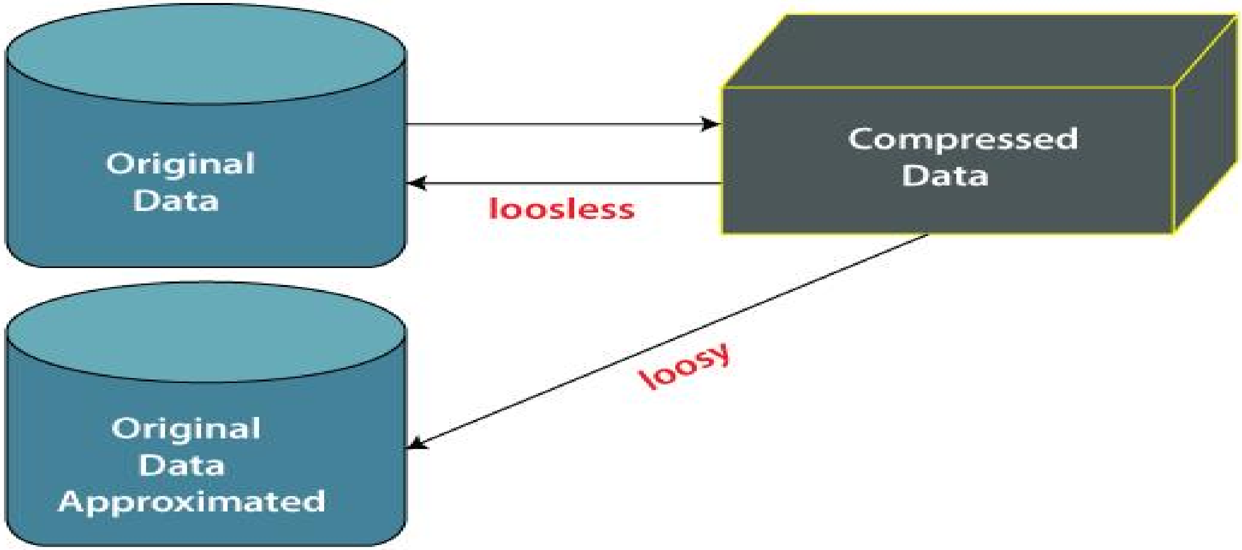
C. Dimensionality Reduction contd..

ii. Principal Component Analysis (PCA)

- Given N data vectors from n-dimensions, find $k \leq n$ orthogonal vectors (*principal components*) that can be best used to represent data
  - Normalize input data: Each attribute falls within the same range
  - Compute k orthonormal (unit) vectors, i.e., *principal components*
  - Each input data (vector) is a linear combination of the k principal component vectors
  - The principal components are sorted in order of decreasing "significance" or strength
  - Since the components are sorted, the size of the data can be reduced by eliminating the *weak components*, i.e., those with low variance (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data)
- Works for numeric data only

3. Data Reduction contd...
D. Data Compression

- Data compression employs modification, encoding, and converting the structure of data in a way that consumes less space.
- It involves building a compact representation of information by removing redundancy and representing data in binary form.
- Data that can be restored successfully from its compressed form is called Lossless compression. In contrast, the opposite where it is not possible to restore the original form from the compressed form is Lossy compression.



3. Data Reduction contd...

D. Data Compression contd...

- Data compression technique reduces the size of the files using different encoding mechanisms. Based on their compression techniques it can be divides into two types

1. Lossless Compression: Encoding techniques like Run Length Encoding and Huffman Encoding allows a simple and minimal data size reduction. Lossless data compression uses algorithms to restore the precise original data from the compressed data.

1. Lossy Compression: In lossy-data compression, the decompressed data may differ from the original data but are useful enough to retrieve information from them. For example, the JPEG image format is a lossy compression, but we can find the meaning equivalent to the original image. Methods such as the Discrete Wavelet transform technique PCA (principal component analysis) are examples of this compression.

3. Data Reduction contd...

E. Numerosity Reduction

The numerosity reduction reduces the original data volume by alternative smaller data representations. This technique includes two types parametric and non-parametric numerosity reduction.

i. Parametric methods

- Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
- E.g.: Regression Model: Linear, Multiple, Log-linear regression

