

Data Mining and Web Algorithms

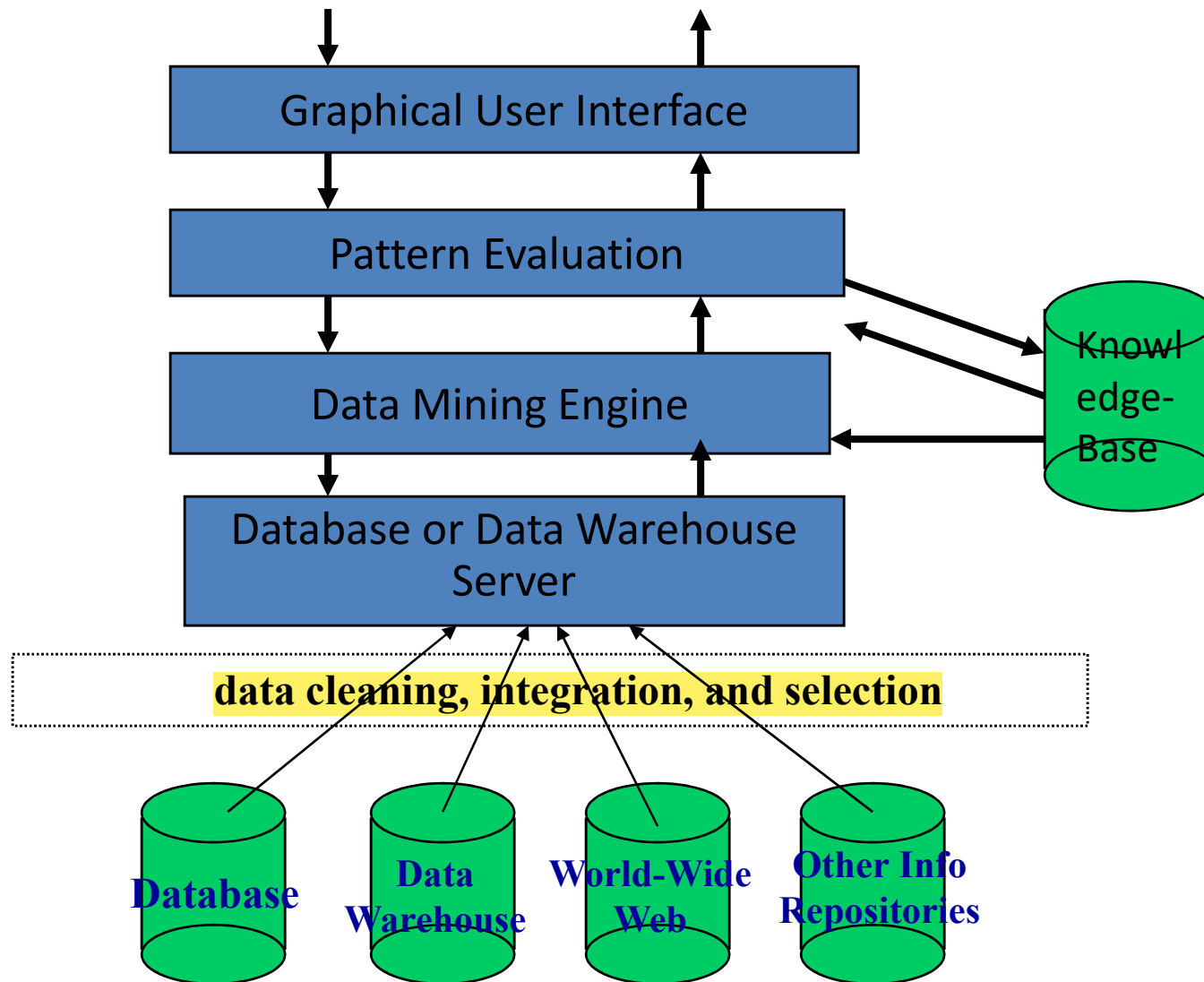
Course Code: 15B22CI621

Credits: 4 [3+ 1]


Where we are....

1. Introduction to Data Mining and Web Algorithms
2. Architecture of Data Mining System
3. Applications of Data Mining
4. Data set
5. Different type of attributes
6. Different Type of Data for DataMining
- 7. Revisit of Data Mining Engine :Data Preprocessing**

Data Preprocessing : data cleaning, integration, and selection



Data Preprocessing

- Why preprocess the data! 
- Data cleaning
- Data integration and transformation
- Data reduction
- Summary

Why Data Preprocessing?

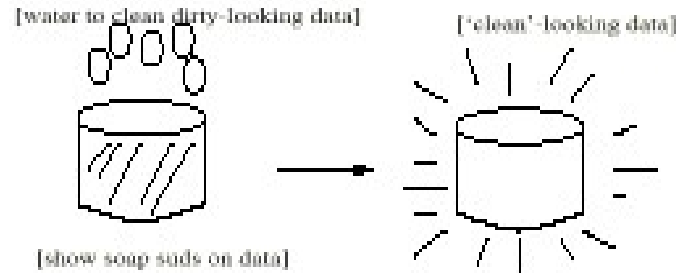
- Data in the real world is dirty
 - **incomplete**: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
 - e.g., occupation=""
 - **noisy**: containing errors or outliers
 - e.g., Salary="-10"
 - **inconsistent**: containing discrepancies in codes or names
 - e.g., Age="42" Birthday="03/07/1997"
 - e.g., Was rating "1,2,3", now rating "A, B, C"
- Data comes from Multiple sources
 - Multiple databases
 - Files , data cubes etc.
- Huge Size of Data

Why Is Data Preprocessing Important?

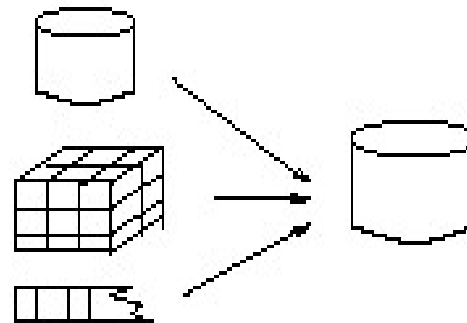
- No quality data, no quality mining results!
 - Quality decisions must be based on quality data
 - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
 - Garbage IN-> Garbage Out

Forms of Data Preprocessing

Data Cleaning



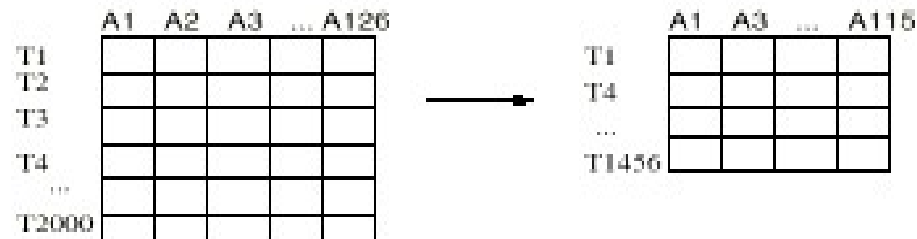
Data Integration



Data Transformation


-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



Source : [1]

Data Preprocessing

- Why preprocess the data?
- Data cleaning 
- Data integration and transformation
- Data reduction
- Summary

Why Data Cleaning?

- **Incomplete data** may come from
 - “Not applicable” data value when collected
 - Different considerations between the time when the data was collected and when it is analyzed.
 - Human/hardware/software problems
- **Noisy data** (incorrect values) may come from
 - Faulty data collection instruments
 - Human or computer error at data entry
 - Errors in data transmission
- **Inconsistent** data may come from
 - Different data sources
 - Functional dependency violation (e.g., modify some linked data)

Data Cleaning

- Data cleaning tasks
 - Fill in missing values
 - Identify outliers and smooth out noisy data
 - Correct inconsistent data

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
 - certain data may not be considered important at the time of entry
 - changes of the data
- Missing data may need to be inferred.

How to Handle Missing Data?

- **Ignore the tuple**: usually done when class label is missing (assuming the tasks in classification—**not effective** when the percentage 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 mode** or mode Median
 - 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**[**Popular**]

Mode

❖ Value that occurs **most frequently** in the data

❖ Example:

3	7	2	5	7	6	9	2	7
---	---	---	---	---	---	---	---	---

The value with the highest frequency is 7, **Mode = 7**

❖ Mode can be one or more depending on the data.

