DWDM

UNIT-1

Dr. Shalini Gambhir

Data Preprocessing

I-Data Cleaning,
II-Data Integration,
III-Data Reduction,
III-Data Transformation and Data Discretization,

Data Preprocessing

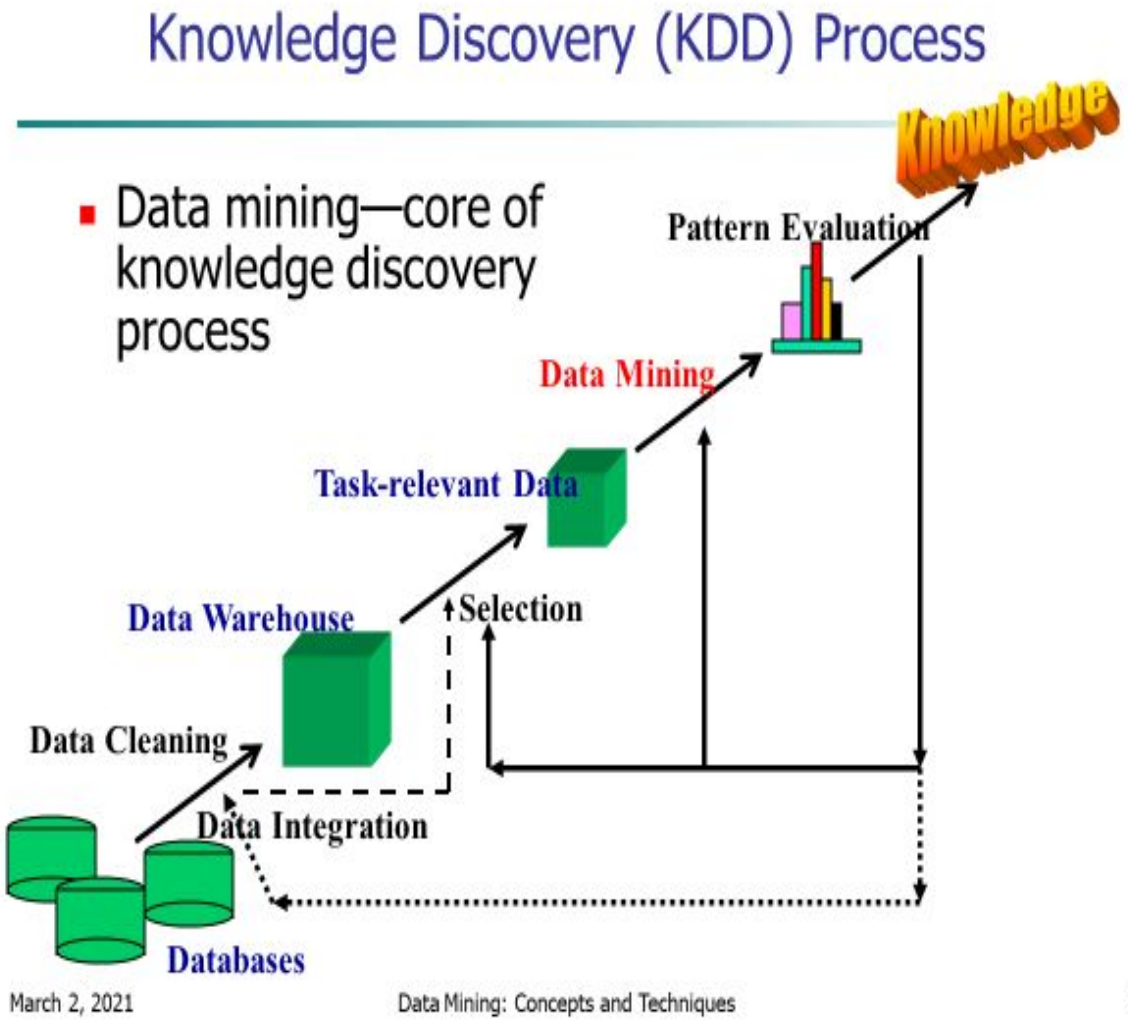
- Data cleaning
  - Missing values
  - Noisy Data
    - (i) Binning
    - (ii) outlier
    - (iii) Regression
- Integration
  - Schema integration
  - Entity identification
  - Values
  - Handling redundancy
    - (i) Covariance
    - (ii) correlation
    - (iii) chi-square test
- Transformation
  - Normalization
    - (i) Max-min
    - (ii) Z-score
    - (iii) Decimal scaling
  - Desorization
    - (i) Binning
    - (ii) Clustering
    - (iii) classification
    - (iv) Concept Hierarchy
- Reduction.
  - Dimensionality
    - (i) Wavelets
    - (ii) PCA
    - (iii) Compression
    - (iv) Feature Subset selection
    - (v) Feature generation
  - Numerosity
    - (i) Data cubes Aggregation
    - (ii) Sampling

Data Quality: Why Preprocess the Data?

- There are many factors/measures comprising **data quality**
- Measures for data quality: A multidimensional view
  - **Accuracy:** correct or wrong, accurate or not, errors from instruments , data transmission errors
  - **Completeness**: not recorded, unavailable, user interested attributes may not be available causing unfilled data
  - **Consistency**: some modified but some not, dangling, …
  - **Timeliness:** timely update? Several sales representatives, however, fail to submit their sales records on time at the end of the month.
    - For a period of time following each month, the data stored in the database are incomplete.
    - However, once all of the data are received, it is correct.
  - **Believability**: how trustable the data are correct?
  - **Interpretability**: how easily the data can be understood?
    - Suppose that a database, at one point, had several errors, all of which have since been corrected.
    - The data also use many accounting codes, which the sales department does not know how to interpret. Even though the database is

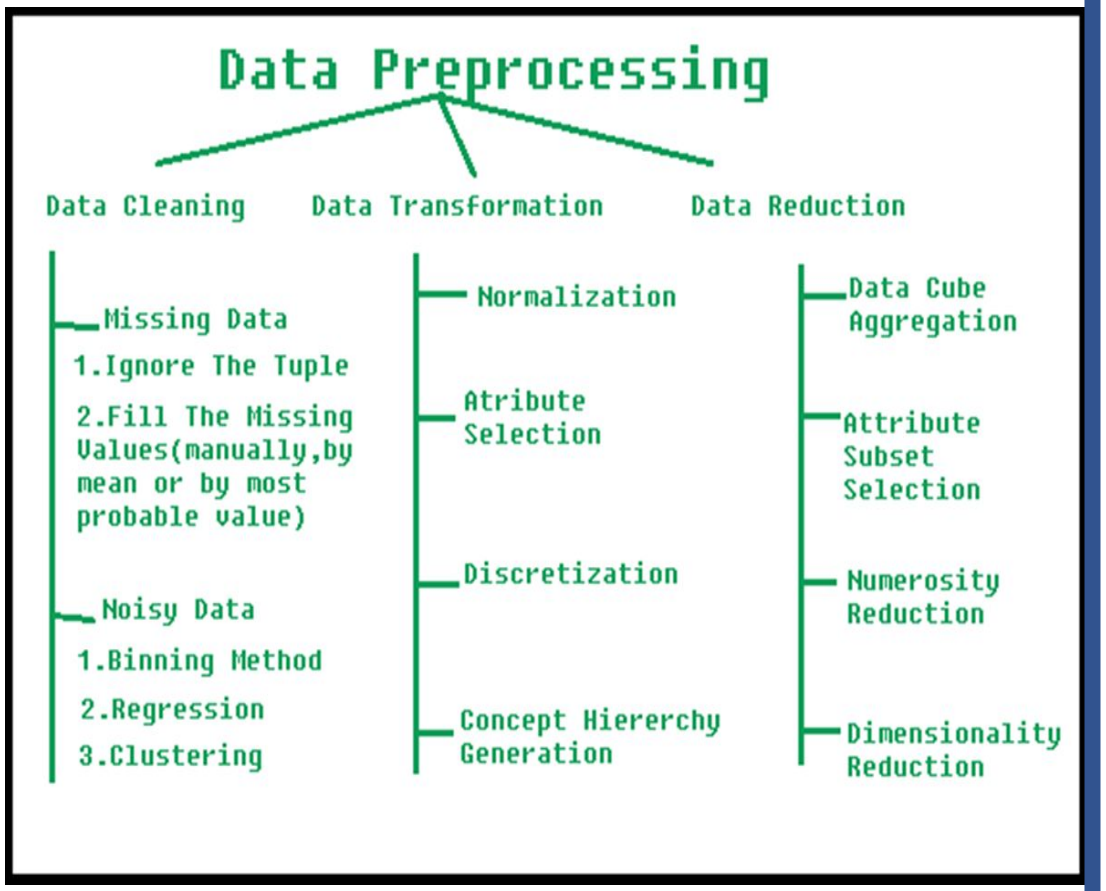
Data Mining as Knowledge Discovery

- **Data cleaning** - to remove noise or irrelevant data
- **Data integration** - where multiple data sources may be combined
- **Data selection-** where data relevant to the analysis task are retrieved from the database
- **Data transformation** -where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations
- **Data mining** - an essential process where intelligent methods are applied in order to extract data patterns
- **Pattern evaluation** to identify the truly interesting patterns representing knowledge based
- **Knowledge presentation** - where visualization and knowledge representation techniques are used to present



Major Tasks/Steps in Data Preprocessing

- **Data cleaning**
  - Fill in missing values, smooth noisy data,
  - identify or remove outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data reduction**
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation



I-Data Cleaning

I- Data Cleaning: Incomplete (Missing) Data

- Data is not always available
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - 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 may need to be inferred

Name	Age	Sex	Income	Class
Mike	40	Male	150k	Big spender
Jenny	20	Female	?	Regular
...				

Data Cleaning: How to Handle Missing Data?

- **Ignore the tuple:** usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably
- **Fill in the missing value manually**: tedious + infeasible?
- Fill in it automatically with
  - **a global constant** : e.g., “unknown”, a new class?!
  - **the attribute mean**
  - **the attribute mean for all samples belonging to the same class:** smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree, or **use regression to fill**

Data Cleaning: Noisy Data

- Noise: random error or variance in a measured variable
- Incorrect attribute values may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

Data Cleaning: How to Handle Noisy Data?

Binning

first sort data and partition into (equal-frequency) bins

then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

Regression

smooth by fitting the data into regression functions

Clustering/Outlier

detect and remove outliers

Combined computer and human inspection

detect suspicious values and check by human (e.g., deal with possible outliers)

Data Cleaning: Handling Noisy Data: Binning

- "Data smoothing is the technique to handle noisy data
- "smooth" out the data to remove the noise
- data smoothing techniques.
  - **Binning**
  - **Regression**
  - **Outlier analysis**

- **Binning:**
- This method works on sorted data
- The sorted data is divided into equal frequency buckets/bins
- Binning is of three types:
  1. **smoothing by bin means**
     - each value in a bin is replaced by the mean value of the bin.
  2. **smoothing by bin medians**
     - each bin value is replaced by the bin median.
  3. **smoothing by bin boundaries**
     - the minimum and maximum values in a given bin are identified as the *bin boundaries*.
     - Each bin value is then replaced by the closest boundary value.

