UNIT II

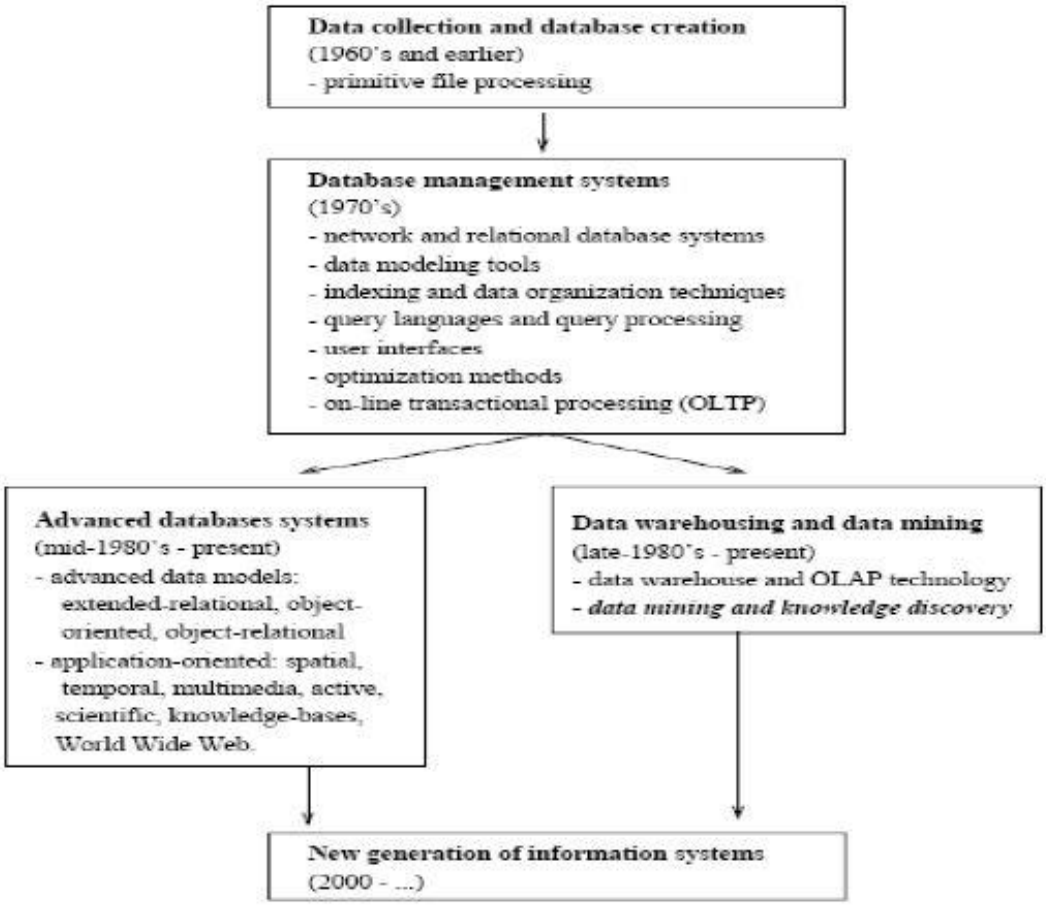
DATA MINING

CONTENTS

2.1INTRODUCTION

The major reason that data mining has attracted a great deal of attention in information Industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from business management, production control, and market analysis, to engineering design and science exploration.

The evolution of database technology
Figure 2.1. The evolution of database technology



Data mining refers to extracting or mining" knowledge from large amounts of data. There are many other terms related to data mining, such as knowledge mining, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery in Databases", or KDD

Essential step in the process of knowledge discovery in databases

Knowledge discovery as a process is depicted in following figure and consists of an iterative sequence of the following steps:

- 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 on some interestingness measures
 Knowledge presentation: where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