→ **Unimodal** (one mode)

→ **Bi-modal**(2 modes)/ **trimodal**(3modes)

❖ In general, two or more than two → Multimodal data set

❖ Empirical formula:

$$\text{mean} - \text{mode} = 3 \times (\text{mean} - \text{median})$$

Calculation of the Mean

- The mean is defined as the sum of the observations in the set of data divided by the total number of observations in the set of data
- 8, 9, 4, 6, 8, 2, 4, 10

$$\bar{X} = \frac{\sum_{i=1}^n x_i}{n}$$

$$\text{Mean} = \frac{8+9+4+6+8+2+4+10}{8}$$

$$= \frac{51}{8} = 6.375$$

Note: Mean is not good estimate in case of skewed data

Disadvantage of Mean

- The mean has one main disadvantage: it is particularly susceptible to the influence of outliers. These are values that are unusual compared to the rest of the data set by being especially small or large in numerical value. For example, consider the wages of staff at a factory below:

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

- The mean salary for these ten staff is \$30.7k.
- However, inspecting the raw data suggests that this mean value might not be the best way to accurately reflect the typical salary of a worker, as most workers have salaries in the \$12k to 18k range.
- The mean is being skewed by the two large salaries.

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data

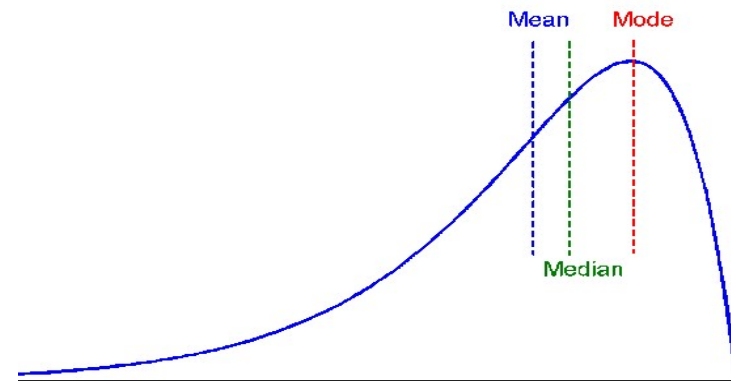
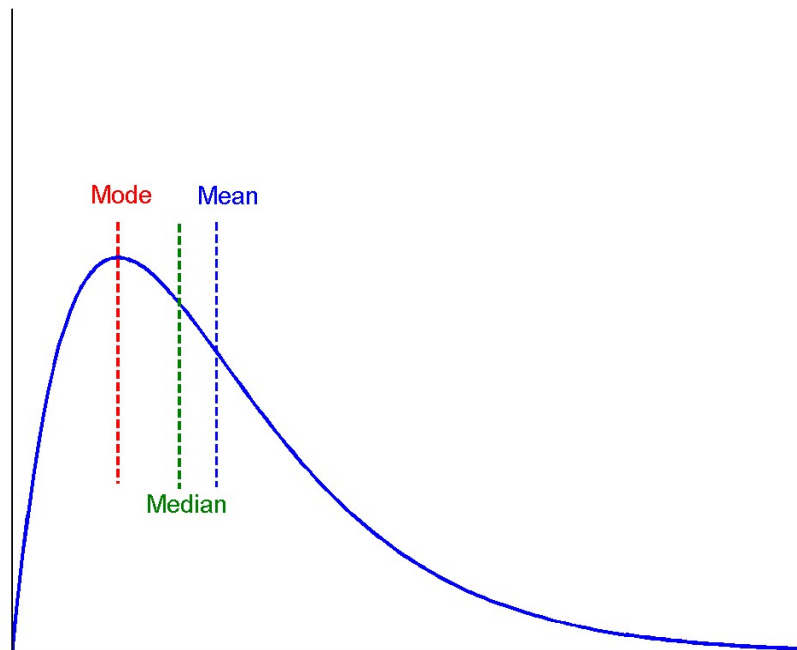
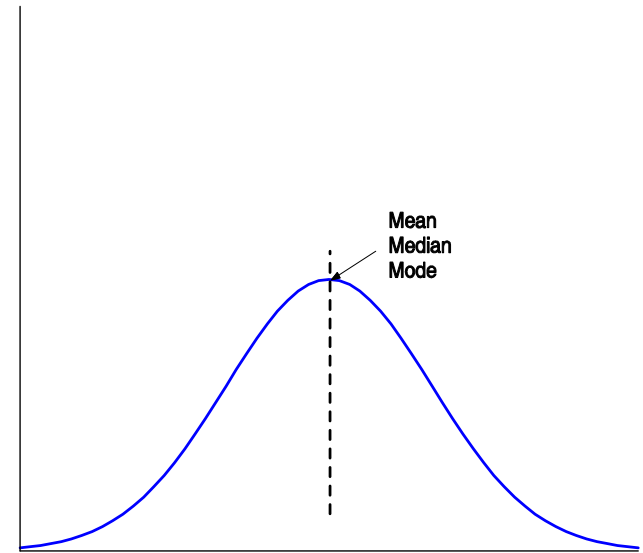


Figure Source : [1]

Median

Median:

Middle value if odd number of values, or average of the middle two values otherwise.

It is used when data is skewed.

For grouped data, Median

$$= l + \frac{\frac{N}{2} - C}{f} \times i$$

l is lower limit of the median class

f is the frequency of the median class

i is the width of the class-interval,

C is the total of all the preceding frequencies of the class

N is the total frequency of the data

Example :Calculate the median for the ungrouped data.

Find the median of each data set.

1) 2, 5, 6, 0, 9, 15, 12

2) 4, 5, 2, 6, 9, 16

Solutions

1) Arrange the data values from least to greatest.

0, 2, 5, 6, 9, 12, 15

Find the middle data point. Since there are seven (odd) observations, the median is the $(7+1)/2 = 4$ th data point. The 4th point is 6 so the median is 6.

2) Arrange the data values from least to greatest.

2, 4, 5, 6, 9, 16

Find the middle data point. Since there are six (even) observations, the median is at $(6+1)/2 = 3.5$, or the average of the 3rd and 4th data points, $(5+6)/2 = 5.5$.

Example :Calculate the median for the grouped data.

Class-interval	frequency	Cumulative frequency
5-10	5	5
10-15	6	11
15-20	15	26
20-25	10	36
25-30	5	41
30-35	4	45
35-40	2	47
40-45	2	49
	49	

Solution

$N/2=24.5$ and therefore, median class is 15-20,

$l=15,$ $i=5,$ $C=11,$ $f=15,$ $N=49$

$$\text{median} = 15 + \frac{\frac{49}{2} - 11}{15} \times 5 = 19.5$$

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

How to Handle Noisy Data?

- Using BoxPlot Analysis
- 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
 - detect and remove outliers
- Combined computer and human inspection
 - detect suspicious values and check by human (e.g., deal with possible outliers)

Boxplot Analysis

- Five-number summary of a distribution:
Minimum, Q1, Q2(Median), Q3, Maximum
- Boxplot
 - Data is represented with a box
 - The ends of the box are at the first and third quartiles, i.e., the height (width) of the box is IQR(Interquartile Range)
 - The median is marked by a line within the box