Data Cleaning: Binning methods

- **Equal-width (distance) partitioning**
  - Divides the range into *N* intervals of equal size: uniform grid
  - if *A* and *B* are the lowest and highest values of the attribute, the width of intervals will be: $W = (B - A)/N$.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well
- **Equal-depth (frequency) partitioning**
  - Divides the range into *N* intervals, each containing approximately same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

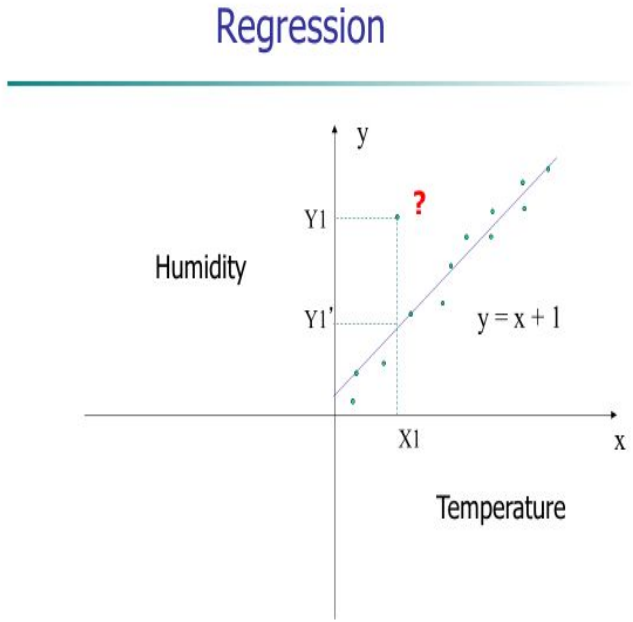
Data Cleaning: Binning Example:

- **Sorted data for *price* (in dollars)**: 4, 8, 15, 21, 21, 24, 25, 28, 34
- **Partition into (equal-frequency-length 3) bins**:
  - Bin1: 4,8,15 (max=15, min=4)
  - Bin 2: 21, 21, 24 (max=24, min=21)
  - Bin 3: 25, 28, 34 (max=34,min=25)
- **Smoothing by bin means**:
  - Bin1: 9,9,9. (mean of Bin1=4+8+15/3 =9)
  - Bin 2: 22, 22, 22 (mean of Bin2=21+21+24/3 =22)
  - Bin 3: 29, 29, 29
- **Smoothing by bin boundaries**:
  - Bin 1: 4,4,15. (4 is closer to min=4,replace 4 by 4. 8 is closer to min=4 so replace 8 by 4, 15 is closer to max=15,replace 15 by 15)
  - Bin 2: 21, 21, 24
  - Bin 3: 25, 25, 34

Use |x2-x1| as closeness measure

Data Cleaning: Handling Noisy Data: Regression

- **Regression:** Data smoothing can also be done by regression, a technique that conforms data values to a function.
- ***Linear regression*** involves finding the "best" line to fit two attributes (or variables) so that one attribute can be used to predict the other.
  - Y=mX+c
- ***Multiple linear regression*** is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.
  - Y=m1X1+m2X2+c
- **Example:**
  - Using various normal equations of Stats the best fitted line can be Y=2X.
  - Use Y=2X to predict the correct value at X=3 Which is 6 and replace(smooth) 5.6 by 6



X	0	1	2	3	4
Y	0	2	4	5.6	8

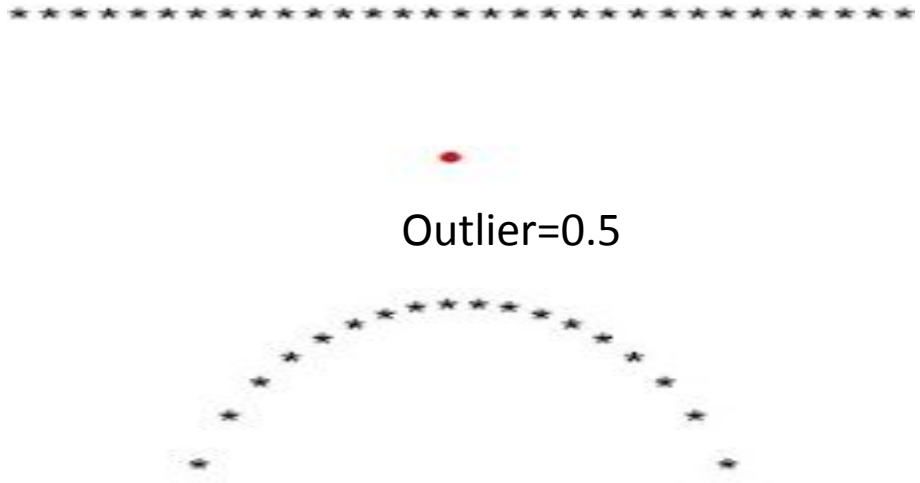
Error data

Data Cleaning: Handling Noisy Data: Outlier analysis

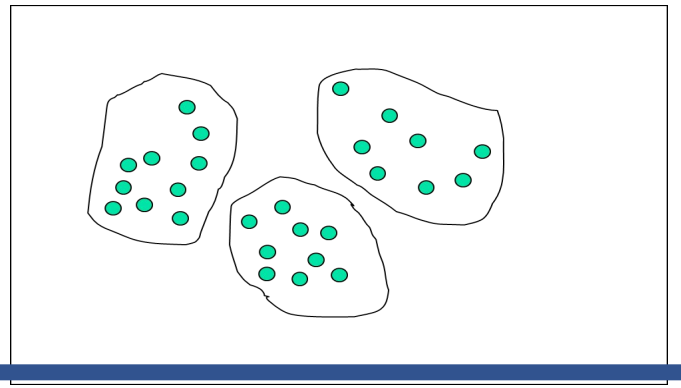
Outlier analysis: Outliers may be detected by clustering,

- where similar values are organized into groups, or “clusters.”
- Intuitively, values that fall outside of the set of clusters may be considered outliers

X	0	1	2	3	4
Y=x	0	1	0.5	3	4
$Y=x^2$	0	1	0.5	9	16



- When Y=X and Y=X2 are fitted then the
  Points satisfying these two equations
  are taken as two clusters and the errored value 0.5
  Which is not into any cluster is identified as outlier .



Handle missing values -Cleaning and Munging

- During **cleaning and munging in data science**, handling missing values is one of the most common tasks. The real-life data might contain missing values which need a fix before the data can be used for analysis. We can handle missing values by:

1. Either removing the records that have missing values or
2. Filling the missing values using some statistical technique or by gathering data understanding.

- A rule of thumb is that you can drop the missing values if they make up for less than five percent of the total number of records but however it depends on the analysis, the importance of the missing values, the size of the data, and the use case we are working on.

II-Data Integration

Data Integration

- Just imagine: your organization collects data from sales, customer service, website clicks, and third-party apps. All of these sources are different, but as data scientists, we want a single, unified view. That's exactly what **data integration** does- it combines data from multiple sources into one coherent dataset
- **Importance**: Without integration, analysis will be fragmented and misleading.
- **Real-world example**:
  Think of a bank:
- Loan records in one system
- Customer personal info in another
- Transactions in another
  → Integration gives a **single customer view**.

Data Integration:

Data integration:

Combines data from multiple sources into a coherent store.

Approaches in Data Integration

1. Entity identification problem

2. Tuple Duplication

3. Detecting and resolving data value conflicts

4. Redundancy and Correlation Analysis

1. Entity identification problem

- **(i) Entity Identification Problem**
- Different sources may refer to the same entity differently.
- Example: *William Clinton* vs *Bill Clinton*; *Customer_ID* vs *Cust#*.
- Schema mismatches (naming conventions, formats).

- **Schema integration:**

  Mismatching attribute names

  Identify real world entities from multiple data sources,

  e.g., Bill Clinton = William Clinton

  e.g., A.cust-id ≡ B.cust-#

  Integrate metadata from different sources
- **Object matching:** Mismatching structure of data

  Ex: Discount issues

  Currency type

2. Tuple Duplication:

- The use of denormalized tables is another source of data redundancy. Inconsistencies often arise between various duplicates due to inaccurate data entry.

- Sometimes the same record appears multiple times across sources.
- Example: A customer appears in both *sales* and *support* database but with a small spelling error.
- Duplicate records → data redundancy → inconsistent analytics.
- 👉 **Analogy**: It's like having two contacts for the same friend in your phone — one with email, one with phone number.

Name	DOB	Branch	Occupation	Address
A	25	HYD	Govt	TPG
B	30	TH	Govt	RJY
A	25	HYD	Private	TPG
D	30	IBP	Private	RJY

3. Detecting and resolving data value conflicts

- Data about the same entity may differ between sources.
- Causes:
  - Different **representations** (USD vs INR, GPA scale US vs China).
  - Different **aggregation levels** (monthly sales for *one store* vs *all stores*).
- 👉 **Example**: If one source says height is *180 cm* and another says *5 feet 11 inches*, you need a rule to resolve it.

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

Possible reasons: different representations, Ex: Total sales for month single store/all stores different scales, e.g., metric vs. British units (e.g., GPA in US and China)

4. Redundancy and Correlation Analysis

- Redundant data occur often when integration of multiple databases
  - ***Object identification*:** The same attribute or object may have different names in different databases
  - ***Derivable data:*** One attribute may be a "derived" attribute in another table, e.g., annual revenue
- **Redundant attributes may be able to be detected by *correlation analysis***
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

- When integrating multiple databases, redundant data often arises.
- Some attributes may be **derivable** from others.
  - Example: *Annual Revenue* vs *Monthly Revenue*.
- Correlation analysis helps detect redundant attributes.
- 👉 **Example with numbers**:
  If "Age" and "Glucose Level" show only **52.9% correlation**, we can't strongly conclude one causes the other - but redundancy can still be identified.

4. Redundancy and Correlation Analysis

Covariance is a measure of the relationship between two random variables and to what extent, they change together. Or we can say, in other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable. This is the property of a function of maintaining its form when the variables are linearly transformed. Covariance is measured in units, which are calculated by multiplying the units of the two variables.

Types of Covariance

Covariance can have both positive and negative values. Based on this, it has two types:

1. Positive Covariance
2. Negative Covariance

Positive Covariance

If the covariance for any two variables is positive, that means, both the variables move in the same direction. Here, the variables show similar behaviour. That means, if the values (greater or lesser) of one variable corresponds to the values of another variable, then they are said to be in positive covariance.

Negative Covariance

If the covariance for any two variables is negative, that means, both the variables move in the opposite direction. It is the opposite case of positive covariance, where greater values of one variable correspond to lesser values of another variable and vice-versa.

Techniques in Redundancy & Correlation Analysis

- **Correlation coefficient (r)** → ranges from -1 to +1.
  - +1 = strong positive relationship
  - -1 = strong negative relationship
  - 0 = no relationship
- **Chi-square test** → used for categorical attributes.
- **Example**:
  Survey 10 students: “Do you like Sci-Fi?” vs “Do you play chess?”
  Then show how chi-square can detect correlation.

Population Covariance Formula

$$Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Sample Covariance

$$Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - y)}{N - 1}$$

Where,

- xi = data value of x
- yi = data value of y
- $\bar{x}$ = mean of x
- $\bar{y}$ = mean of y
- N = number of data values.

If cov(X, Y) is greater than zero, then we can say that the covariance for any two variables is positive and both the variables move in the same direction.