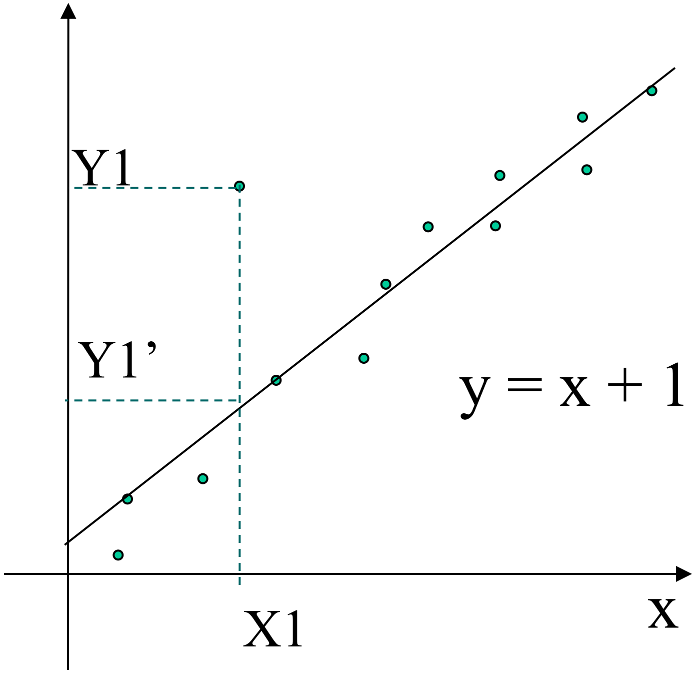
ii. Non-parametric methods

- Do not assume models. This technique results in a more uniform reduction, irrespective of data size, but it may not achieve a high volume of data reduction like the parametric.
- E.g.: Histograms, clustering, sampling

E. Numerosity Reduction contd..

i. Parametric Method: Regression Model

- **Regression analysis:** A collective name of techniques for the modeling and analyzing of numerical data consisting of values of a dependent variable/response variable/measurement Vs. one or more independent variables /explanatory variables/predictors.



- The parameters are estimated to give a **"best fit"** of the data
- Most commonly the best fit is evaluated by using the **least squares method**, but other criteria have also been used
- **Application:** prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

E. Numerosity Reduction contd..

i. Parametric Method: Regression Model

- **Linear regression**: It attempts to model the relationship between two variables by fitting a linear equation to observed data.

$$Y = \alpha + \beta X$$

  - ➢ Two parameters , α and β specify the line and are to be estimated by using the data at hand and then the fitness of the model is evaluated using least squares criterion.
  - ➢ Let the modeler wants to relate the weights of individuals to their heights using a linear regression model.

- **Multiple regression**: It is used to predict the outcome of a variable based on the value of two or more variables. It is an extension of linear regression.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \dots \dots + \beta_n x_n + e$$

Let the modeler wants to relate age and intelligence of individuals to their CGPA using a multiple-linear regression model.

E. Numerosity Reduction contd..

i. Parametric Method: Regression Model

- **Log-Linear regression:** Logistic Regression is used when the dependent variable (target) is categorical.

It measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function which is the cumulative logistic distribution. The predicted values are probabilities and thus are restricted to (0, 1)

There are 3 types of Logistic Regression

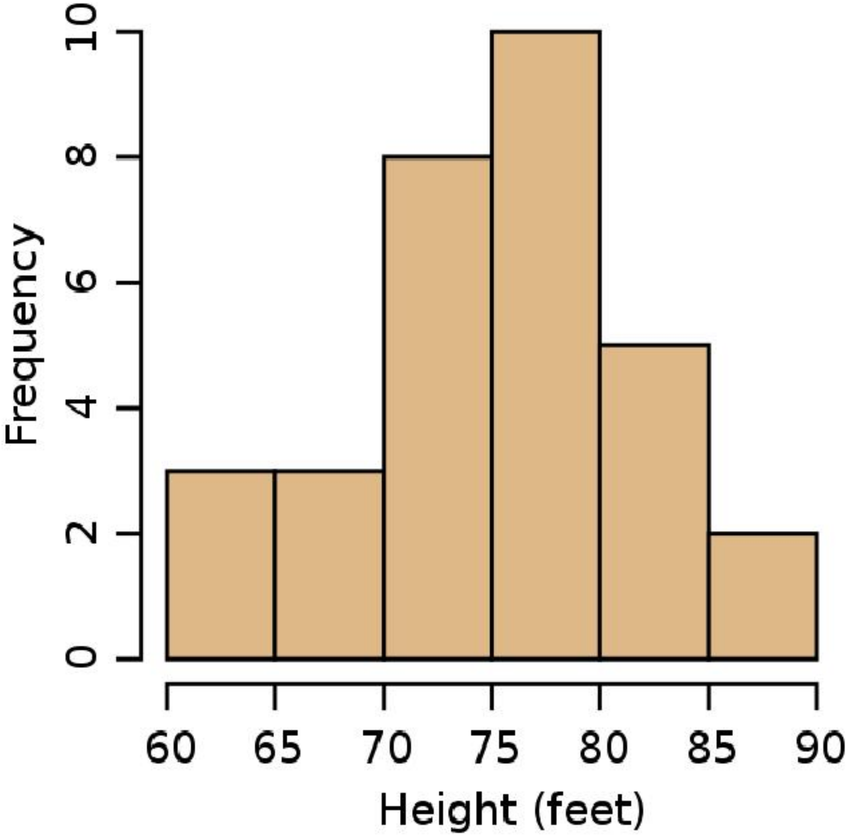
- **Binary Logistic Regression:** The categorical response has only two 2 possible outcomes. Example: Spam or Not.
- **Multinomial Logistic Regression:** Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan)
- **Ordinal Logistic Regression:** Three or more categories with ordering. Example: Movie rating from 1 to 5.