Box Plot Analysis

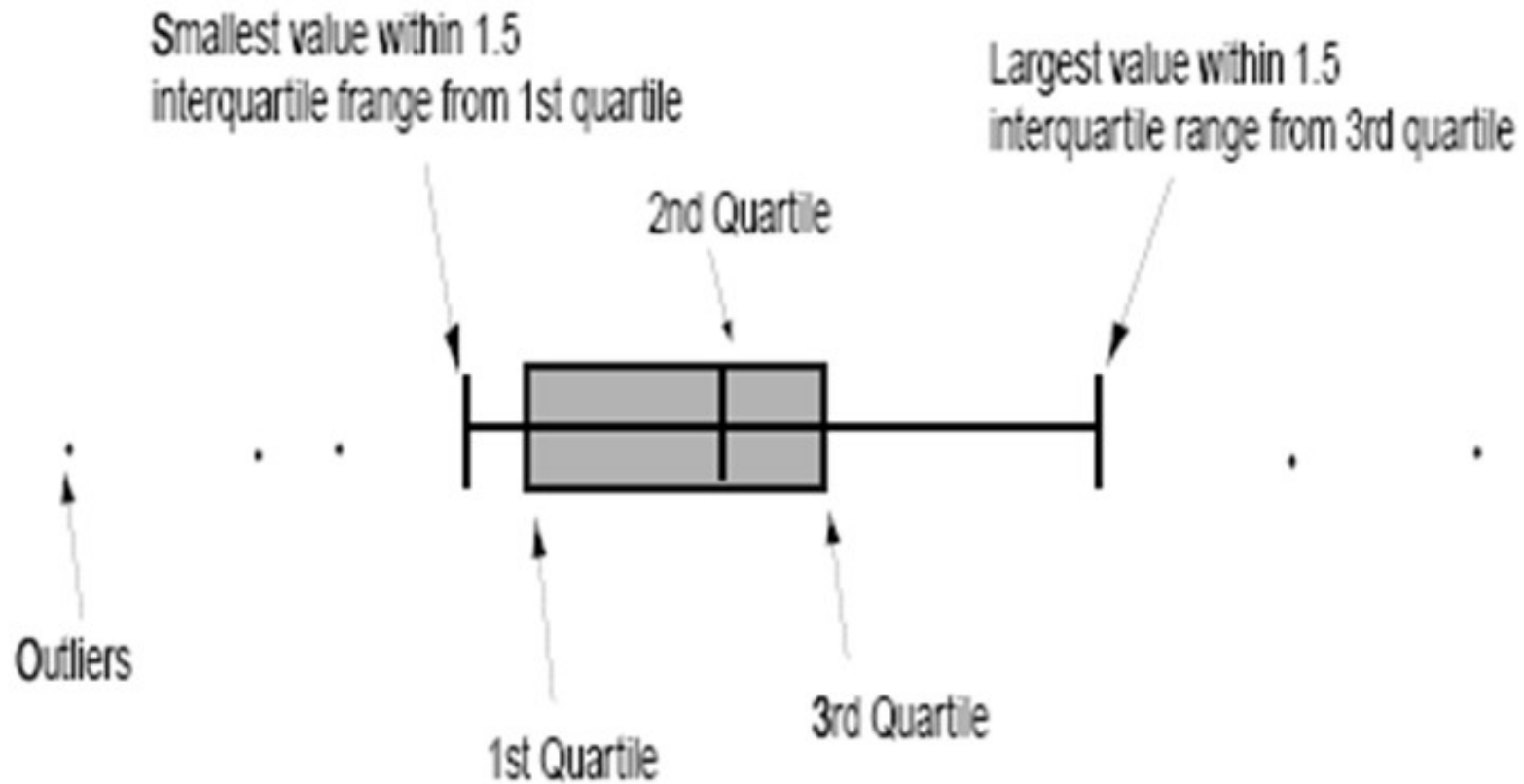


Figure Source : [3]

Quartiles

For ungrouped data

The **lower quartile(Q1)** is the middle value of the lower half.

The middle **quartile(Q2)** is the median

The **upper quartile(Q3)** is the middle value of the upper half.

For grouped data

The observation for the first quartile corresponds to $N/4$ th observation

$$\text{For grouped data, } Q_1 = l + \frac{\frac{N}{4} - C}{f} \times i$$

l is lower limit of the quartile class

f is the frequency of the median class

i is the width of the class-interval,

C is the total of all the preceding frequencies of the class

N is the total frequency of the data

Example: Median, Quartiles And Percentiles (Ungrouped Data)

Find the median, lower quartile and upper quartile of the following numbers.

12, 5, 22, 30, 7, 36, 14, 42, 15, 53, 25

Solution:

First, arrange the data in ascending order:

5,	7,	12,	14,	15,	22,	25,	30,	36,	42,	53
		↑			↑			↑		
		lower quartile			median			upper quartile		

In case of odd no of elements, median = $(n+1)/2$ th observation

In case of even no of elements, median = $[(n/2) \text{ th} + (n+1)/2 \text{ th observation}]/2$

Median (middle value) = 22

Lower quartile (middle value of the lower half) = 12

Upper quartile (middle value of the upper half) = 36

If there is an even number of data items, then we need to get the average of the middle numbers.

Interquartile Range

- Can eliminate some outlier problems by using the **interquartile range**
- Eliminate some high- and low-valued observations and calculate the range from the remaining values
- Interquartile range = 3rd quartile – 1st quartile
= $Q_3 - Q_1$

Binning Methods for Data Smoothing

❑ Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

Partition into equi-depth bins (divide the range into N intervals, each containing same number of samples):

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

* Smoothing by bin means:

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

* Smoothing by bin boundaries:

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

Regression : Noise Detection

Regression Formula:

Regression Equation(y on x) = $a + bx$

$$b_{yx} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2}$$

$$a = \bar{y} - b_{yx} \bar{x}$$

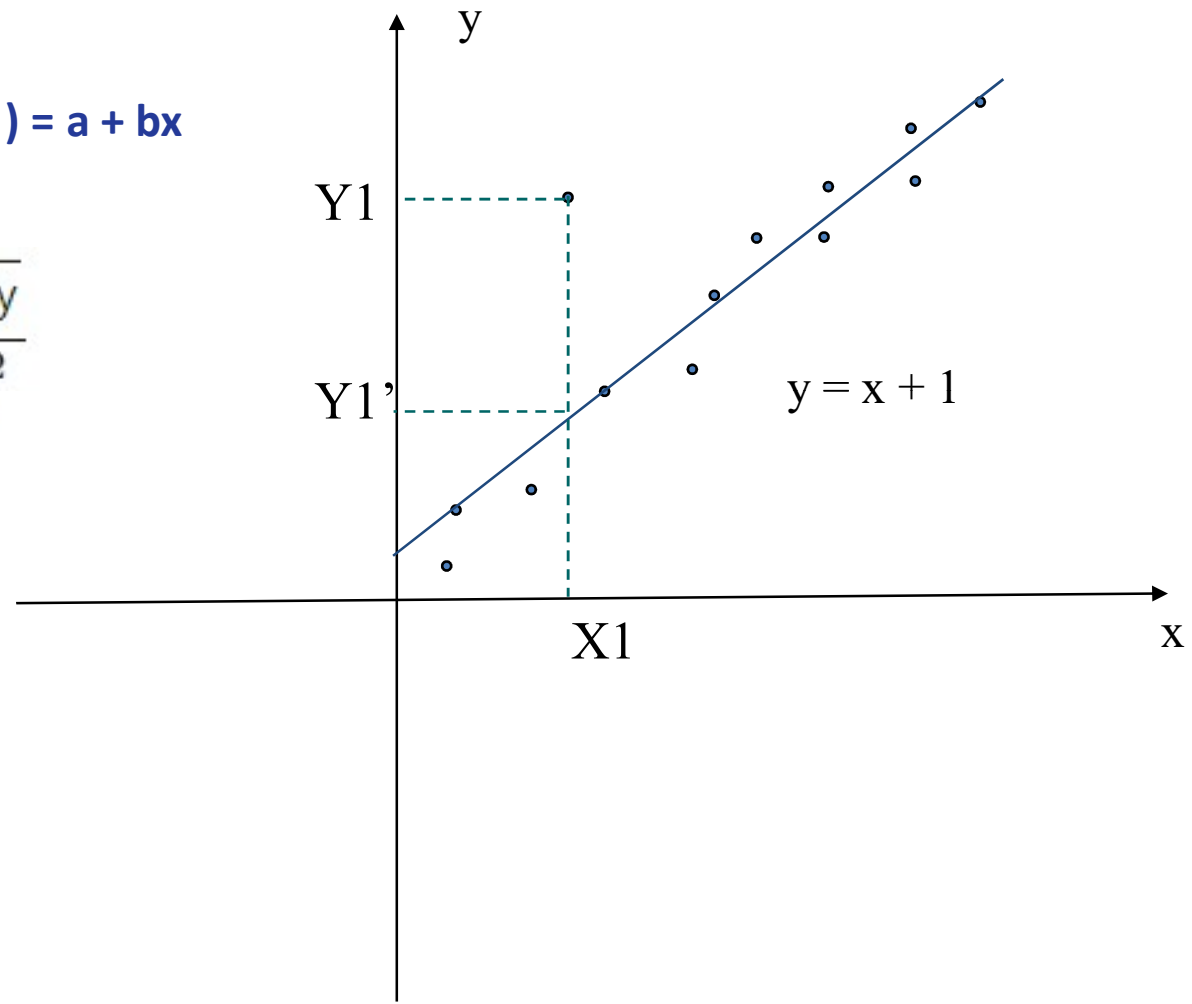


Figure Source : [1]

Example

Calculate regression equation Y on X.