If cov(X, Y) is less than zero, then we can say that the covariance for any two variables is negative and both the variables move in the opposite direction.

If cov(X, Y) is zero, then we can say that there is no relation between two variables.

Covariance Example

Below example helps in better understanding of the covariance of among two variables.

Question:

Calculate the coefficient of covariance for the following data:

X	2	8	18	20	28	30
Y	5	12	18	23	45	50

Solution:

Number of observations = 6

Mean of X = 17.67

Mean of Y = 25.5

$Cov(X, Y) = (1/6)\ [(2 - 17.67)(5 - 25.5) + (8 - 17.67)(12 - 25.5) + (18 - 17.67)(18 - 25.5) + (20 - 17.67)(23 - 25.5) + (28 - 17.67)(45 - 25.5) + (30 - 17.67)(50 - 25.5)]$

$= 157.83$

Interpretation

The covariance is **positive (157.83)** → this means **X and Y move together**.
When X (say, age, time, or investment) increases, Y also tends to increase.
The magnitude (157.83) depends on the units of X and Y, so it cannot be compared directly with other datasets.
That's why we usually convert this into **correlation** (by dividing by the product of standard deviations), which gives a value between -1 and +1.

4. Redundancy and Correlation Analysis

Correlation

Correlation analysis is a method of statistical evaluation used to study the strength of a relationship between two, numerically measured, continuous variables.

- Computed as Correlation co-efficient
- Value ranges between – (-1) to (+1)
- Positively Correlated, Negatively correlated, Not correlated
- The strength of a correlation indicates how strong the relationship is between the two variables. The strength is determined by the numerical value of the correlation

The Formula for Correlation Is

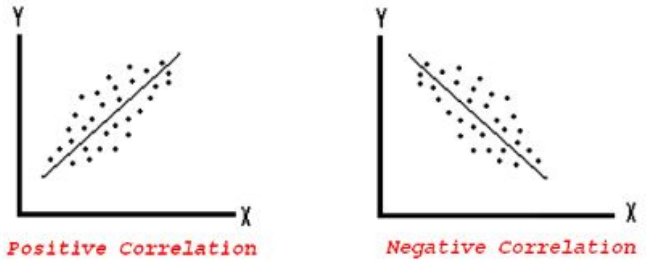
$$r = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum(X - \overline{X})^2}\sqrt{(Y - \overline{Y})^2}}$$

where:

r = the correlation coefficient

$\overline{X}$ = the average of observations of variable X

$\overline{Y}$ = the average of observations of variable Y



The Pearson correlation coefficient is denoted by the letter "r". The formula for Pearson correlation coefficient r is given by:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Where,

r = Pearson correlation coefficient

x = Values in the first set of data

y = Values in the second set of data

n = Total number of values.

Types of Correlation

The scatter plot explains the correlation between the two attributes or variables. It represents how closely the two variables are connected. There can be three such situations to see the relation between the two variables –

- Positive Correlation – when the values of the two variables move in the same direction so that an increase/decrease in the value of one variable is followed by an increase/decrease in the value of the other variable.
- Negative Correlation – when the values of the two variables move in the opposite direction so that an increase/decrease in the value of one variable is followed by decrease/increase in the value of the other variable.
- No Correlation – when there is no linear dependence or no relation between the two variables.

Solved Example

Question: Marks obtained by 5 students in algebra and trigonometry as given below:

Algebra	15	16	12	10	8
Trigonometry	18	11	10	20	17

Calculate the Pearson correlation coefficient.

Solution:

Construct the following table:

x	y	x^2	y^2	xy
15	18	225	324	270
16	11	256	121	176
12	10	144	100	120
10	20	100	400	200
8	17	64	289	136
$\sum x$ = 61	$\sum y$ = 76	$\sum x^2$ = 789	$\sum y^2$ = 1234	$\sum xy$ = 902

Formula for Pearson correlation coefficient is given by:

r =

$$\frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}}$$

r =

$$\frac{5\times902-61\times76}{\sqrt{[5\times789-(61)^2][5\times1234-(76)^2]}}$$

r = -0.424

Interpretation

The result **r = -0.424** means:

There is a **moderate negative linear relationship** between X and Y.

As **X increases, Y tends to decrease**, but not perfectly.

The strength is **not very strong** (since -0.424 is closer to 0 than -1).

Example problem:

Business problem: The healthcare industry want to develop a medication to control glucose levels. For this it want to study does age have impact on raise in glucose levels

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	X^2	Y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

From our table:
$\Sigma x = 247$
$\Sigma y = 486$
$\Sigma xy = 20,485$
$\Sigma x^2 = 11,409$
$\Sigma y^2 = 40,022$
n is the sample size=6

The correlation coefficient =
$6(20,485) - (247 \times 486) / [\sqrt{[[6(11,409) - (247^2)] \times [6(40,022) - 486^2]]}]$. = 0.5298 (strength and direction)
The range of the correlation coefficient is from -1 to 1.
here result is 0.5298 or 52.98%, which means the variables have a moderate positive correlation(some what more). We cant infer with 52.98% correlation that age has impact on rise in glucose levels. **We need more data to analyze**

14 b. Calculate and analyze correlation coefficient

$$r = \frac{\Sigma XY}{\sqrt{\Sigma x^2 \Sigma y^2}}$$ where $X = x - \bar{x}$, $Y = y - \bar{y}$

Let Number of study hours = x

number of sleeping hours = y.

$N = 5$

x	y	$X = x - 6$	$Y = y - 8$	XY	x^2	y^2
2	10	-4	2	-8	16	4
4	9	-2	1	-2	4	1
6	8	0	0	0	0	0
8	7	2	-1	-2	4	1
10	6	4	-2	-8	16	4
				$\Sigma XY = -20$	$\Sigma X^2 = 40$	$\Sigma Y^2 = 10$

$\bar{x} = \frac{\Sigma x}{N}$ $\bar{y} = \frac{40}{5}$

$= \frac{30}{5} = 6$ $= 8$.

$\therefore r = \frac{-20}{\sqrt{40 \times 10}} = \frac{-20}{20} = -1$

$r = -1$ means weak relation between study hours and sleeping hours.

Covariance and Correlation

Below table shows the comparison among covariance and correlation in brief.

Covariance	Correlation
It is a measure to show the extent to which given two random variables change with respect to each other.	It is a measure used to describe how strongly the given two random variables are related to each other.
It is a measure of correlation.	It is defined as the scaled form of covariance.
The value of covariance lies between -∞ and +∞.	The value of correlation lies between -1 and +1.
It indicates the direction of the linear relationship between the given two variables.	It measures the direction and strength of the linear relationship between the given two variables.

Chi Square Test

- 🔷 **Purpose**
- The **Chi-square test** is used to check whether there is a **correlation/association between two categorical (discrete) variables**.
- Example: Checking if **gender (male/female)** is related to **preference of a product (yes/no)**.

Redundancy and Correlation Analysis: Chi-square Test

- A correlation relationship between two categorical (discrete) attributes, *A* and *B*, can be discovered by a X^2 (**chi-square**) test.

Properties

The following are the important properties of the chi-square test:

- Two times the number of degrees of freedom is equal to the variance.
- The number of degree of freedom is equal to the mean distribution
- The chi-square distribution curve approaches the normal distribution when the degree of freedom increases.

Formula

The chi-squared test is done to check if there is any difference between the observed value and expected value. The formula for chi-square can be written as;

$$X^2 = \sum \frac{(\text{Observed value} - \text{Expected value})^2}{\text{Expected value}}$$

How It’s Used

- Construct a **contingency table** (rows = one categorical variable, columns = another).
- Calculate the **expected values** based on row and column totals.
- Apply the χ^2 formula.
- Compare χ^2 with a **critical value** from the chi-square distribution table (based on degrees of freedom and significance level, e.g., 0.05).
  - If $\chi^2 >$ critical value $\rightarrow$ reject null hypothesis $\rightarrow$ **there is a relationship**.
  - If $\chi^2 \leq$ critical value $\rightarrow$ fail to reject null hypothesis $\rightarrow$ **no relationship**.

Example of Categorical Data

Let us take an example of a categorical data where there is a society of 1000 residents with four neighbourhoods, P, Q, R and S. A random sample of 650 residents of the society is taken whose occupations are doctors, engineers and teachers. The null hypothesis is that each person's neighbourhood of residency is independent of the person's professional division. The data are categorised as:

Categories	P	Q	R	S	Total
Doctors	90	60	104	95	349
Engineers	30	50	51	20	151
Teachers	30	40	45	35	150
Total	150	150	200	150	650

Assume the sample living in neighbourhood P, 150, to estimate what proportion of the whole 1,000 people live in neighbourhood P. In the same way, we take 349/650 to calculate what ratio of the 1,000 are doctors. By the supposition of independence under the hypothesis, we should "expect" the number of doctors in neighbourhood P is;

150 x 349/650 ≈ 80.54

$$E = \frac{(\text{Row Total} \times \text{Column Total})}{\text{Grand Total}}$$

So by the chi-square test formula for that particular cell in the table, we get;

(Observed – Expected)^2/Expected Value = (90-80.54)^2/80.54 ≈ 1.11

Chi-Square Calculation: An Example

- Suppose that a group of 1,500 people was surveyed.
- The observed frequency (or count) of each possible joint event is summarized in the contingency table shown

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- The numbers in parentheses are the expected frequencies (calculated based on the data distribution for both attributes using Equation e_{ij}).
- Are ***like_science_fiction*** and ***play_chess*** correlated?

Chi-Square Calculation: An Example

- For example, the expected frequency for the cell (play_chess, fiction) is

$$e_{11} = \frac{count(play\_chess) * count(like\_science\_fiction)}{N} = \frac{300 * 450}{1500} = 90$$

- Notice that
  - the sum of the expected frequencies must equal the total observed frequency for that row, and
  - the sum of the expected frequencies in any column must also equal the total observed frequency for that column.

Chi-Square Calculation: An Example

- We can get X^2 by:

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$
$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.$$

- For this 2 x 2 table, the degrees of freedom are (2-1)(2-1) = 1. ((no. of rows-1)*(no.of columns-1))
- For 1 degree of freedom, the X^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 X^2 distribution, typically available from any textbook on statistics).
- Since our computed value is above this, we can reject the hypothesis that ***play chess*** and ***preferred reading*** are independent and conclude that the two attributes are (strongly) correlated for the given group of people.

III- Data Reduction

Data Reduction strategies:

- **Data reduction**: Obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.
- **Data reduction strategies**

1.Data cube aggregation

3. Dimensionality reduction-remove unimportant attributes/variables

Eliminate the redundant attributes: which are weekly important across the data.