E. Numerosity Reduction contd..

ii. Non-Parametric Method: a) Histograms

- Histograms use binning to approximate data distributions and are a popular form of data reduction
- A histogram for an attribute A partitions the data distribution of A into disjoint subsets, or buckets.

- The buckets are displayed on a horizontal axis, while the height (and area) typically reflects the average frequency of the values represented by the bucket [Often buckets represent continuous ranges for the given attribute].
- Can be constructed optimally in one dimension using dynamic programming
- Related to quantization problems.



E. Numerosity Reduction contd..

ii. Non-Parametric Method: **b) Cluster**

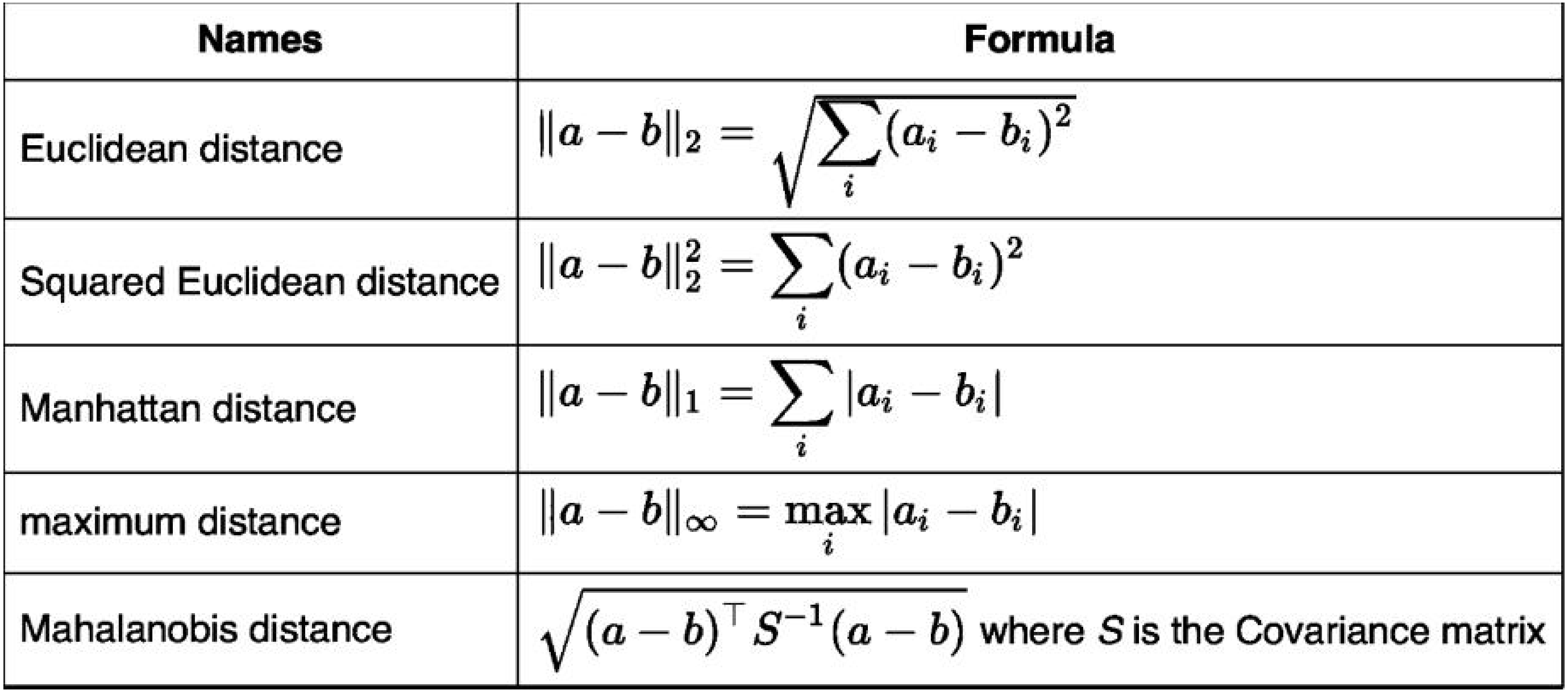
- Partition data set into clusters/groups, by finding similarity and dissimilarity among objects.
- The Similarity is measured interns of the closeness of objects in space based on a **distance function**.

The popular types of clustering are:

- Connectivity models: data points closer in data space exhibit more similarity.
- Centroid models: by the closeness of a data point to the centroid of the clusters
- Distribution models: how probable is it that all data points
- Density models: It isolates different density regions

- In data reduction, cluster representations are used to replace actual data
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms.