X	10	12	13	17	18
Y	5	6	7	9	13

$$b_{YX} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2}$$
$$= \frac{600 - 5(14)(8)}{1026 - 5(14)^2} = \frac{600 - 560}{1026 - 980} = \frac{40}{46} = 0.87$$

$$\text{Also, } a = \bar{y} - b_{YX} \bar{x}$$
$$= 8 - 0.87 \times 14 = 8 - 12.18 = -4.18$$

\therefore The regression equation of Y on X is

Cluster Analysis : Noise Detection

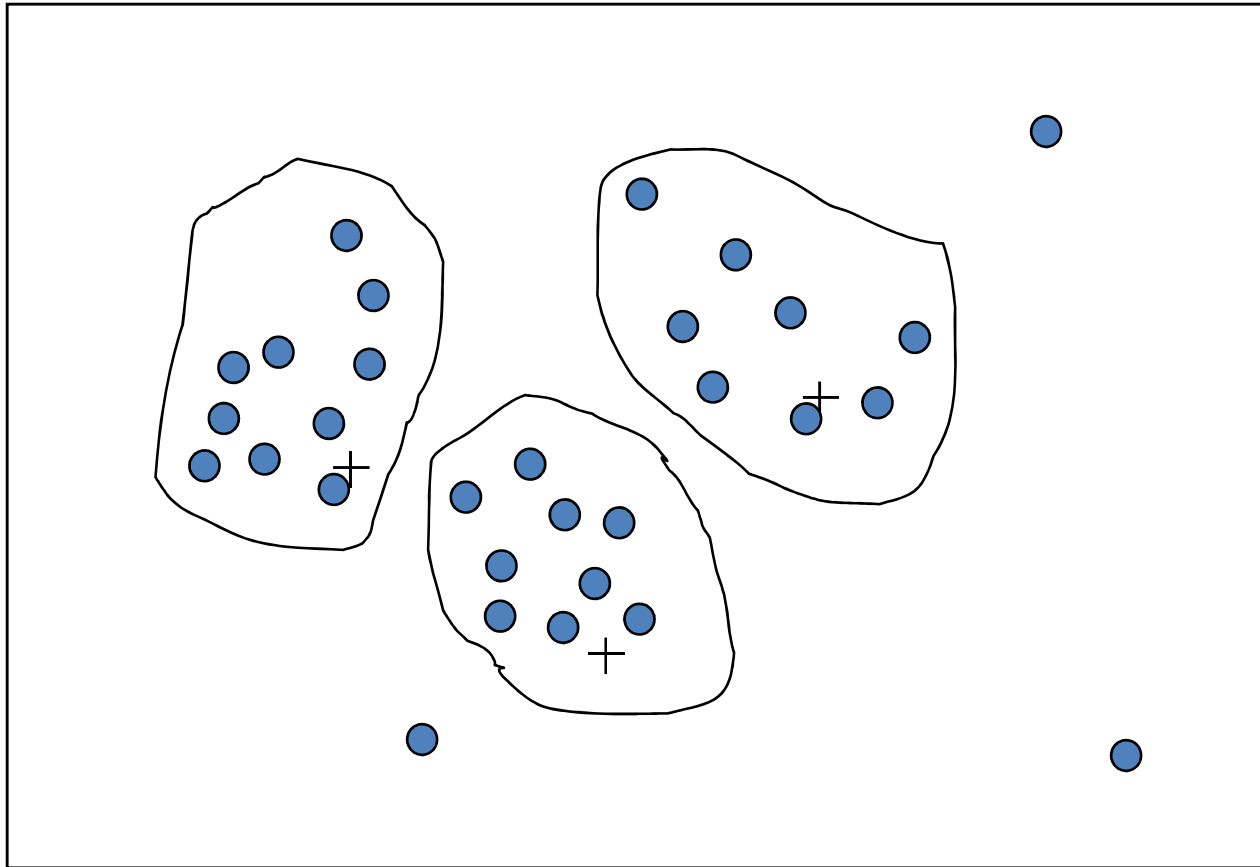


Figure Source : [1]

Inconsistent Data

- Use metadata (e.g., domain, range, dependency, distribution) to correct it manually.
- Functional dependencies can be used to find any violations.
- Use commercial tools
 - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
 - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)

Data Integration

- Combines data from multiple sources(Data Files, Data Cubes that allows data to be viewed in multiple dimensions , Multiple data bases) into a coherent store
- Problems
 - Entity identification problem : Identify real world entities from multiple data sources. E.g., A.cust-id \equiv B.cust-#
 - Data Redundancy Problem
 - The same attribute or object may have different names in different databases.
 - One attribute may be a “derived” attribute in another table, e.g., annual revenue
 - Duplication may be at tuple level.
 - Data value conflicts
 - For the same real world entity, attribute values from different sources are different
 - Possible reasons: different representations, different scales

Handling Redundancy in Data Integration

- Integrate metadata from different sources to handle entity integrity problem.
- Redundant tuples must be eliminated .
- 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

Correlation Analysis (Numerical Data)

- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum (A - \bar{A})(B - \bar{B})}{(n-1)\sigma_A\sigma_B} = \frac{\sum (AB) - n\bar{A}\bar{B}}{(n-1)\sigma_A\sigma_B}$$

where n is the number of tuples, \bar{A} and \bar{B} are the respective means of A and B , σ_A and σ_B are the respective standard deviation of A and B , and $\sum(AB)$ is the sum of the AB cross-product.

- If $r_{A,B} > 0$, A and B are positively correlated (A 's values increase as B 's). The higher, the stronger correlation.
- $r_{A,B} = 0$: independent; $r_{A,B} < 0$: negatively correlated.
- Standard deviation s (or σ) is the square root of variance s^2 (or σ^2)

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right]$$

What is the value of Co-relation coefficient??

Number of Study hours	2	4	6	8	10
Number of sleeping hours	10	9	8	7	6

Correlation Example

X	Y	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$
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	1
ΣX = 30	ΣY = 40	$\Sigma(X - \bar{X})$ = 0	$\Sigma(Y - \bar{Y})$ = 0	$\Sigma(X - \bar{X})(Y - \bar{Y})$ = -20	$\Sigma(X - \bar{X})^2$ = 40	$\Sigma(Y - \bar{Y})^2$ = 10

$$\bar{X} = \frac{\Sigma X}{n} = \frac{30}{5} = 6 \quad \text{and} \quad \bar{Y} = \frac{\Sigma Y}{n} = \frac{40}{5} = 8$$

$$r_{XY} = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}} = \frac{-20}{20} = -1$$

Correlation Analysis (Nominal Data)

- **χ^2 (chi-square) test** : It is used to examine the relationship between two or more qualitative or categorical variables

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

- The larger the χ^2 value, the more likely the variables are related
- Degree of freedom = (No of rows - 1) * (No of columns - 1)
- The cells that contribute the most to the χ^2 value are those whose actual count is very different from the expected count

H_0 : The two attributes A and B are independent,

H_0 : The null hypothesis is the hypothesis that there is no relationship between row and column frequencies.

H_1 : Alternate Hypothesis

Calculation of Expected Frequency

- **Expected Frequency:**

$$E_{ij} = \frac{O_{i*} O_{*j}}{N}, i = 1, 2, \dots, m; j = 1, 2, \dots, n$$

Total of observed freq in the ith row × total of a observed

$$E_{ij} = \frac{\text{freq. in the } j\text{th column}}{\text{Total of all cell frequencies}}$$

- **Degree of freedom:** Degrees of Freedom refers to the maximum number of logically independent values, which are values that have the freedom to vary, in the data sample.

$$\nu = (m - 1)(n - 1)$$

if $\chi^2 < \chi^2_{\alpha}$, H_0 is accepted at α % LOS, i.e. the attributes

A and B are independent.

If the calculated test statistic (χ^2_{test}) > the critical value from tables χ^2_{table} , then the null hypothesis is rejected and the alternative hypothesis is favoured.

Chi-Square Calculation: An Example

We want to know that science fiction and chess attributes are independent to each other

	Play chess	Not play chess	Sum (row)
Like science fiction	250	200	
Not like science fiction	50	1000	
Sum(col.)			

Degree of freedom = (row-1) * (col -1) = (2-1)*(2-1) = 1

and

chi- square value at 1% LOS = .00016

Chi-Square Calculation: An Example

	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

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

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

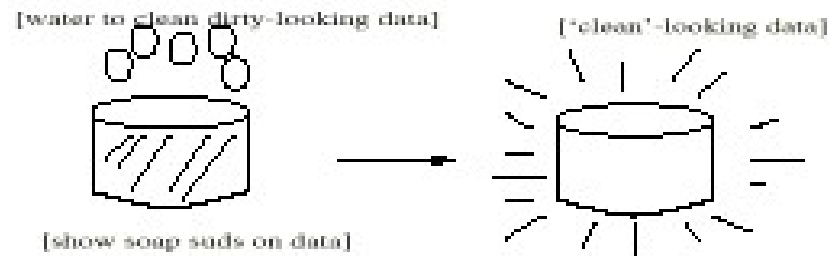
- It shows that like_science_fiction and play_chess are correlated in the group

<i>df</i> <i>lp</i>	.005	.01	.025	.05	.10	.90	.95	.975	.99	.995
1	.0000 4	.0001 6	.0009 8	.003 9	.0158	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.102 6	.2107	4.61	5.99	7.38	9.21	10.60
3	.0717	.115	.216	.352	.584	6.25	7.81	9.35	11.34	12.84
4	.207	.297	.484	.711	1.064	7.78	9.49	11.14	13.28	14.86
5	.412	.554	.831	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	.676	.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82