- Wavelet transforms/ Data compression
- Principal Components Analysis (PCA)
- Feature subset selection, feature creation

4. Numerosity reduction- replace original data volume by smaller forms of data

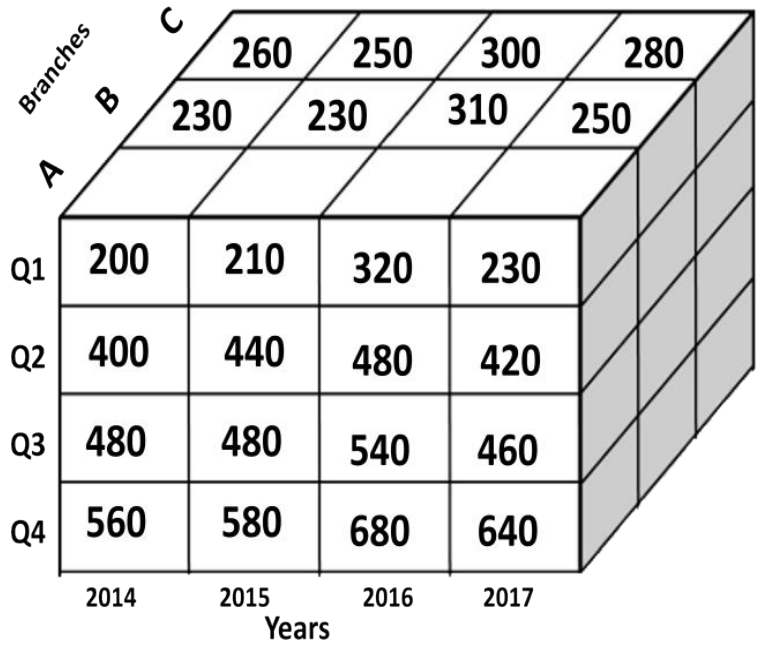
- Regression and Log-Linear Models
- Histograms, clustering, sampling
- Data cube aggregation

1. Data cube aggregation

For example, the data consists of All Electronics sales per quarter for the years 2014 to 2017.You are, however, interested in the annual sales, rather than the total per quarter. Thus, the data can be ***aggregated*** so that the resulting data summarize the total sales per year instead of per quarter

Year/Quarter	2014	2015	2016	2017
Quarter 1	200	210	320	230
Quarter 2	400	440	480	420
Quarter 3	480	480	540	460
Quarter 4	560	580	680	640

Year	Sales
2014	1640
2015	1710
2016	2020
2017	1750



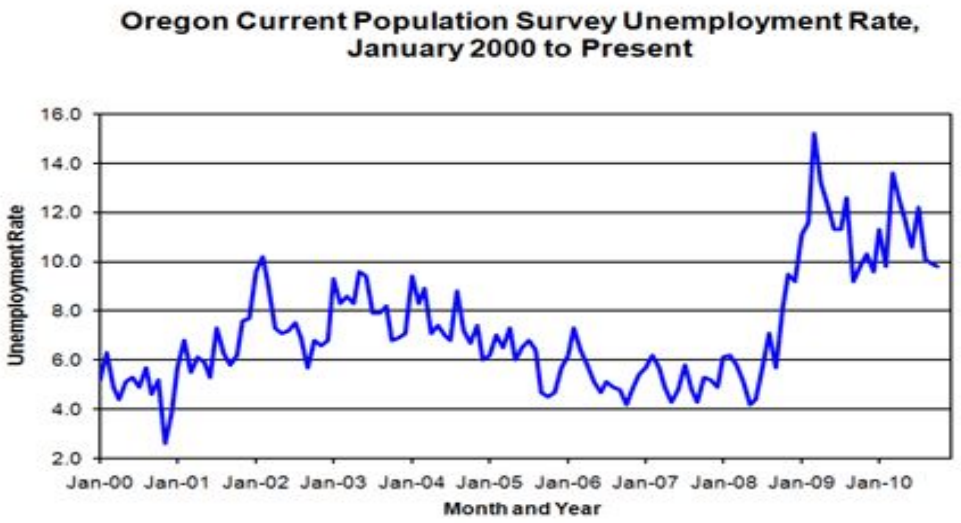
2. Dimensionality Reduction

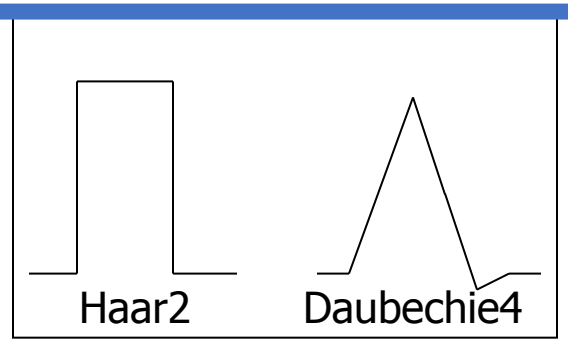
- **Know about Curse of dimensionality**
  - When dimensionality increases, data becomes increasingly sparse
  - Density and distance between points, which is critical to clustering, outlier analysis, becomes less meaningful
  - The possible combinations of subspaces will grow exponentially
- **What is Dimensionality reduction**
  - Method of eliminating irrelevant features so as to reduce noise
  - Is proposed to avoid the curse of dimensionality
  - Reduce time and space required in data mining
  - Allow easier visualization of data (quite messy to visualize huge data)
- **Dimensionality reduction techniques**
  - Wavelet transforms
  - Principal Component Analysis
  - Supervised and nonlinear techniques (e.g., feature selection)

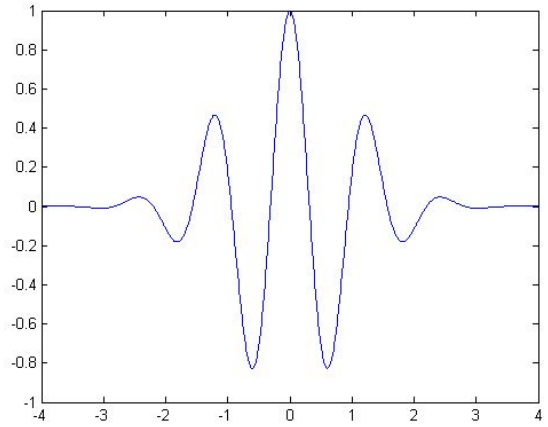
Dimensionality Reduction: Wavelet Transform

- The discrete wavelet transform (DWT) is a linear signal processing technique that, when applied to a data vector, transforms it to a numerically different vector, of wavelet coefficients.
- The two vectors are of the same length. When applying this technique to data reduction, we consider each tuple as an n-dimensional data vector, that is, $X=(x_1,x_2,\ldots\ldots\ldots\ldots,x_n)$, depicting n measurements made on the tuple from n database attributes .
- **For example**, all wavelet coefficients larger than some user-specified threshold can be retained. All other coefficients are set to 0. The resulting data representation is therefore very sparse, so that can take advantage of data sparsity are computationally very fast if performed in wavelet space.
- Given a set of coefficients, an approximation of the original data can be constructed by applying the inverse of the DWT used.

Dimensionality Reduction: Sequence Data and Wavelet Function







A continuous wavelet function

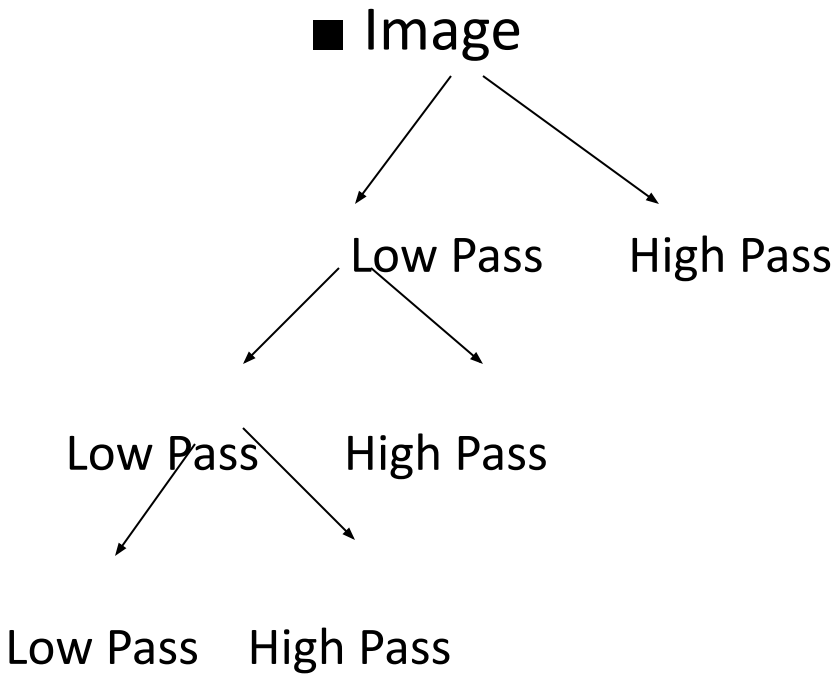
Dimensionality Reduction: Wavelet Transformation

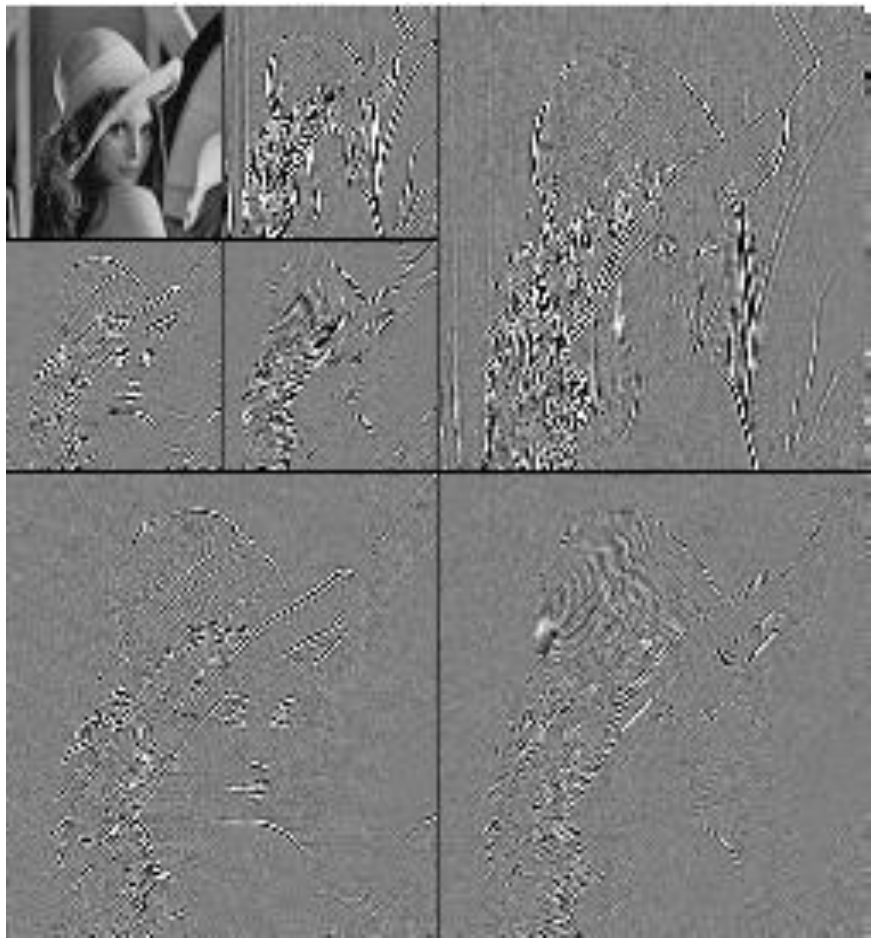
- Discrete wavelet transform (DWT): linear signal processing, multi-resolutional analysis
- Similar to discrete Fourier transform (DFT), but better lossy compression, localized in space
- Method:
  - Length, L, must be an integer power of 2 (padding with 0's, when necessary)
  - Each transform has 2 functions: sum, difference
  - Applies to pairs of data, resulting in two set of data of length L/2
  - Applies two functions recursively, until reaches the desired length
- Compressed approximation: store only a small fraction of the strongest of the wavelet coefficients