Distance Function



Names	Formula
Euclidean distance	$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
Squared Euclidean distance	$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
Manhattan distance	$\|a-b\|_1 = \sum_i \lvert a_i - b_i \rvert$
maximum distance	$\|a-b\|_\infty = \max_i \lvert a_i - b_i \rvert$
Mahalanobis distance	$\sqrt{(a-b)^\top S^{-1}(a-b)}$ where *S* is the Covariance matrix

E. Numerosity Reduction contd..

ii. Non-Parametric Method: **c) Sampling**

- The Sampling technique selects a subset of data to be analyzed. Instead of dealing with an entire data, it selects the instances at periodic intervals.
- Sampling is used to compute the expected values of the data sets. It reduces the amount of data to be processed and the computational costs
- The main problem is to obtain a representative sample i.e. a subset of data that has approximately the same properties of the original data.

Types of Sampling:

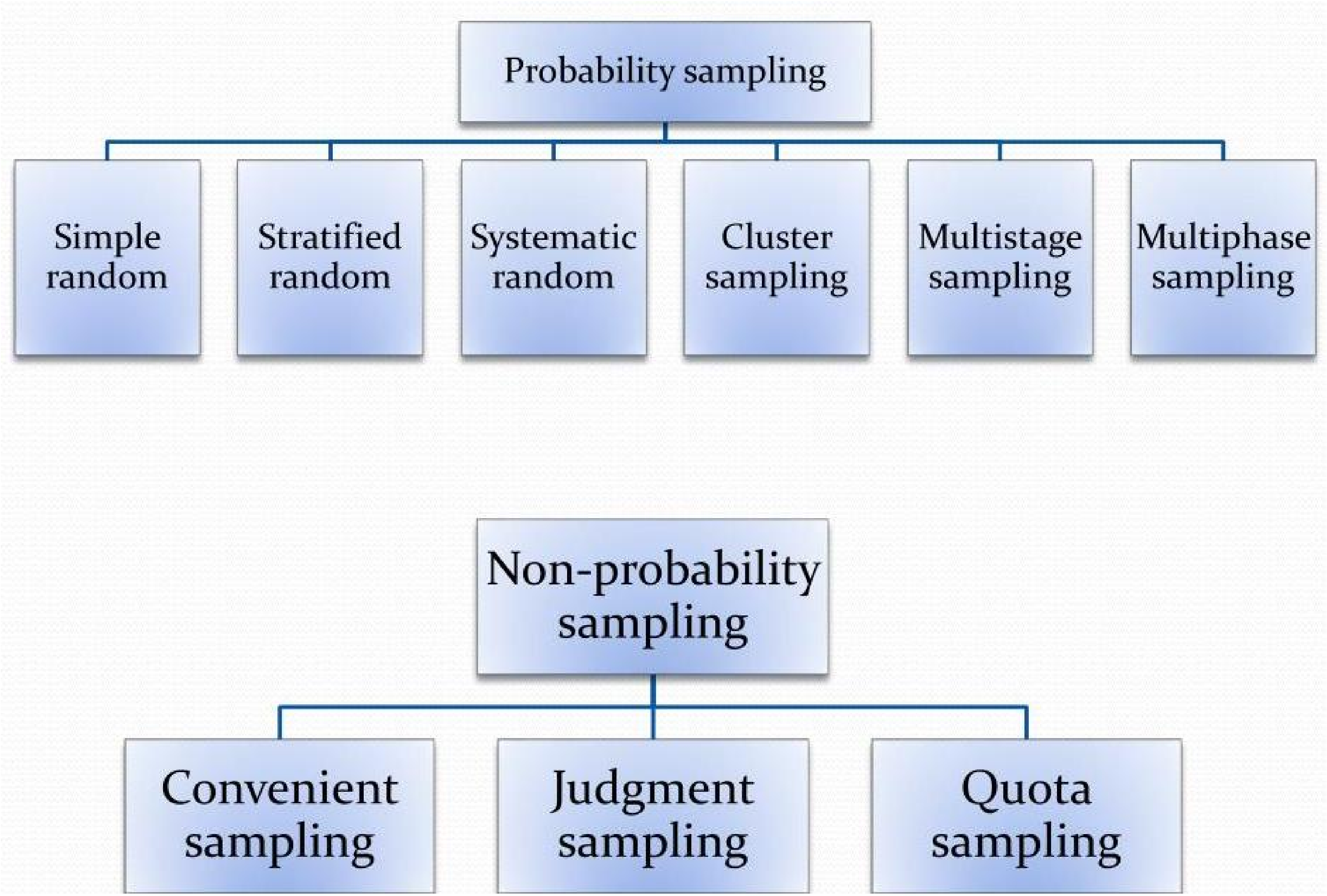
1. Probability sampling: It is random sampling in which each item in the population has an equal chance of getting selected.