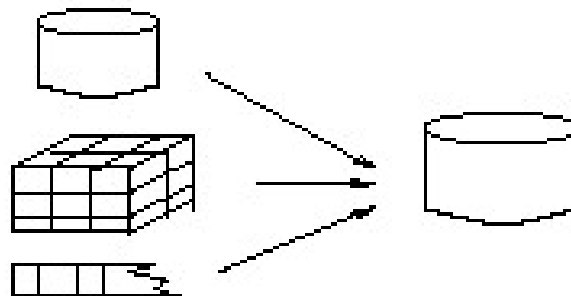
<i>df</i> <i>lp</i>	.005	.01	.025	.05	.10	.90	.95	.975	.99	.995
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.6	5.23	6.26	7.26	8.55	22.31	25	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
20	7.43	8.26	9.59	10.8 5	12.44	28.41	31.41	34.17	37.57	40.00
24	9.89	10.86	12.40	13.8 5	15.66	33.20	36.42	39.36	42.98	45.56
30	13.79	14.95	16.79	18.4 9	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.5 1	29.05	51.81	55.76	59.34	63.69	66.77
60	35.53	37.48	40.48	43.1 9	46.46	74.40	79.08	83.30	88.38	91.95
120	83.85	86.92	91.58	95.7 0	100.6 2	140.2 3	146.5 7	152.2 1	158.9 5	163.6 4

Forms of Data Preprocessing

Data Cleaning



Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction

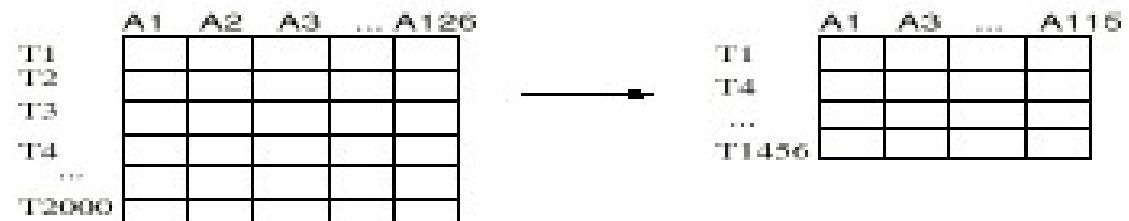


Figure Source : [1]

Data Transformation

- **Smoothing**: remove noise from data
- **Normalization**: scaled to fall within a small, specified range
 - min-max normalization
 - z-score normalization
 - normalization by decimal scaling
- **Aggregation**: summarization, data cube construction
- **Discretization**: where the raw values of a numeric attribute (e.g., age) are replaced by interval labels (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., youth, adult, senior).
- **Generalization**: concept hierarchy climbing
- **Attribute/feature construction**
 - New attributes constructed from the given ones.

Data Transformation: Normalization

The measurement unit used can affect the data analysis. Normalizing the data attempts to give all attributes an equal weight.

- **Min-max normalization:** to $[\text{new_min}_A, \text{new_max}_A]$

$$v' = \frac{v - \text{min}_A}{\text{max}_A - \text{min}_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

– Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then

\$73,000 is mapped to $\frac{73,600 - 12,000}{98,000 - 12,000} (1.0 - 0) + 0 = 0.716$

- Min-max normalization preserves the relationships among the original data values.
- It will encounter an “out-of-bounds” error if a future input case for normalization falls outside of the original data range for A.

Data Transformation: Normalization

The measurement unit used can affect the data analysis. Normalizing the data attempts to give all attributes an equal weight.

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

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

- Ex. Let income range \$12,000 to \$98,000 ,then \$73,000 is mapped to :
(Let $\mu = 54,000$, $\sigma = 16,000$)

$$\frac{73,600 - 54,000}{16,000} = 1.225$$

This method of normalization is useful when the actual minimum and maximum of attribute *A* are *unknown*, or when there are outliers that dominate the min-max normalization.

Source : [1]

Data Transformation: Normalization

The measurement unit used can affect the data analysis. Normalizing the data attempts to give all attributes an equal weight.

- Normalization by **decimal scaling**

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

Suppose that the recorded values of A range from 986 to 917. The maximum absolute value of A is 986. To normalize by decimal scaling, we therefore divide each value by 1000 (i.e., $j = 3$) so that 986 normalizes to 0.986 and 917 normalizes to 0.917.

Data Reduction Strategies

- Why data reduction?
 - A database/data warehouse may store terabytes of data
 - Complex data analysis/mining may take a very long time to run on the complete data set
- Data reduction
 - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- Data reduction strategies
 - Data cube aggregation:
 - Dimensionality reduction — e.g., remove unimportant attributes
 - Data Compression
 - Numerosity reduction — e.g., fit data into models
 - Discretization and concept hierarchy generation

Data Cube

- Data cube
 - Multidimensional view of aggregated data.
 - Defined by **dimensions** and **facts**.
 - **Dimensions** are the entities with respect to organization(location, time, product by sales department)
 - **Facts** are the measure/values in some unit.
 - **Facts** are **calculated** by **three operations** namely **aggregate - algebraic** (**mean, min, max**) and **holistic** (**median, mode**) functions.
 - The lowest level of a data cube (base cuboid)
 - can be obtained by OLAP operations such as **Roll up(drill up), drill down, slice & dice**.
 - The aggregated data for an **individual entity of interest** E.g., a customer in a phone calling data warehouse

Roll up/drill down/slice & dice

- **Roll-up**: The roll-up operation (also called the drill-up operation by some vendors) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction.
- **Drill-down**: Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either stepping down a concept hierarchy for a dimension or introducing additional dimensions
- **Slice and dice**: The slice operation performs a selection on one dimension of the given cube, resulting in a subcube. The dice operation defines a subcube by performing a selection on two or more dimensions.
- **Pivot (rotate)**: Pivot (also called rotate) is a visualization operation that rotates the data axes in view to provide an alternative data presentation.