An example

			0		1		2		3		4		3		2		1	
1		1		-1		-1					1		5		7		3	
0	0				2		-2				4		-4		6		10	
0		**0**		**-4**		**0**					**-8**		**0**				**4**	**16**

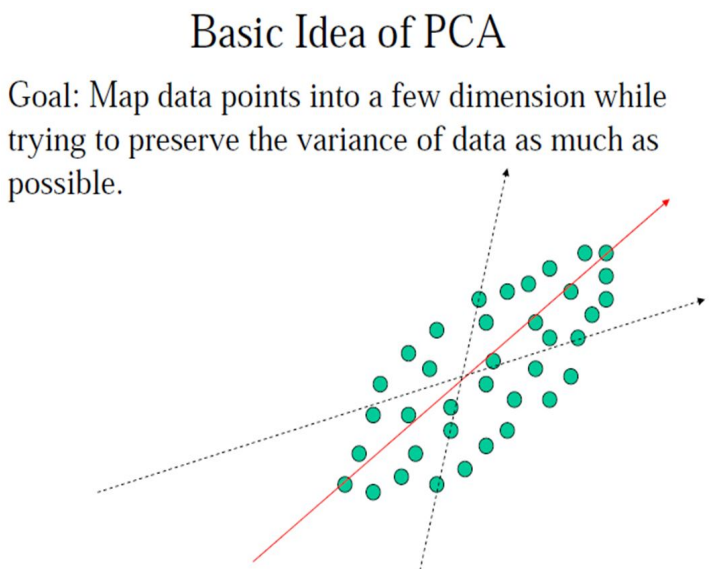
DWT for Image Compression





Dimensionality Reduction: Principal Component Analysis (PCA)

- Given *N* data vectors from *d*-dimensions, find $k \leq d$ orthogonal vectors (*principal components*) that can be best used to represent data
- Steps
  - Normalize input data
  - Compute *k* orthonormal (unit) vectors, i.e., *principal components*
  - 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)
- Used when the number of dimensions is large



Dimensionality Reduction: PCA Method

- Given a data matrix X (n × d, n data points, d dimension).
- Normalize X by subtracting mean from each data point
- Construct a covariance matrix $C=X^TX / (n-1)$ (d × d)
- Calculate the eigenvectors and eigenvalues of the covariance matrix C. ($C\ v = v\ \lambda$).
- Sort eigenvectors by eigenvalues in decreasing order
- Map data point *x* to the direction *v* by computing the dot product.
- A well studied problem. Implementation in many software such as MatLab.

Dimensionality Reduction: Attribute Subset Selection

- Attribute subset selection reduces the data set size by removing irrelevant or redundant attributes (or dimensions). The goal of attribute subset selection is to find a minimum set of attributes. It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand.

•*To find out a 'good' subset from the original attributes*

- Another way to reduce dimensionality of data
- **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 often irrelevant to the task of predicting students' GPA

Dimensionality Reduction: Heuristic Search in Attribute Selection

- There are 2^d possible attribute combinations of d attributes
- Typical heuristic attribute selection methods:
  - Best single attribute under the attribute independence assumption: choose by significance tests
  - **Best step-wise(forward) feature selection**:
    - The best single-attribute is picked first
    - Then next best attribute condition to the first, ...
  - **Step-wise attribute(backward) elimination**:
    - Repeatedly eliminate the worst attribute
  - **Best combined attribute selection and elimination**
  - **Decision tree induction**
    - Use attribute elimination and backtracking

Heuristic Search in Attribute Selection

Forward selection

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:

$\{A_1, A_4, A_6\}$

backward selection

Initial attribute set:

=> $\{A_1, A_2, A_3, A_4, A_5, A_6\}$

=> $\{A_1, A_4, A_5, A_6\}$

=> Reduced attribute

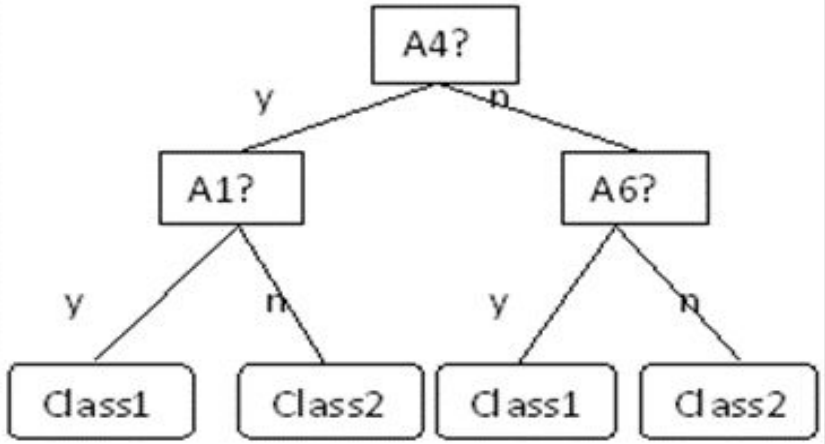
set

$\{A_1, A_4, A_6\}$

decision tree introduction

Initial attribute set:

$\{A_1, A_2, A_3, A_4, A_5, A_6\}$



Reduced attribute set

$\{A_1, A_4, A_6\}$

Dimensionality Reduction: Attribute Creation (Feature Generation)

- Create new attributes (features) that can capture the important information in a data set more effectively than the original ones
- Three general methodologies
  - Attribute extraction
    - Domain-specific
  - Mapping data to new space (see: data reduction)
    - E.g., Fourier transformation, wavelet transformation, manifold approaches (not covered)
  - Attribute construction
    - Combining features (see: discriminative frequent patterns in Chapter 7)
    - Data discretization

3. Numerosity Reduction

- Reduce data volume by choosing alternative, *smaller forms* of data representation
- **Parametric methods**
  - **Regression:**

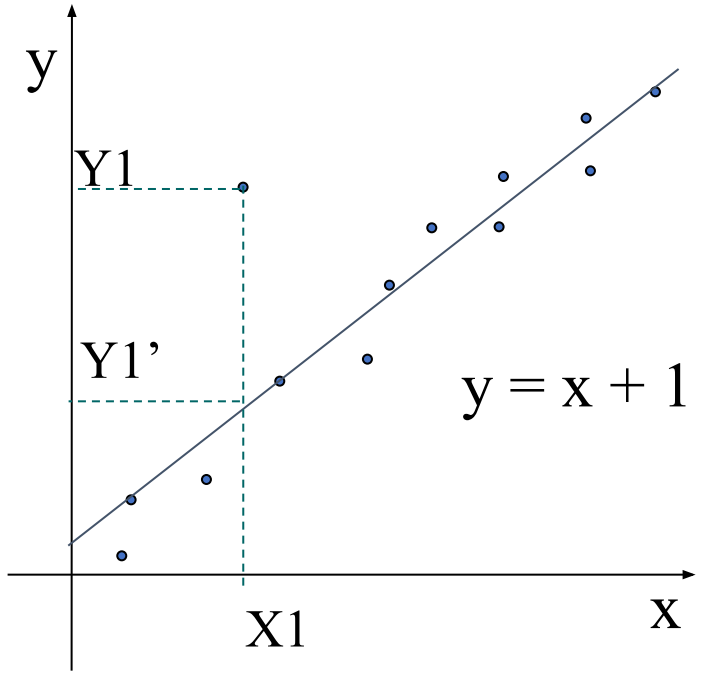
    Simple linear regression

    Multiple linear regression

    Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  - **Log-linear models**—obtain value at a point in *m*-D space as the product on appropriate marginal subspaces
- **Non-parametric methods**
  - Do not assume models
  - Major families: **Histograms, Clustering, Sampling, Cube aggregation**

Parametric methods: Regression Analysis

- **Regression analysis:** A collective name for techniques for the modeling and analysis of numerical data consisting of values of a ***dependent variable*** (also called ***response variable*** or *measurement*) and of one or more *independent variables* (aka. ***explanatory variables*** or ***predictors***)
- The parameters are estimated so as 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



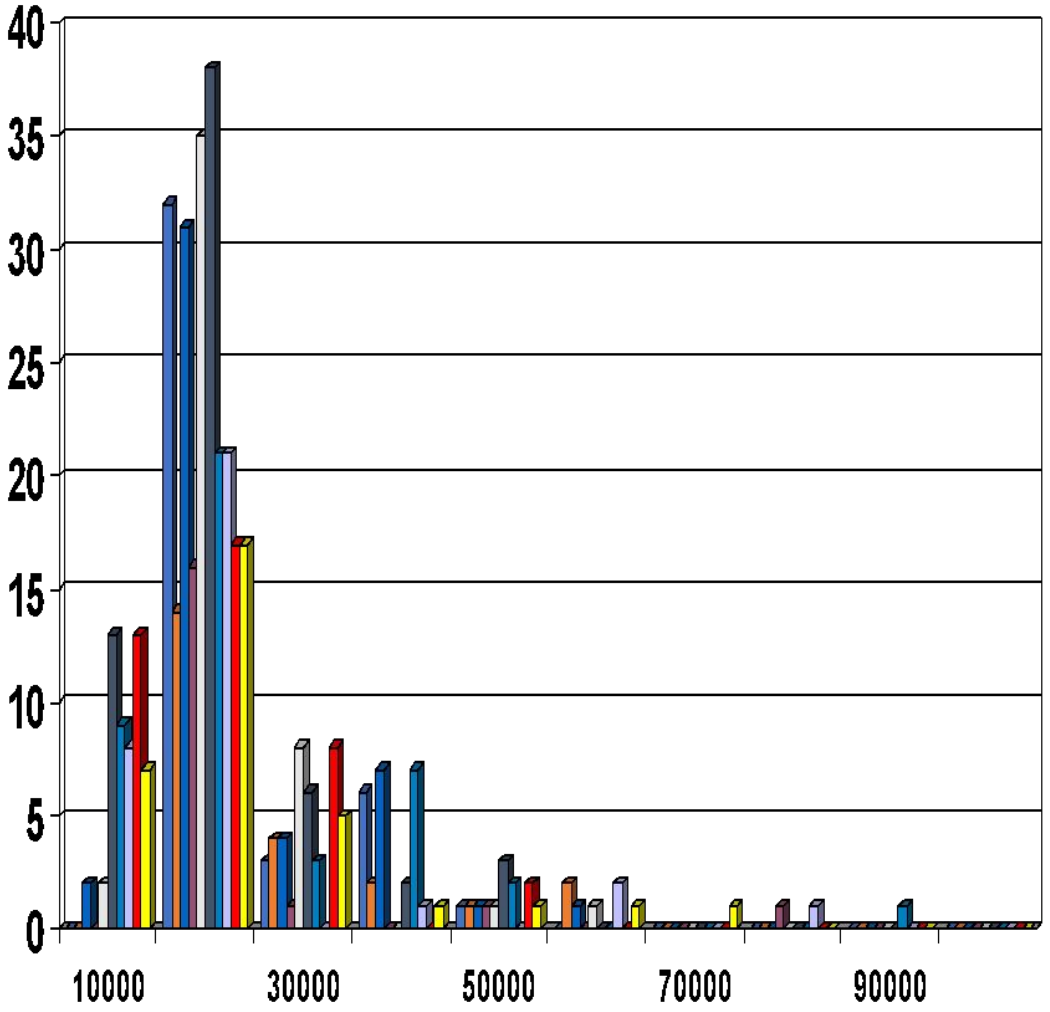
- Used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships

Parametric methods : Regress Analysis and Log-Linear Models

- **Linear regression**: $Y = w X + b$
  - Two regression coefficients, *w* and *b,* specify the line and are to be estimated by using the data at hand
  - Using the least squares criterion to the known values of $Y_1, Y_2, \ldots, X_1, X_2, \ldots$
- **Multiple regression**: $Y = b_0 + b_1 X_1 + b_2 X_2$
  - Many nonlinear functions can be transformed into the above
- **Log-linear models:**
  - Approximate discrete multidimensional probability distributions
  - Estimate the probability of each point (tuple) in a multi-dimensional space for a set of discretized attributes, based on a smaller subset of dimensional combinations
  - Useful for dimensionality reduction and data smoothing

Nonparametric methods : Histogram Analysis

- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
  - Equal-width: equal bucket range
  - Equal-frequency (or equal-depth)



Nonparametric methods : Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can be very effective if data is clustered but not if data is “smeared”
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms

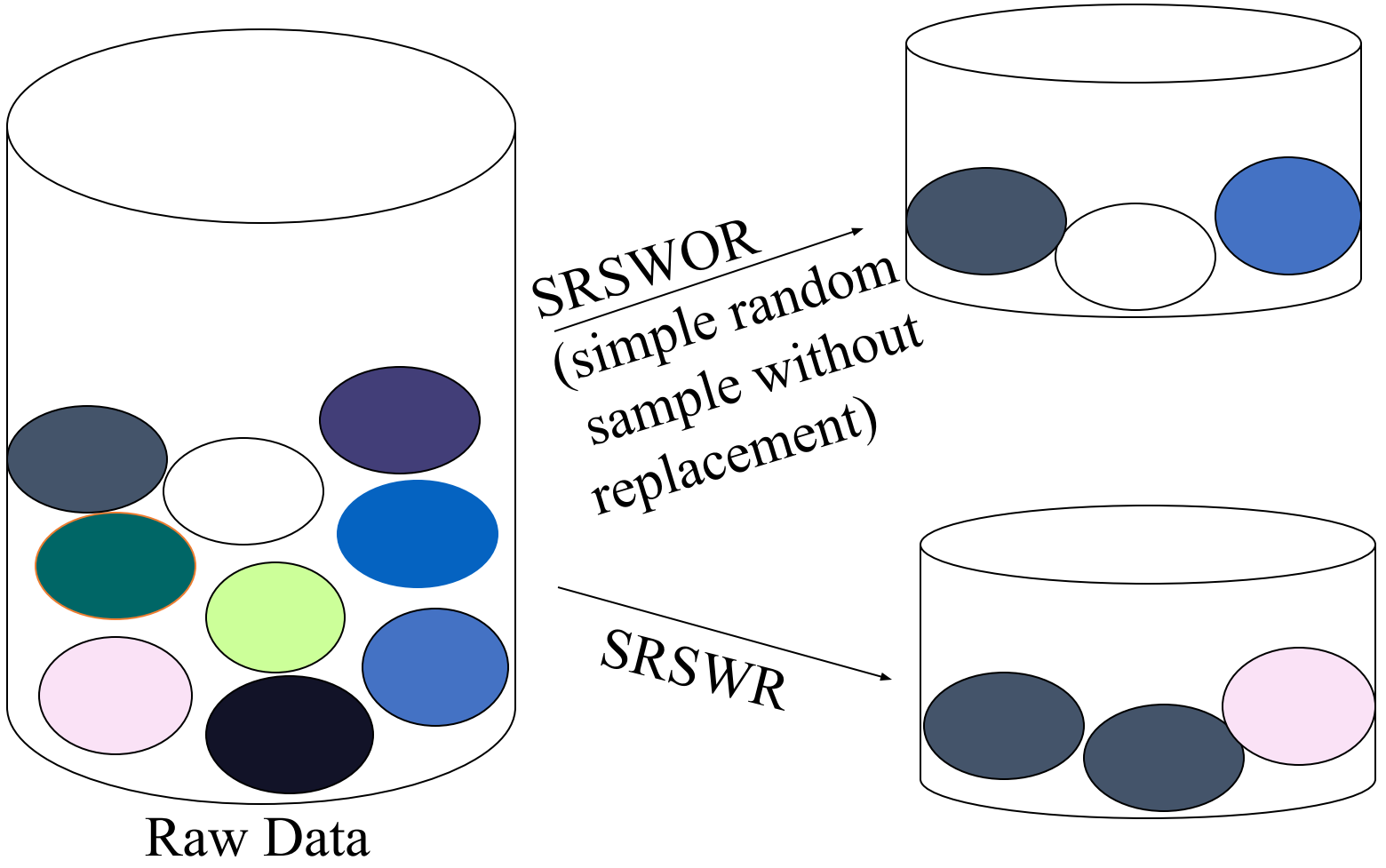
Nonparametric methods : Sampling

- **Sampling**: obtaining a small sample *s* to represent the whole data set *N*
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- **Key principle**: Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
  - Develop adaptive sampling methods, e.g., stratified sampling:
- Note: Sampling may not reduce database I/Os (page at a time)

Nonparametric methods : Types of Sampling

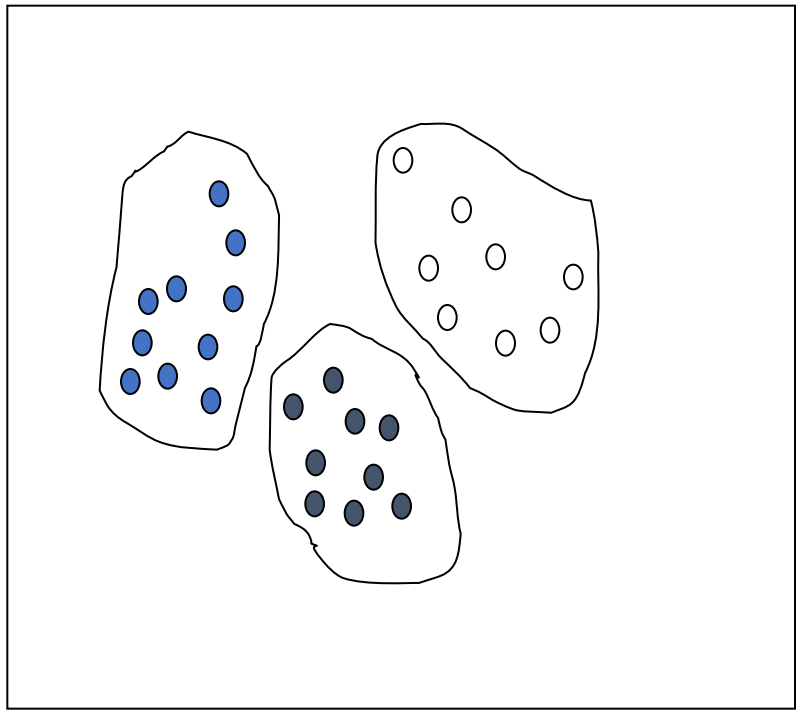
- **Simple random sampling**
  - There is an equal probability of selecting any particular item
- **Sampling without replacement**
  - Once an object is selected, it is removed from the population
- **Sampling with replacement**
  - A selected object is not removed from the population
- **Stratified sampling:**
  - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
  - Used in conjunction with skewed data

Nonparametric methods : Sampling: With or without Replacement

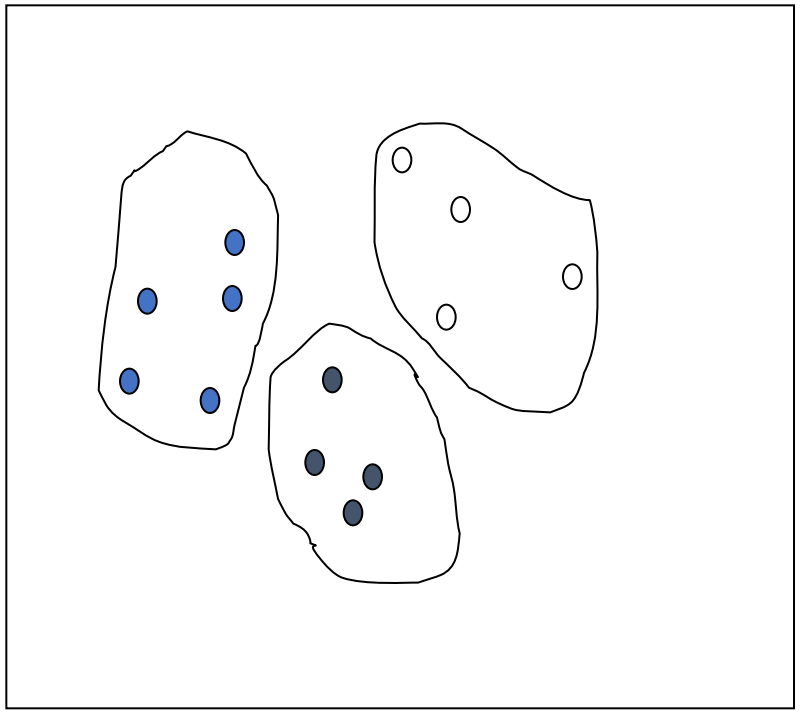


Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample



Nonparametric methods : Data Cube Aggregation

- Data cube aggregation, where aggregation operations are applied to the data for construction of a data cube.
- Data cubes store multidimensional aggregated information. Each cell holds an aggregate data value, corresponding to the data point in multidimensional space.
- Concept hierarchies may exist for each attribute, allowing the analysis of data at multiple level of abstraction.
- Data cubes provide fast access to pre computed summarized data, thereby benefiting on-line analytical processing as well as data mining.

Nonparametric methods : Data Cube Aggregation

The cube can be created in three ways:

- **Based cuboid-** The cube created at the lowest level of abstraction is referred to as base cuboid.
- **Lattice of cuboids-** Data cubes created for varying levels of abstraction are often referred to as cuboids.
- **Apex cuboid** - A cube at highest level of abstraction is the apex cuboid

Year 2004

Quarter	sales

Year 2003

Quarter	sales

Year 2002

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

Year	sales
2002	$1,568,000
2003	$2,356,000
2004	$3,594,000

Fig .2.3(a) on the left, the sales are shown per quarter. On the right ,the data are aggregated to provide the annual sales.