2. Non-probability sampling: A non-probability sampling method is the way of sampling that relies on the subjective judgment of the researcher.

- Sampling may not reduce database I/Os but is a natural choice for progressive refinement of a reduced data set.

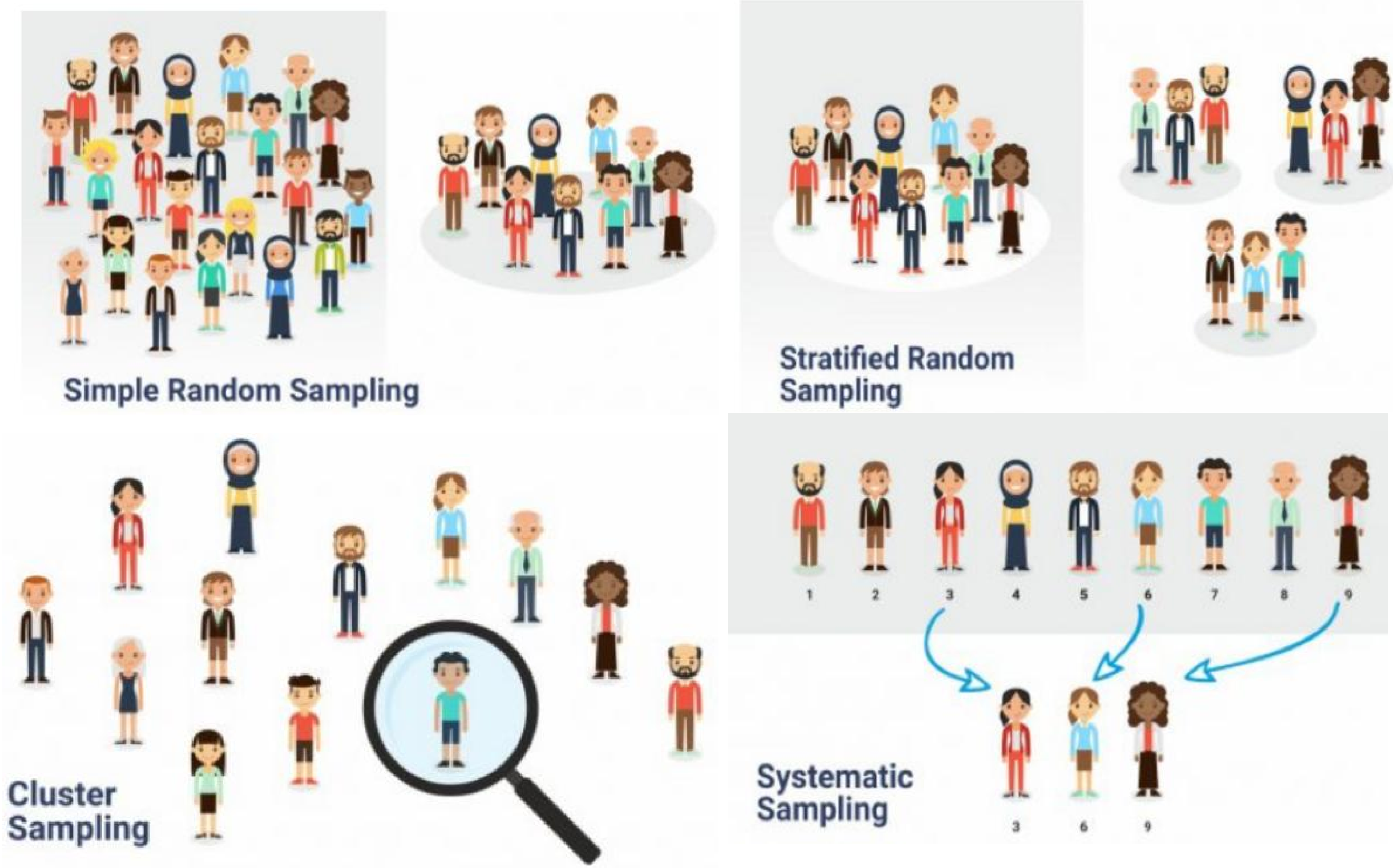
E. Numerosity Reduction contd..