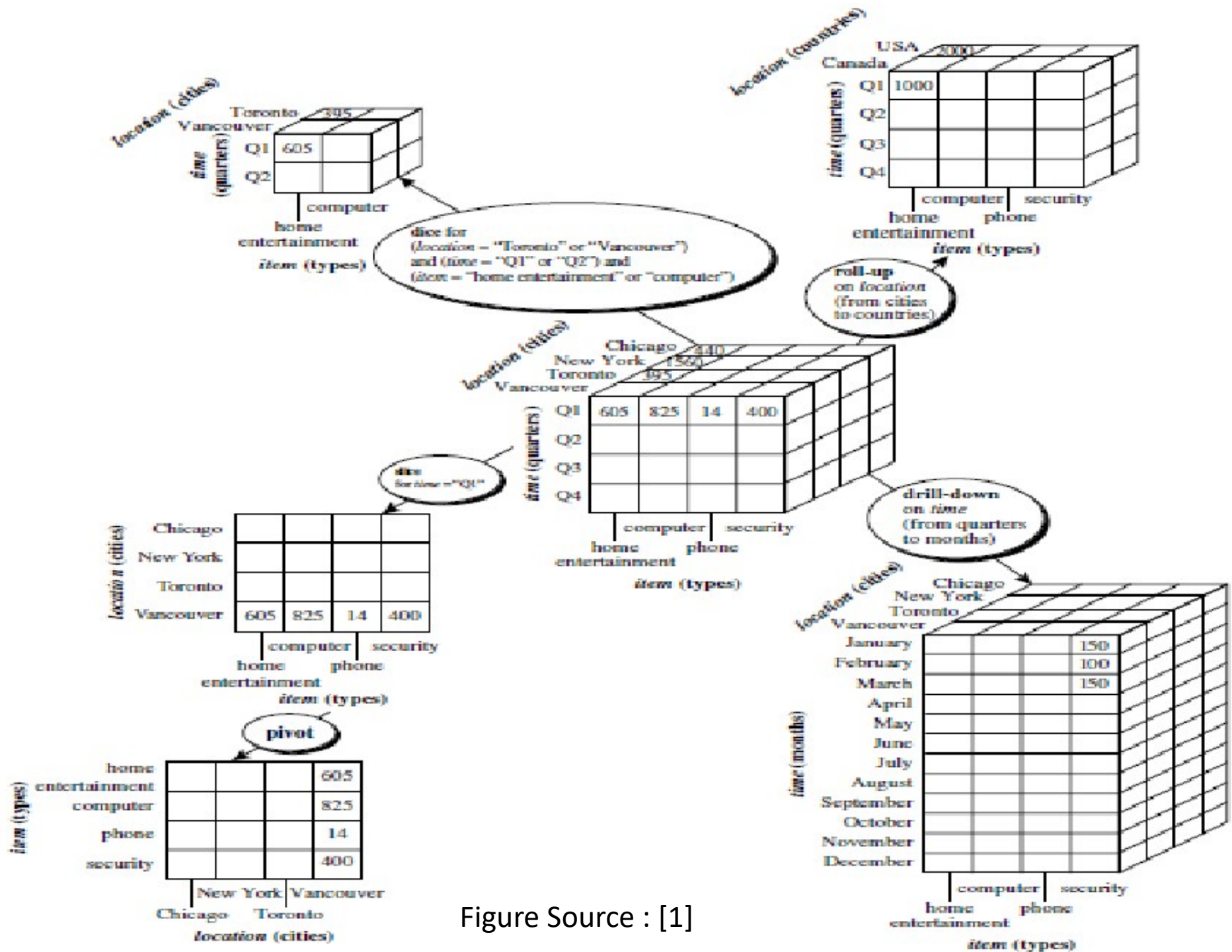
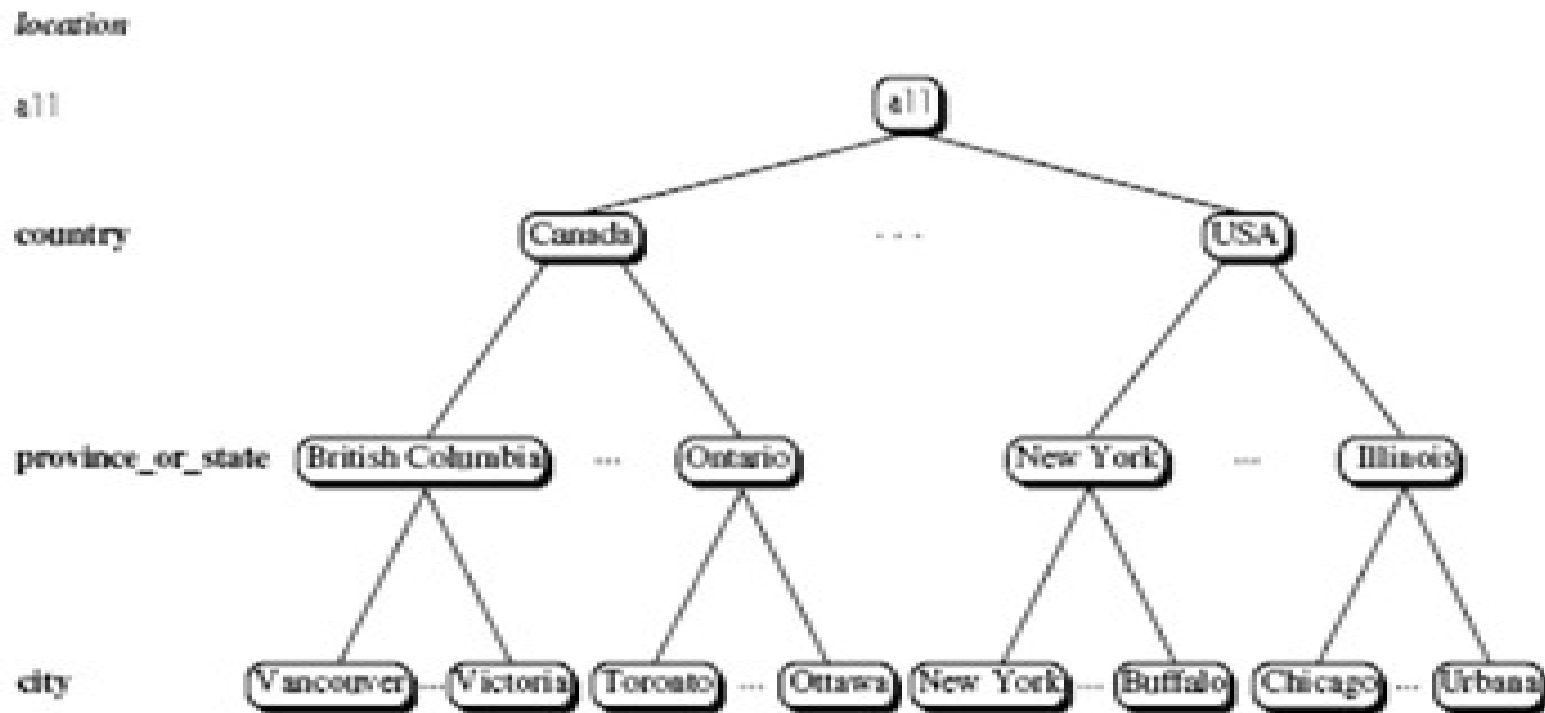


Figure Source : [1]

Data Cube Aggregation

- Multiple levels of aggregation in data cubes
 - Use of concept hierarchy(multi-level organization of concepts (attribute values)
 - Further reduce the size of data to deal with
- Queries regarding aggregated information should be answered using ROLL up and other operations , when possible

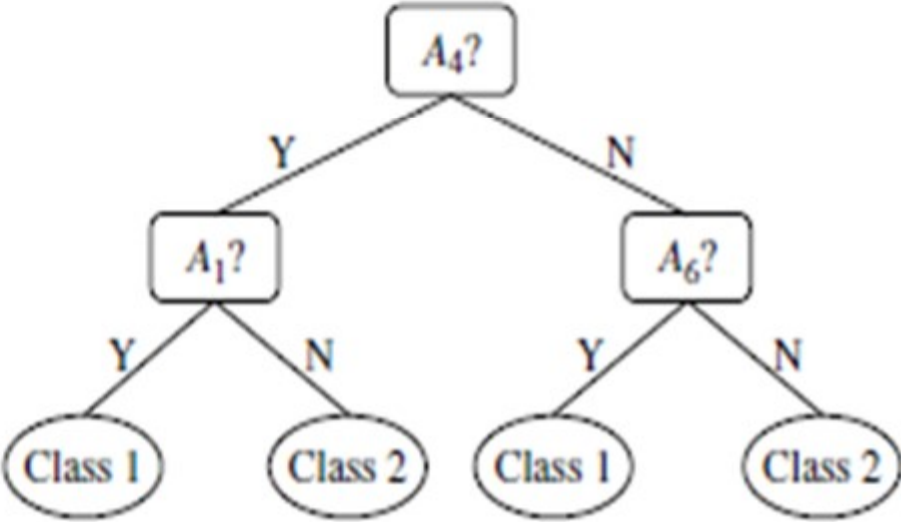


A concept hierarchy for the dimension *location*. Due to space limitations, not all of the nodes of the hierarchy are shown (as indicated by the use of “ellipsis” between nodes).

Figure Source : [1]

Dimensionality Reduction : Feature Selection Method

Feature selection works by removing features that are not relevant or are redundant.

Forward selection	Backward elimination	Decision tree induction
<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p> <p>Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>	<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p> <p>$\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>	<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p>  <pre> graph TD A4["A4?"] -- Y --> A1["A1?"] A4 -- N --> A6["A6?"] A1 -- Y --> C1_1((Class 1)) A1 -- N --> C2_1((Class 2)) A6 -- Y --> C1_2((Class 1)) A6 -- N --> C2_2((Class 2)) </pre> <p>\Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>

Source : [1,5]

Dimensionality Reduction : Feature Selection Method

Feature selection works by removing features that are not relevant or are redundant.

Filter Approaches: Features are selected before the data mining algorithm is run, using some approach independent of data mining task. E.g.: correlation, ...

Wrapper approaches: These methods use the target data mining algorithm as a black box to find the best subset of attributes, in a way similar to that of the ideal algorithm, but without enumerating all possible sets.

Embedded approaches: Feature selection occurs as a part of the data mining algorithm. During the operation of data mining algorithm, the algorithm itself decides which attributes to use and which to drop. E.g. decision tree classifiers often operate in this manner

Dimensionality Reduction : Feature Extraction Method

Feature extraction creates new variables as combinations of others to reduce the dimensionality of the selected feature.

- Principal Components Analysis (PCA)[Also used for Data Compression]
 - PCA finds the “principal components” in the data which are uncorrelated eigenvectors each representing some proportion of variance in the data.
 - Low variance PC can be removed.
- Wavelet transforms [Also used for Data Compression]
 - It is a linear signal processing technique that transforms a data vector D to numerically different vector D' of wavelet coefficients.
 - Lower wavelet coefficient can be removed.
- Others

Dimensionality Reduction :-Principal component analysis(PCA)

- ❖ **Principal Component Analysis** is a dimensionality reduction technique used in Machine Learning applications including Feature Engineering and Feature Extraction.
- ❖ PCA is a transformation procedure that converts a data matrix with possibly correlated features into a set of linearly uncorrelated variables called **principal components**.
- ❖ PCA finds a linear projection of high dimensional data into a lower dimensional subspace such that the variance retained is maximized and the least square reconstruction error is minimized.