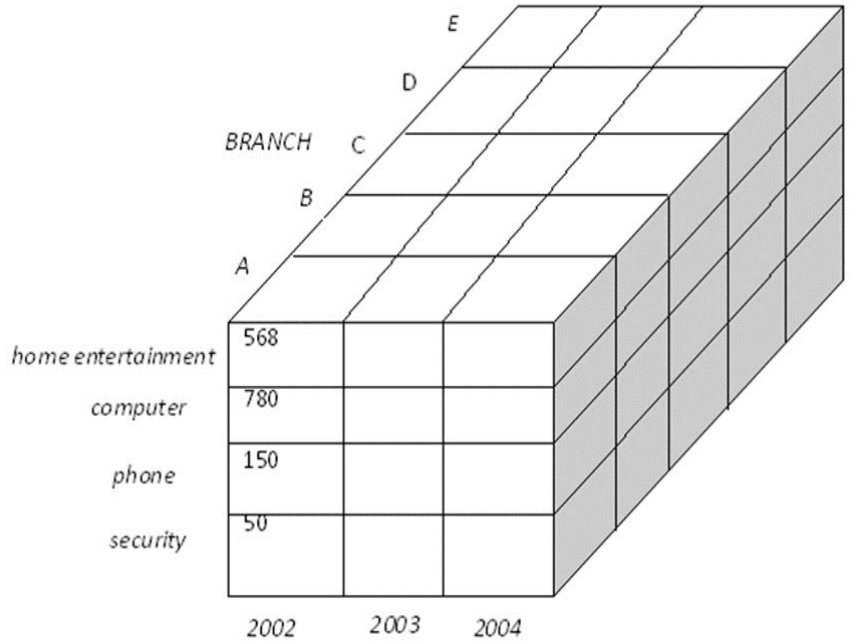


Fig 2.3(b) A data cube for sales

IV- Data Transformation

- **Data transformation**

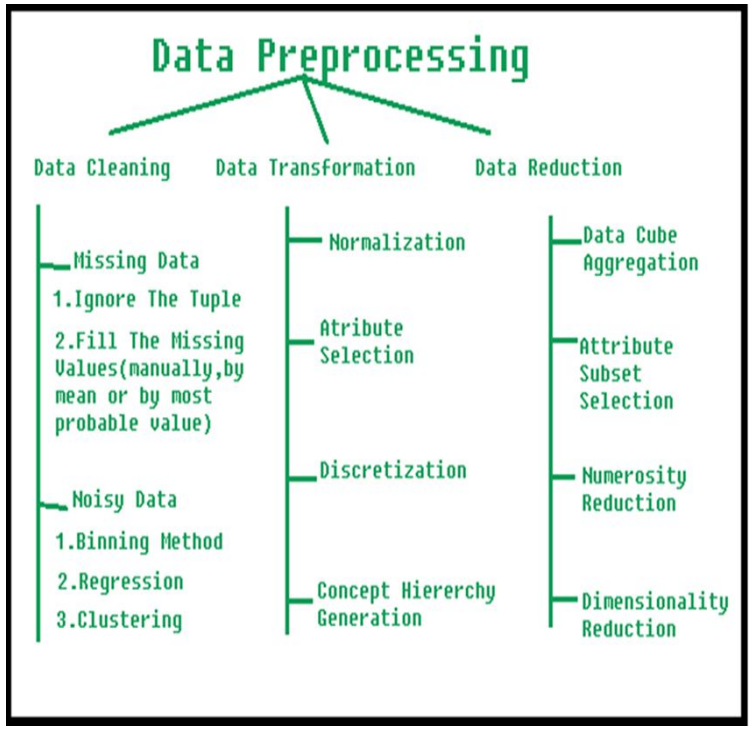
  The data are transformed into forms appropriate for mining.

- **Data transformation tasks:**

  1. **Smoothing**: Remove the noise from the data. Techniques includes Binning, Regression, Clustering.
  2. **Normalization**
  3. **Attribute construction**, Subset selection
  4. **Aggregation**
  5. **Discretization**
  6. **Generalization**

Data Transformation Tasks

- **Normalization**
  - the attribute data are scaled so as to fall within a small specified range, such as -1.0 to 1.0, 0.0 to 1.0
- **Attribute construction** (or feature construction)
  - new attributes are constructed and added from the given set of attributes to help the mining process.
- **Aggregation**
  - summary or aggregation operations are applied to the data.
  - For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts.



Data Transformation Tasks

- **Discretization**
  - Dividing the range of a continuous attribute into intervals
  - **For example,** values for numerical attributes, like **age**, may be mapped to higher-level concepts, like **youth**, **middle- aged**, and **senior**.
- **Generalization**
  - where low-level or "primitive" (raw) data are replaced by higher-level concepts through the use of concept hierarchies.
  - **For example**, categorical attributes, like **street**, can be generalized to higher-level concepts, like **city** or **country**.

2. Normalization

- An attribute is normalized by scaling its values so that they fall within a small specified range, such as 0.0 to 1.0.
- Normalization is particularly useful for classification algorithms involving
  - neural networks
  - distance measurements such as nearest-neighbor classification and clustering.
- If using the neural network backpropagation algorithm for classification mining, normalizing the input values for each attribute measured in the training instances will help speed up the learning phase.
- For distance-based methods, normalization helps prevent attributes with initially large ranges (e.g., ***income***) from out-weighing attributes with initially smaller ranges (e.g., binary attributes).

- **Normalization methods**
  I. Min-max normalization
  II. z-score normalization

Min-max Normalization

- Min-max normalization
  - performs a linear transformation on the original data.
- Suppose that:
  - $\min_A$ and $\max_A$ are the minimum and maximum values of an attribute, A.
- Min-max normalization maps a value, v, of A to v′ in the range [new_min_A, new_max_A] by computing:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

Normalization

- **Min-max normalization:** to [new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

- **Z-score normalization** (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

- **Normalization by decimal scaling**

$$v' = \frac{v}{10^j} \quad \text{Where } j \text{ is the smallest integer such that Max}(|v'|) < 1$$

3. Data Aggregation

- On the left, the sales are shown per quarter. On the right, the data are aggregated to provide the annual sales
- Sales data for a given branch of *AllElectronics* for the years 2002 to 2004.

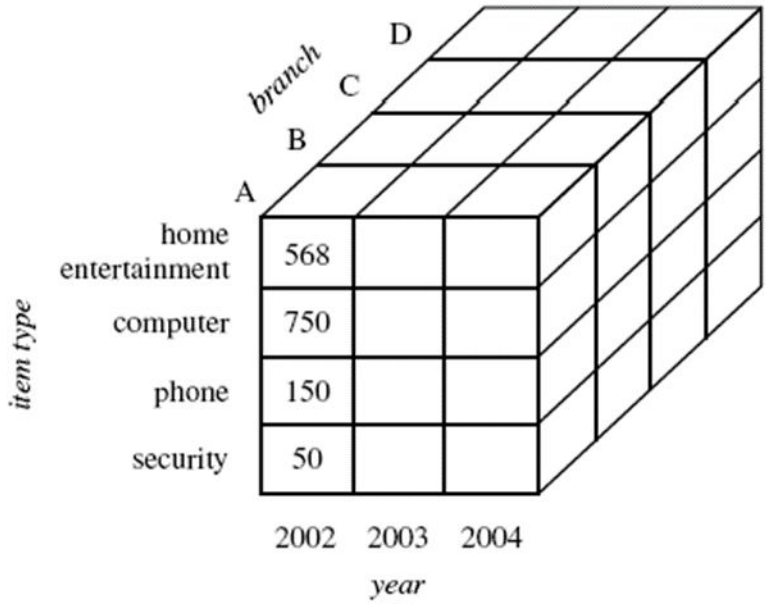
Year 2004

Year 2003

Year 2002

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

Year	Sales
2002	$1,568,000
2003	$2,356,000
2004	$3,594,000



- Data cubes store multidimensional aggregated information.
- Data cubes provide fast access to precomputed, summarized data, thereby benefiting on-line analytical processing as well as data mining.
- A data cube for sales at *AllElectronics*.

4. Attribute Construction

- Attribute construction (feature construction)
  - new attributes are constructed from the given attributes and added in order to help improve the accuracy and understanding of structure in high-dimensional data.
- Example
  - we may wish to add the attribute ***area*** based on the attributes ***height*** and ***width***.
- By attribute construction can discover missing information.
- Why attribute subset selection
  - Data sets for analysis may contain hundreds of attributes, many of which may be irrelevant to the mining task or redundant.
- For example,
  - if the task is to classify customers as to whether or not they are likely to purchase a popular new CD at *AllElectronics* when notified of a sale, attributes such as the **customer's telephone number** are likely to be irrelevant, unlike attributes such as ***age*** or ***music_taste***.

Attribute Subset Selection

- Using domain expert to pick out some of the useful attributes
  - Sometimes this can be a difficult and time-consuming task, especially when the behavior of the data is not well known.
- Leaving out relevant attributes or keeping irrelevant attributes result in discovered patterns of poor quality.
- In addition, the added volume of irrelevant or redundant attributes can **slow down** the mining process.

- Attribute subset selection (feature selection):
  - Reduce the data set size by removing irrelevant or redundant attributes

- **Goal**:
  - select a minimum set of features (attributes) such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
  - It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand.

Attribute Subset Selection

- 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 prohibitively expensive, especially as n increase.
  - Heuristic methods that explore a reduced search space are commonly used for attribute subset selection.
  - These methods are typically greedy in that, while searching through attribute space, they always make what looks to be the best choice at the time.
  - Such greedy methods are effective in practice and may come close to estimating an optimal solution.
- Heuristic methods:
  1. Step-wise forward selection
  2. Step-wise backward elimination
  3. Combining forward selection and backward elimination
  4. Decision-tree induction
- The "best" (and "worst") attributes are typically determined using:
  - the tests of ***statistical significance***, which assume that the attributes are independent of one another.
  - the ***information gain*** measure used in building decision trees for classification.

Attribute Subset Selection

- 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:
$\{A_1, A_4, A_6\}$

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

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

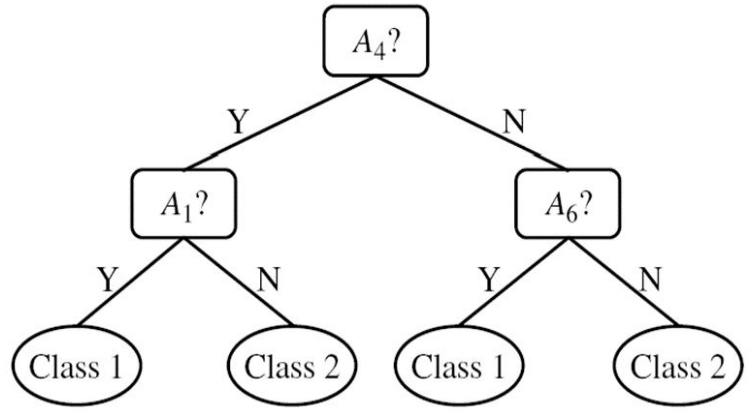
=> $\{A_1, A_3, A_4, A_5, A_6\}$
=> $\{A_1, A_4, A_5, A_6\}$
=> Reduced attribute set:
$\{A_1, A_4, A_6\}$

- 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.

Attribute Subset Selection

- Decision tree induction:
  - Decision tree algorithms, such as ID3, C4.5, and CART, were originally intended for classification.
  - Decision tree induction constructs a flowchart-like structure where each internal (nonleaf) node denotes a test on an attribute, each branch corresponds to an outcome of the test, and each external (leaf) node denotes a class prediction.
  - At each node, the algorithm chooses the "best" attribute to partition the data into individual classes.
  - When decision tree induction is used for attribute subset selection, a tree is constructed from the given data.
  - All attributes that do not appear in the tree are assumed to be irrelevant.