ii. Non-Parametric Method: **c) Sampling**



E. Numerosity Reduction contd..

ii. Non-Parametric Method: **c) Sampling**



Agenda

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

4. Discretization and concept hierarchy generation

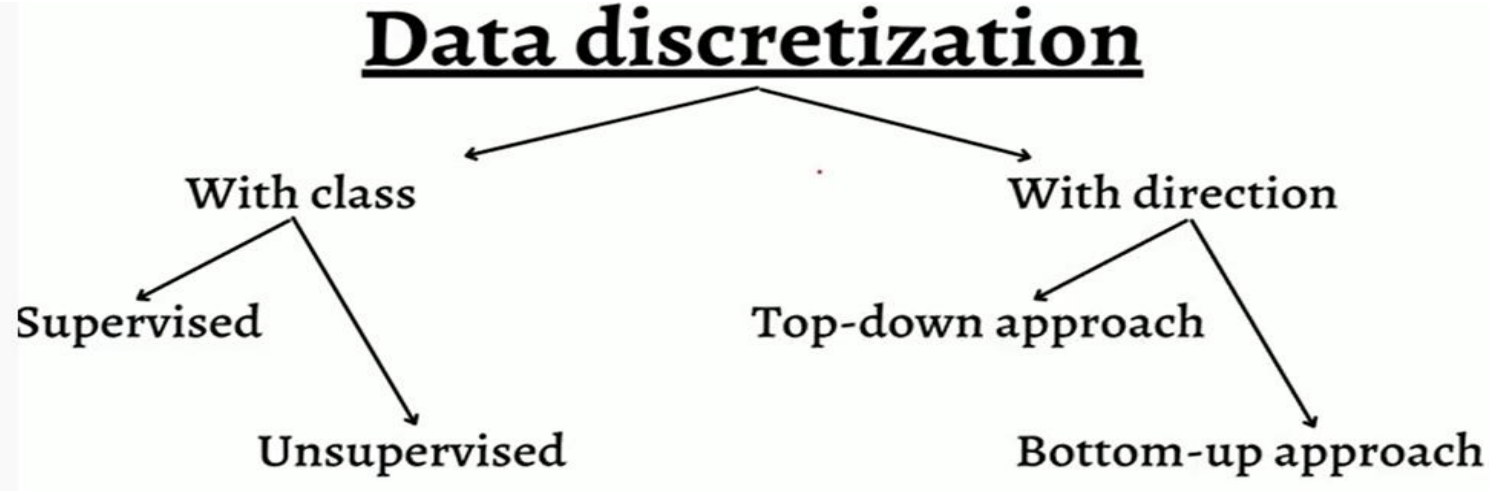
Discretization techniques: It is the process to reduce the number of values for a given continuous attribute, by dividing the attribute into a range of intervals. Interval value labels can be used to replace actual data values.

Concept hierarchies: It reduces the data by collecting and replacing low-level concepts (such as numeric values for the attribute age) by higher-level concepts (such as young, middle-aged, or senior).

This leads to a concise, easy-to use, knowledge-level representation of mining results.

4. Discretization and concept hierarchy generation contd..

These are recursive methods where a large amount of time is spent on sorting the data at each step. The smaller the number of distinct values to sort, the faster these methods can be.



Data discretization

With class

Supervised

Unsupervised

With direction

Top-down approach

Bottom-up approach

Discretization techniques can be categorized based on whether it uses class information or not such as follows:

- **Supervised Discretization** - This discretization process uses class information.
- **Unsupervised Discretization** - This discretization process does not use class information.

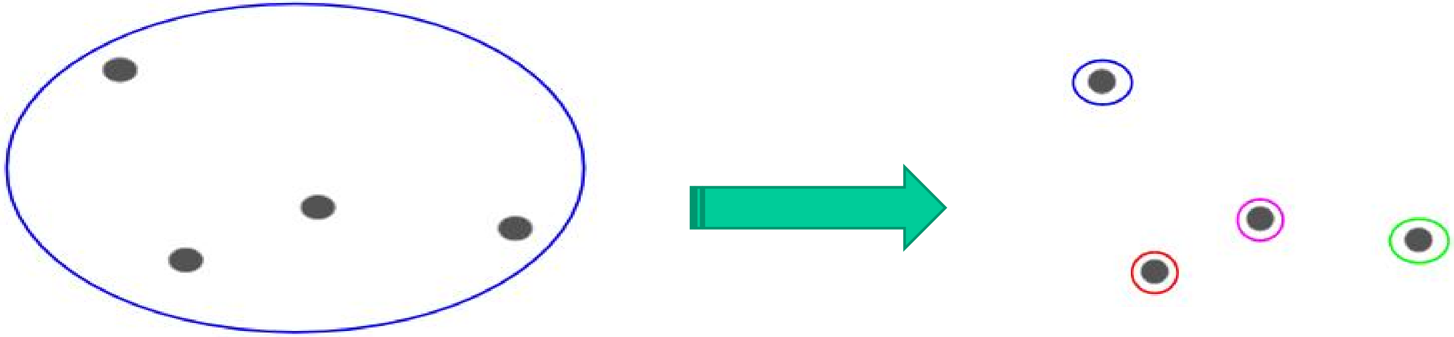
Discretization techniques can be categorized based on which direction it proceeds as follows:

- Top-down approach
- Bottom-up Approach

4. Discretization and concept hierarchy generation contd..

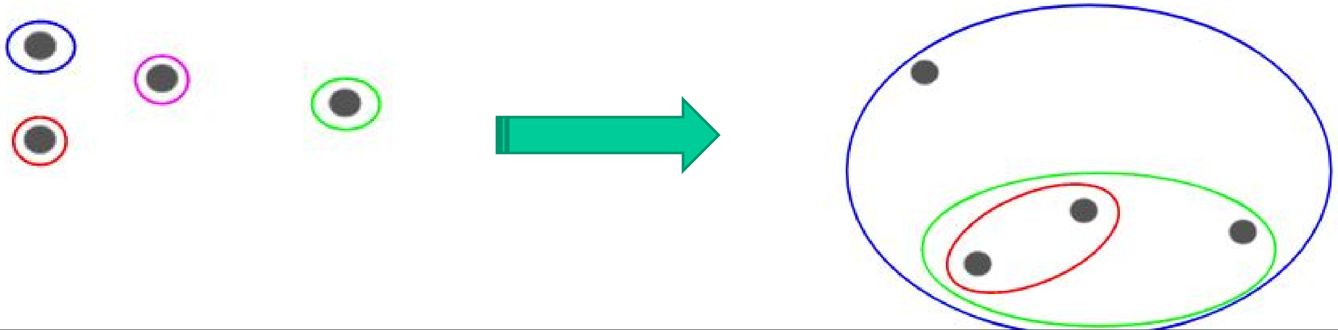
Top-down Discretization (Splitting) -

If the process starts by first finding one or a few points called split points or cut points to split the entire attribute range and then repeat this recursively on the resulting intervals.



Bottom-up Discretization (Merging) -

Starts by considering all of the continuous values as potential split-points. Removes some by merging neighborhood values to form intervals, and then recursively applies this process to the resulting intervals.



4.. Discretization and concept hierarchy generation **contd..**

Many discretization techniques can be applied recursively to provide a hierarchical or multiresolution partitioning of the attribute values known as concept hierarchy.

A **concept hierarchy** for a given numeric attribute defines a discretization of the attribute.

Concept hierarchies can be used to reduce the data y collecting and replacing low-level concepts (such as numeric value for the attribute age) with higher-level concepts (such as young, middle-aged, or senior).

Although detail is lost by such generalization, it becomes meaningful and it is easier to interpret.

Concept hierarchy formation: Recursively reduce the data by collecting and replacing low-level concepts (such as numeric values for *age*) by higher-level concepts (such as *youth, adult*, or *senior*)

4. Discretization and concept hierarchy generation **contd..**

Five methods for discretization & concept hierarchy f**or numeric data** are defined:

i. Binning

ii. Histogram

iii. Cluster Analysis

iv. Decision Tree/Entropy-Based Discretization

v. Correlation Analysis (Chi square)

i. Binning:

It is a top-down un-supervised splitting technique because it is based on a specified number of bins but without having any class information.

The sorted values are distributed into several buckets or bins and then replaced with each bin value by the use of bin mean/median/ boundaries. It is further classified into

- Equal-width (distance) partitioning
- Equal-depth (frequency) partitioning

4. Discretization and concept hierarchy generation **contd..**

ii Histogram Analysis:

It is a top-down unsupervised discretization technique because histogram analysis does not use class information.

Histograms partition the values for an attribute into disjoint ranges called buckets.

It is also further classified into

- Equal-width histogram
- Equal frequency histogram

The histogram analysis algorithm can be applied recursively (top-down) to each partition to automatically generate a multilevel concept hierarchy, with the procedure terminating once a pre-specified number of concept levels has been reached.

4. Discretization and concept hierarchy generation contd..

iii Cluster Analysis:

It is a popular data discretization method that can be applied to discretize a numerical attribute of A by partitioning the values of A into clusters or groups.

Cluster Analysis is unsupervised learning and uses top-down/bottom-up approach.

Clustering considers the distribution of A, as well as the closeness of data points, and therefore can produce high-quality discretization results.

Each initial cluster or partition may be further decomposed into several subcultures, forming a lower level of the hierarchy.

There are mainly two types of hierarchical clustering:

- Agglomerative hierarchical clustering
- Divisive Hierarchical clustering

4. Discretization and concept hierarchy generation **contd..**

iv. Decision Tree/Entropy-Based Discretization (supervised & top-down)

- The Decision tree is an entropy-based discretization splitting technique that explores class distribution information and determines exact split points using calculation.
- To discretize a numeric attribute, A, this method selects the value of A that has minimum entropy [a measure] as a split-point and recursively partitions the resulting intervals to arrive at a hierachical discretization. This forms a concept hierarchy for A.