Principal component analysis

- ❖ The objectives of PCA includes
 - finding relationships between observations,
 - extracting the most important information from the data,
 - outlier detection and removal, and
 - reducing the dimension of the data by keeping only the important information.
- ❖ All these **objectives** are achieved by finding the PCA space, which represents the direction of the **maximum variance of the given data**
- ❖ The PCA space consists of orthogonal principal components, i.e. axes or vectors.
- ❖ The principal components (PCs) are calculated by solving the covariance matrix or using Singular Value Decomposition (SVD).

Principal component analysis: Orthogonality

Two vectors are *orthogonal* to each other if their inner product equals zero. In two-dimensional space this is equivalent to saying that the vectors are perpendicular, or that the only angle between them is a 90° angle. For example, the vectors $[2, 1, -2, 4]$ and $[3, -6, 4, 2]$ are orthogonal because

$$[2, 1, -2, 4] \cdot [3, -6, 4, 2] = 2(3) + 1(-6) - 2(4) + 4(2) = 0$$

Steps in Principal component analysis

Step 1: Form a matrix of size $n \times p$ having n observations on p variables.

Step 2: Form a matrix with deviation from mean (mean centred matrix).

Step 3: Calculate the covariance matrix.

Step 4: Calculate eigen value and eigen vectors from the covariance matrix.

Step 5: Choose the PC (eigen vector) with high contribution of variance and form a feature vector. Contribution is calculated by dividing eigen value of specific PC with sum of eigen values of all PC.

Step 6: Derive the new data set by multiplying the mean centred matrix with feature vector. i.e.

$Z = X[A]$ where X is a mean centered matrix and A is a feature vector.

Eigen Vector and Eigen Values

- Let A be a square matrix. If λ is a scalar and X is a non-zero column vector satisfying

$$AX = \lambda X$$

X is an eigenvector of A ; λ is an eigenvalue of A .

- Eigenvectors are possible only for square matrices.
- Eigenvectors of a matrix are orthogonal.

- λ is an eigenvalue of an $n \times n$ matrix A , with corresponding eigenvector X .

$(A - \lambda I)X = 0$, with $X \neq 0$ leads to

$$|A - \lambda I| = 0$$

- There are at most n distinct eigenvalues of A .

Example

Obtain the eigenvalues and eigenvectors for the matrix,

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

The eigenvalues are obtained as

$$|A - \lambda I| = 0$$

$$\begin{vmatrix} 1-\lambda & 2 \\ 2 & 1-\lambda \end{vmatrix} = 0$$

$$(1-\lambda)(1-\lambda) - 4 = 0$$

$$\lambda^2 - 2\lambda - 3 = 0$$

Solving the equation,

$$\lambda = 3, -1$$

The eigenvalues are 3 and -1 for matrix A .

The eigenvector is obtained by

$$(A - \lambda I)X = 0$$

For $\lambda_1 = 3$

$$(A - \lambda_1 I)X_1 = 0$$

$$\begin{bmatrix} 1-3 & 2 \\ 2 & 1-3 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0$$

$$\begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = 0$$

$$-2x_1 + 2y_1 = 0$$

$$2x_1 - 2y_1 = 0$$

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

which has solution $x_1 = y_1$, x_1 arbitrary.

eigenvectors corresponding to $\lambda_1 = 3$ are the vectors $\begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$, with $x_1 \neq 0$.

e.g., if we take $x_1 = 2$ then $y_1 = 2$

The eigenvector is $\begin{bmatrix} 2 \\ 2 \end{bmatrix}$

Example

The procedure is explained with a simple data set of the yearly rainfall and the yearly runoff of a catchment for 15 years.

Year	1	2	3	4	5	6	7	8	9	10
Rainfall (cm)	105	115	103	94	95	104	120	121	127	79
Runoff (cm)	42	46	26	39	29	33	48	58	45	20

Year	11	12	13	14	15
Rainfall (cm)	133	111	127	108	85
Runoff (cm)	54	37	39	34	25

Mean of Rainfall = 108.5 cm
Mean of Runoff = 38.3 cm

Steps 1-3

Original matrix

$$\begin{bmatrix} 105 & 42 \\ 115 & 46 \\ 103 & 26 \\ 94 & 39 \\ 95 & 29 \\ 104 & 33 \\ 120 & 48 \\ 121 & 58 \\ 127 & 45 \\ 79 & 20 \end{bmatrix}$$

Matrix with deviations from mean

$$X = \begin{bmatrix} -1.3 & -3.4 \\ 8.7 & 7.4 \\ -3.3 & -12.6 \\ -12.3 & 0.4 \\ -11.3 & -9.3 \\ -2.3 & -5.6 \\ 13.7 & 9.4 \\ 14.7 & 19.4 \\ 20.7 & 6.4 \\ -27.3 & -18.6 \end{bmatrix}$$

Calculate the covariance matrix

$$\text{cov}(X, Y) = s_{X,Y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$\begin{bmatrix} \text{cov}(X, X) & \text{cov}(X, Y) \\ \text{cov}(Y, X) & \text{cov}(Y, Y) \end{bmatrix} = \begin{bmatrix} 216.67 & 141.35 \\ 141.35 & 133.38 \end{bmatrix}$$

Steps 4-5

Calculate the eigenvalues and eigenvectors of the covariance matrix

$$A = \begin{bmatrix} 216.67 & 141.35 \\ 141.35 & 133.38 \end{bmatrix}$$

Eigenvalues:

$$|A - \lambda I| = 0$$

$$\lambda_1 = 322.4 \text{ and}$$

$$\lambda_2 = 27.7$$

Eigenvectors:

$$(A - \lambda I)X = 0$$

$$X = \begin{bmatrix} 0.801 & -0.599 \\ 0.599 & 0.801 \end{bmatrix}$$

- First column: First Eigen vector
- Second column is second eigen vector

- Eigen vectors are of unit length. i.e. Square root of sum of square of both components of each eigen vector is 1.

The total variance accounted for by the first principal component is

$$\frac{\lambda_1}{\sum \lambda_j} = \frac{322.3}{350.1} = 0.92$$

i.e., 92% of total system variance is represented by the first principal component and the remaining 8% is represented by the second component.

Hence the second principal component can be neglected and only the first one considered.

From the two eigenvectors, the feature vector is selected

$$A = \begin{bmatrix} 0.801 \\ 0.599 \end{bmatrix}$$

Steps 6

$$Z = X A$$
$$\begin{bmatrix} -1.3 & 3.4 \\ 8.7 & 7.4 \\ -3.3 & -12.6 \\ -12.3 & 0.4 \\ -11.3 & -9.3 \\ -2.3 & -5.6 \\ 13.7 & 9.4 \\ 14.7 & 19.4 \\ 20.7 & 6.4 \\ -27.3 & -18.6 \end{bmatrix} \begin{bmatrix} 0.801 \\ 0.599 \end{bmatrix} = \begin{bmatrix} 0.995 \\ 11.39 \\ -10.2 \\ -9.61 \\ -14.8 \\ -5.20 \\ 16.60 \\ 23.39 \\ 20.41 \\ -33.0 \end{bmatrix}$$

Summary: Principal component analysis

Each principal component is a linear combination of each of original input variables.

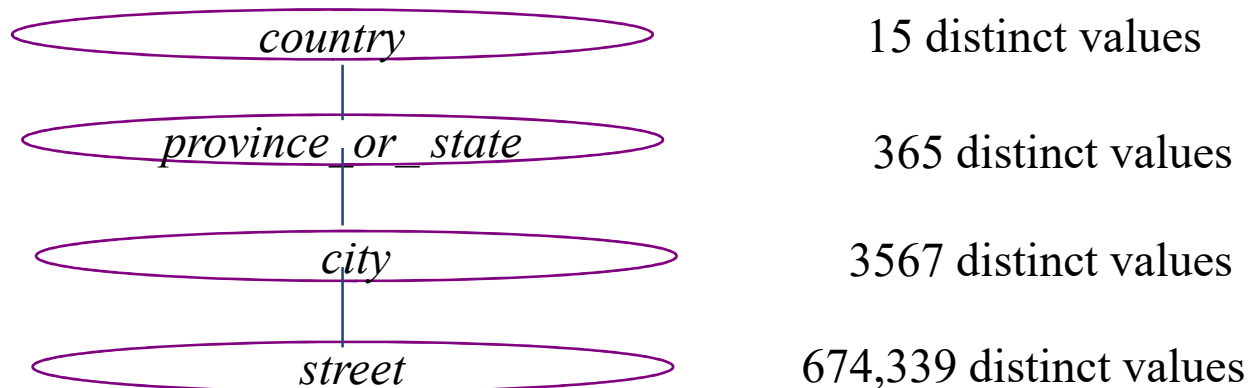
To reduce dimensionality, we can choose PCs with variance and discard lower one.

PCA converts a set of correlated variables in uncorrelated set of components.