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



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

5. Generalization

- Generalization is the generation of concept hierarchies for categorical data
- Categorical attributes have a finite (but possibly large) number of distinct values, with no ordering among the values.
- Examples include
  - geographic location,
  - job category, and
  - itemtype.

- A relational database or a dimension location of a data warehouse may contain the following group of attributes: street, city, province or state, and country.
- A user or expert can easily define a concept hierarchy by specifying ordering of the attributes at the schema level.
- A hierarchy can be defined by specifying the total ordering among these attributes at the schema level, such as:
  - **street < city < province or state < country**

6. Discretization

- **Three types of attributes:**
  - Nominal — values from an unordered set, e.g., color, profession
  - Ordinal — values from an ordered set, e.g., military or academic rank
  - Continuous — real numbers, e.g., integer or real numbers
- **Discretization:**
  - Divide the range of a continuous attribute into intervals
  - Some classification algorithms only accept categorical attributes.
  - Reduce data size by discretization

Discretization and Concept Hierarchy

- **Discretization**
  - Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
- **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 young, middle-aged, or senior)

Discretization and Concept Hierarchy Generation for Numeric Data

- **Typical methods:** All the methods can be applied recursively
  - **Binning**
    - Top-down split, unsupervised,
  - **Histogram analysis**
    - Top-down split, unsupervised
  - **Clustering analysis**
    - Either top-down split or bottom-up merge, unsupervised
  - **Entropy-based discretization**: supervised, top-down split
  - Interval merging by χ^2 Analysis: supervised, bottom-up merge

Entropy-Based Discretization

- Entropy is calculated based on class distribution of the samples in the set. Given *m* classes, the entropy of S_1 is

$$Entropy(S_1) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

  - where p_i is the probability of class *i* in S_1

- Given a set of samples S, if S is partitioned into two intervals S_1 and S_2 using boundary T, the information gain after partitioning is

$$I(S,T) = Entropy(S) - (\frac{|S_1|}{|S|} Entropy(S_1) + \frac{|S_2|}{|S|} Entropy(S_2))$$

- The boundary that minimizes the entropy function over all possible boundaries (i.e. maximize information is selected as a binary discretization)
- The process is recursively applied to partitions obtained until some stopping criterion is met
- Such a boundary may reduce data size and improve classification accuracy

Interval Merge by χ^2 Analysis

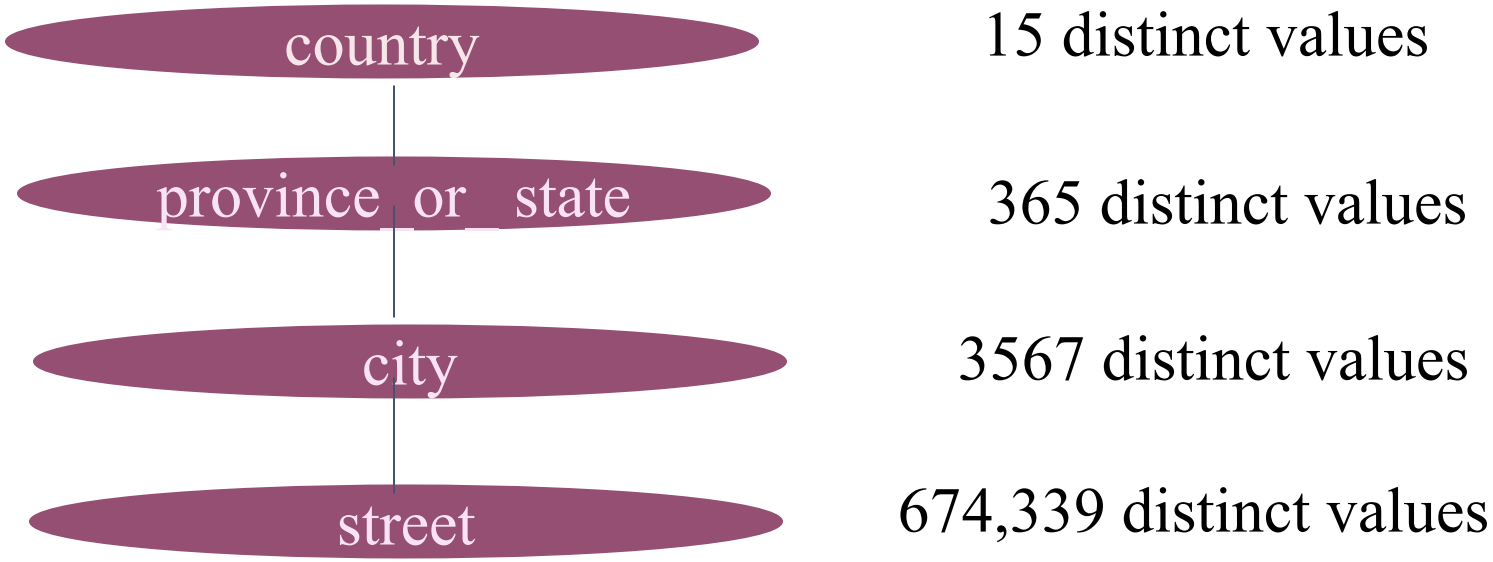
- Merging-based (bottom-up) vs. splitting-based methods
- Merge: Find the best neighboring intervals and merge them to form larger intervals recursively
- **ChiMerge [Kerber AAAI 1992, See also Liu et al. DMKD 2002]**
  - Initially, each distinct value of a numerical attr. A is considered to be one interval
  - χ^2 tests are performed for every pair of adjacent intervals
  - Adjacent intervals with the least χ^2 values are merged together, since low χ^2 values for a pair indicate similar class distributions
  - This merge process proceeds recursively until a predefined stopping criterion is met (such as significance level, max-interval, max inconsistency, etc.)

Concept Hierarchy Generation for Categorical Data

- Specification of a partial/total ordering of attributes explicitly at the schema level by users or experts
  - street < city < state < country
- Specification of a hierarchy for a set of values by explicit data grouping
  - {Urbana, Champaign, Chicago} < Illinois
- Automatic generation of hierarchies (or attribute levels) by the analysis of the number of distinct values
  - E.g., for a set of attributes: {street, city, state, country}

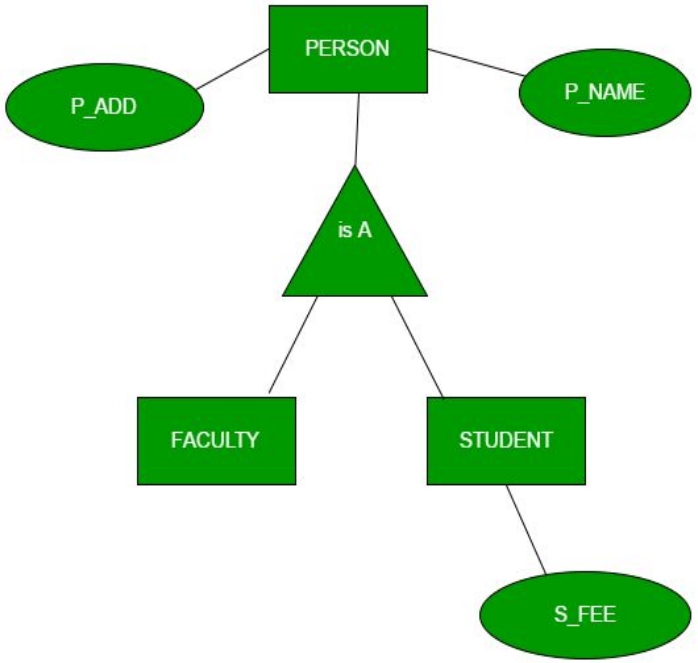
Automatic Concept Hierarchy Generation

- Some hierarchies can be automatically generated based on the analysis of the number of distinct values per attribute in the data set
  - The attribute with the most distinct values is placed at the lowest level of the hierarchy
  - Exceptions, e.g., weekday, month, quarter, year

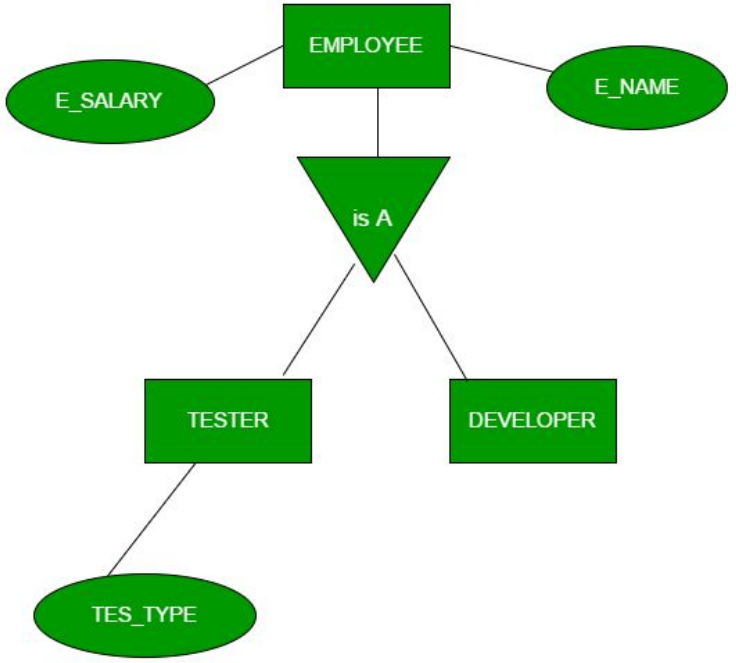


Generalization , specialization, Aggregation

- Generalization is the process of extracting common properties from a set of entities and create a generalized entity from it

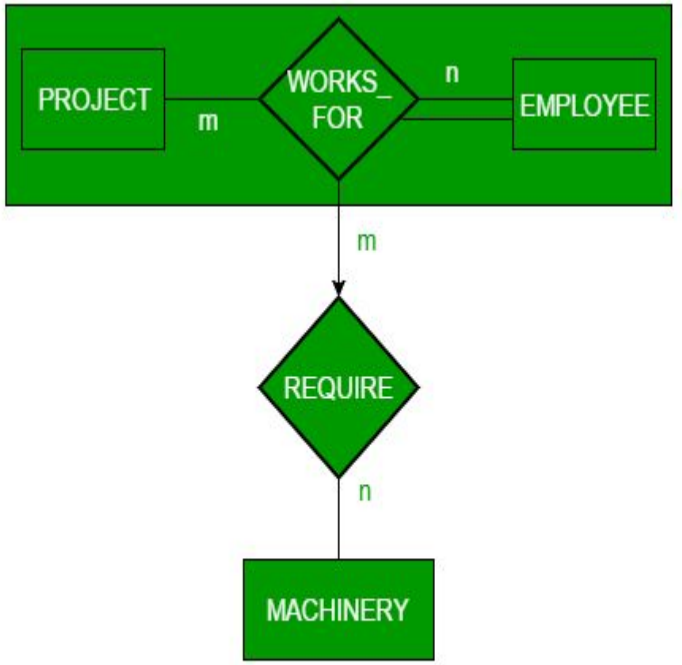


- In specialization, an entity is divided into sub-entities based on their characteristics. It is a top-down approach where higher level entity is specialized into two or more lower level entities



Specialization

- Aggregation is an abstraction through which we can represent relationships as higher level entity sets.



Aggregation