4. Discretization and concept hierarchy generation contd..

v. Correlation analysis (e.g., Chi-merge: χ^2-based discretization)

- Supervised: use class information
- Bottom-up merge: find the best neighboring intervals (those having similar distributions of classes, i.e., low χ^2 values) to merge
- Initially, each numeric distinct value of attribute A is considered to be one interval. Then χ^2 tests are performed on every pair of adjacent intervals and the intervals with the least χ^2 value are merged recursively until a predefined stopping condition

4. Discretization and concept hierarchy generation contd..

Four methods of Concept hierarchy generation for **non-numeric (categorical/nominal) data**
Nominal attributes have a finite but possibly large number of distinct values which is difficult to do partitioning in terms of ordering. For this concept hierarchy is used to transform the data into multiple levels of granularity. Following are the four methods:

i. Specification of a partial ordering of attributes explicitly at the schema level by users or experts

- Users/experts can easily define a concept hierarchy by partial/total ordering of the schema attribute at the schema level.
- At the schema level, a hierarchy can be defined by specifying the ordering among the attribute as:

street < city < state < country

4. Discretization and concept hierarchy generation **contd..**

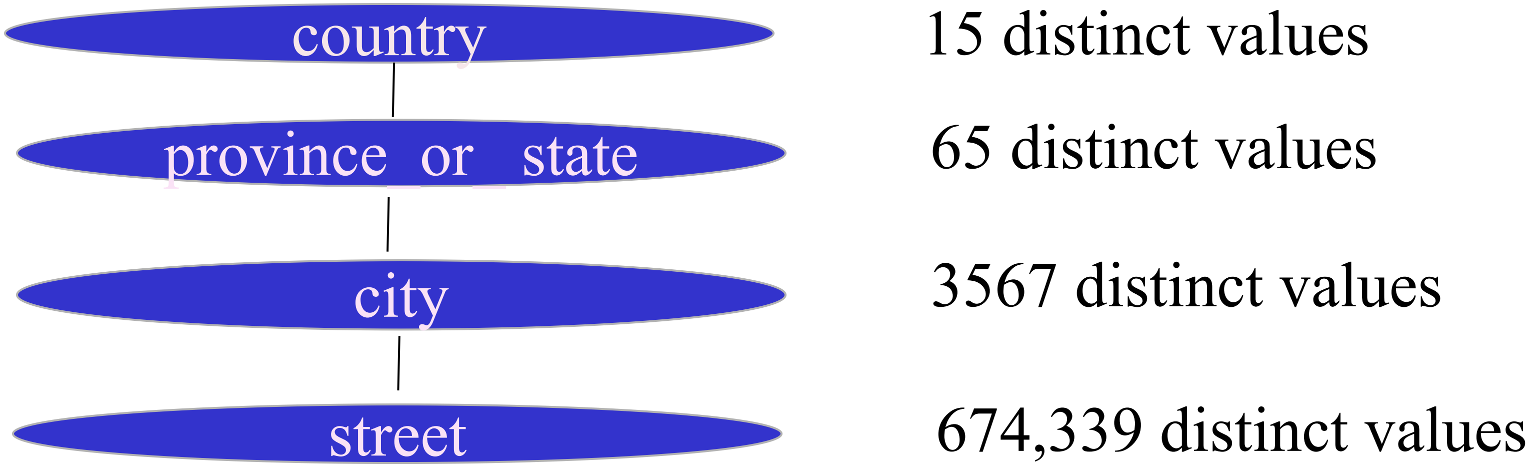
ii. Specification of a portion of a hierarchy by explicit data grouping

- This is the manual definition of a portion of the concept hierarchy.
- On the large database, on the attribute country (example) explicit groupings can be made to achieve a small portion of intermediate level data

{Gujurat, Maharastra, Goa...} ⊂ Western India,
{Assam, Manipur} ⊂ Eastern India

4. Discretization and concept hierarchy generation **contd..**

iii. Specification of a set of attributes, but not of their partial ordering



country — 15 distinct values

province_or_state — 65 distinct values

city — 3567 distinct values

street — 674,339 distinct values

- Based on the observation that high concept level (country) attribute usually contain a smaller number of distinct value than lower concept level (street) attribute, a concept hierarchy can be generated based on the number of distinct values per attribute in the given attribute set.
- The attribute with the most distinct value is placed in the lowest hierarchy and so on.

4. Discretization and concept hierarchy generation contd..

iv. Specification of only a partial set of attributes

- At the time schema design, sometimes ur carelessly include only a small subset of the relevant attribute (partial) in the hierarchy specification.

  e.g. including street & city in hierarchical attribute set location
- Solution: Embed data semantics in the database schema => attributes with tight semantic connection can be pinned together.

Agenda

- Why preprocess the data?
- Data cleaning
- Data integration and transformation
- Data reduction
- Discretization and concept hierarchy generation
- Summary

Summary

- Data preparation is a big issue for both warehousing and mining
- Data preparation includes
  - Data cleaning and data integration
  - Data reduction and feature selection
  - Discretization
- A lot of methods have been developed but still an active area of research