Each PC is not original variable.

Data Reduction: Discretization

- **Discretization**: (Convert continuous attribute into discrete.i.e. categorical attributes)
 - **Unsupervised way**: Binning(equi-width/equi-depth : Interval labels can then be used to replace actual data values)
 - **Supervised way**: **Entropy based** discretization
 - Commonly used in classification / Association analysis
 - Concept hierarchy can be used for discretization
 - 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



Dimensionality Reduction : Histogram Analysis

Divide data into buckets and store average (sum) for each bucket
Buckets are shown on x-axis and y –axis shows average frequency.

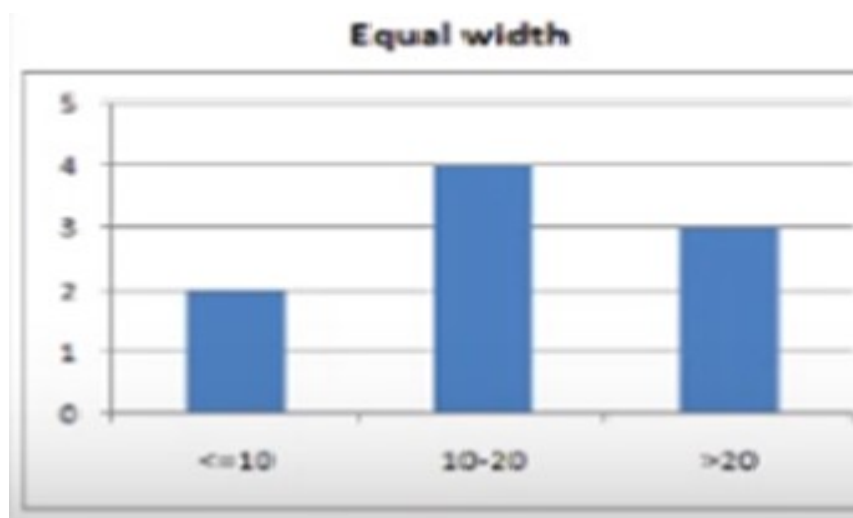
Approaches:

- Equi-width binning
- Equi-depth (equi-frequency) binning
- Others...

Histogram Analysis

Equi-width binning

- This algorithm divides the data into k intervals(groups/bins) of equal size(width).
- Width of interval (w) = $(\text{max}-\text{min})/k$
- Interval boundaries are :
 - $\text{Min}+w, \text{min}+2w\dots, \text{min}+(k-1)w$



Example (equi-width)

Data for given attribute in sorted :

5, 10, 11, 13, 15, 35, 50 ,55, 72, 92, 204, 215

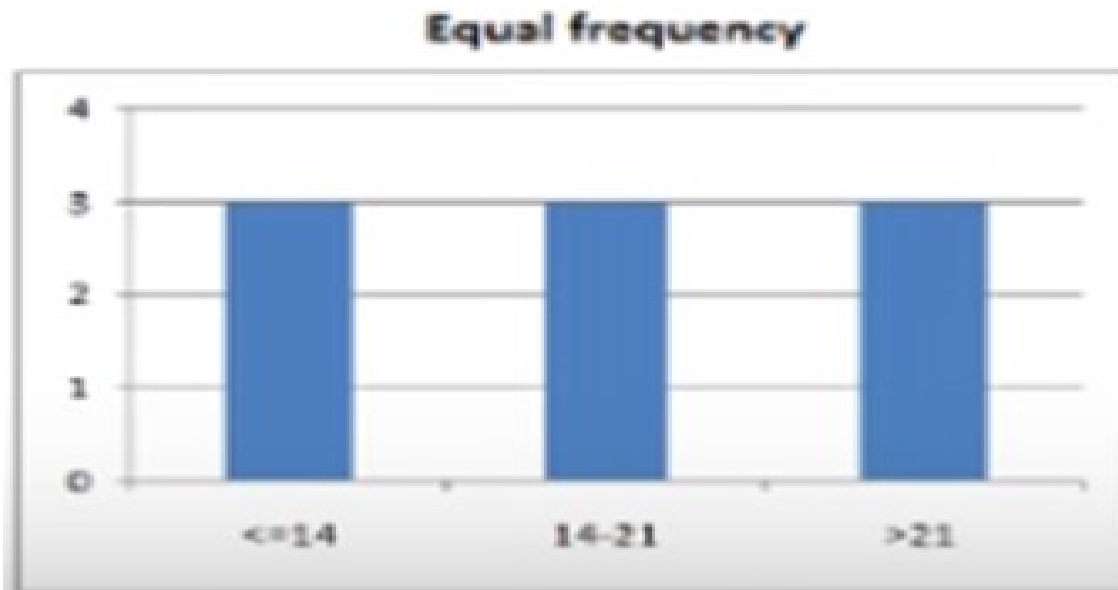
No. Of bins(N) = 3

width = (max-min)/N= (215-5)/3=70

bin1: 5,10,11,13,15,35,50,55,72	[5,75)
bin2: 92	[75, 145)
bin3: 204,215	[145, 215]

Histogram Analysis

- Equal-depth binning
 - ▣ Divides the range into N intervals, each containing approximately *same number* of records
 - ▣ Skewed data is also handled well
- Equal depth is also known as Equal frequency.



Discretization and Binarization

- It is often necessary to transform a continuous attribute to discrete one , and both cont..and discrete to binary attributes => Binarization.
- One way is to convert them into integers and write their binary equivalent.

Categorical value	Integer value	x1	x2	x3
poor	0	0	0	0
good	1	0	0	1
great	2	0	1	0
awful	3	0	1	1

- Another way is to introduce one variable(attribute) for each category.

Categorical value	Integer value	x1	x2	x3	x4
poor	0	1	0	0	0
good	1	0	1	0	0
great	2	0	0	1	0
awful	3	0	0	0	1

If this result into number of attributes, then dimensional reduction techniques can be taken before binarization.

Entropy based Discretization(Supervised way)

Given probabilities p_1, p_2, \dots, p_s whose sum is 1, **Entropy** is defined as:

$$H(p_1, p_2, \dots, p_s) = \sum_{i=1}^s (p_i \log(1/p_i))$$

Entropy measures the amount of randomness or surprise or uncertainty.

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

$$E(S, T) = \frac{|S_1|}{|S|} Ent(S_1) + \frac{|S_2|}{|S|} Ent(S_2)$$

- The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization.
- The process is recursively applied to partitions obtained until some stopping criterion is met, e.g.,

$$Ent(S) - E(T, S) > \delta$$

- Experiments show that it may reduce data size and improve classification accuracy

Example

<u>Color</u>	<u>Size</u>	<u>Shape</u>	<u>Edible?</u>
Yellow	Small	Round	+
Yellow	Small	Round	-
Green	Small	Irregular	+
Green	Large	Irregular	-
Yellow	Large	Round	+
Yellow	Small	Round	+
Yellow	Small	Round	+
Yellow	Small	Round	+
Green	Small	Round	-
Yellow	Large	Round	-
Yellow	Large	Round	+
Yellow	Large	Round	-
Yellow	Large	Round	-
Yellow	Large	Round	-
Yellow	Small	Irregular	+
Yellow	Large	Irregular	+

Edible 2 is a class attribute

1. Choose one split point and divide the data into intervals/groups and calculate the entropy for this .
2. Repeat it for different split points
3. Select the one which is giving the lowest entropy/ highest gain.

16 instances: 9 positive, 7 negative.

$$\text{Ent (S)} = - \left[\left(\frac{9}{16} \right) \log_2 \left(\frac{9}{16} \right) + \left(\frac{7}{16} \right) \log_2 \left(\frac{7}{16} \right) \right] = 0.9836$$

Data Reduction: Other methods:

- **Clustering** :Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only.
- **Sampling** obtaining a small sample s to represent the whole data set N .
 - **Sampling without replacement**
 - Once an object is selected, it is removed from the population
 - Not good for skewed data
 - **Sampling with replacement**
 - A selected object is not removed from the population
 - Not good for skewed data
 - **Stratified sampling:**
 - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
 - good for skewed data

Summary of Data Preprocessing

Data Preprocessing is needed to discover the good quality results. Following tasks are done to process the data before applying data mining tasks:

- Data Cleaning
- Data integration and Transformation
- Data reduction

References

- [1] Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2012(Third Edition).
- [2] Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. Introduction to data mining. Pearson Education India, 2016.
- [3] Dunham, Margaret H. Data mining: Introductory and advanced topics. Pearson Education India, 2006.
- [4] <https://www.javatpoint.com/data-mining-architecture>
- [5] Hira, Zena M., and Duncan F. Gillies. "A review of feature selection and feature extraction methods applied on microarray data." *Advances in bioinformatics* 2015 (2015).