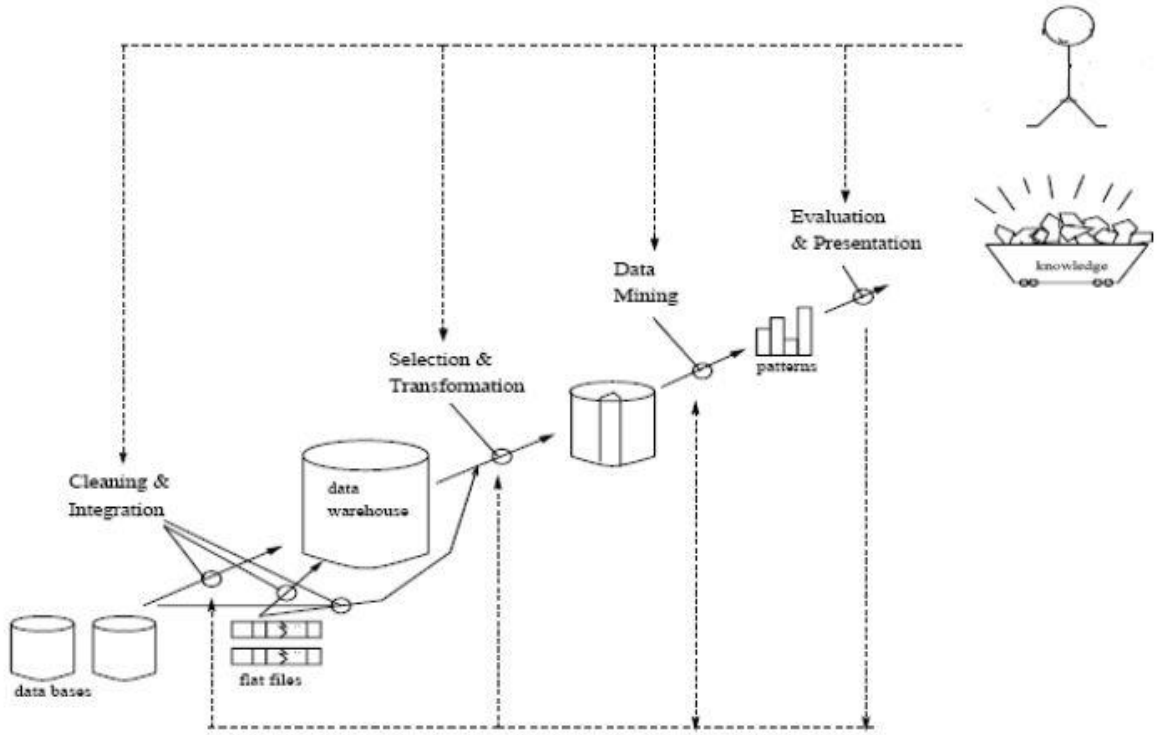


Figure 3.2 Data mining process of knowledge discovery

Architecture of a typical data mining system/Major Components

Data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. Based on this view, the architecture of a typical data mining system may have the following major components:

 A database, data warehouse, or other information repository, which consists of the set of databases, data warehouses, spreadsheets, or other kinds of information repositories containing the student and course information.
 A database or data warehouse server which fetches the relevant data based on users' data mining requests.

 A knowledge base that contains the domain knowledge used to guide the search or to evaluate the interestingness of resulting patterns. For example, the knowledge base may contain metadata which describes data from multiple heterogeneous sources.
 A data mining engine, which consists of a set of functional modules for tasks such as classification, association, classification, cluster analysis, and evolution and deviation analysis.
 A pattern evaluation module that works in tandem with the data mining modules by employing interestingness measures to help focus the search towards interestingness patterns.
 A graphical user interface that allows the user an interactive approach to the data mining system.

Differences between a data warehouse and a database

A data warehouse is a repository of information collected from multiple sources, over a history of time, stored under a unified schema, and used for data analysis and decision support; whereas a database, is a collection of interrelated data that represents the current status of the stored data. There could be multiple heterogeneous databases where the schema of one database may not agree with the schema of another. A database system supports ad-hoc query and on-line transaction processing.

Similarities between a data warehouse and a database

Both are repositories of information, storing huge amounts of persistent data.

2.2.Data

Data, information and knowledge are closely related terms, but each has its own role in relation to the other. Data are collected and analyzed to create information suitable for making decisions, while knowledge is derived from extensive amounts of experience dealing with information on a subject.

2.3.Types of Data

 Business transactions: Every transaction in the business industry is (often) "memorized" for perpetuity. Such transactions are usually time related and can be inter-business deals such as purchases, exchanges, banking, stock, etc., or intra-business operations such as management of in-house wares and assets.
 Medical and personal data: From government census to personnel and customer files, very large collections of information are continuously gathered about individuals and groups. Governments, companies and organizations such as hospitals, are stockpiling very important quantities of personal data to help them manage human resources, better understand a market, or simply assist clientele.
 Surveillance video and pictures: With the amazing collapse of video camera prices, video cameras are becoming ubiquitous. Video tapes from surveillance cameras are usually recycled and thus the content is lost. However, there is a tendency today to store the tapes and even digitize them for future use and analysis.
 Games: Our society is collecting a tremendous amount of data and statistics about games, players and athletes. From hockey scores, basketball passes and car-racing lapses, to swimming times, boxer s pushes and chess positions, all the data are stored. Commentators and journalists are using this information for reporting, but trainers and athletes would want to exploit this data to improve performance and better understand opponents.

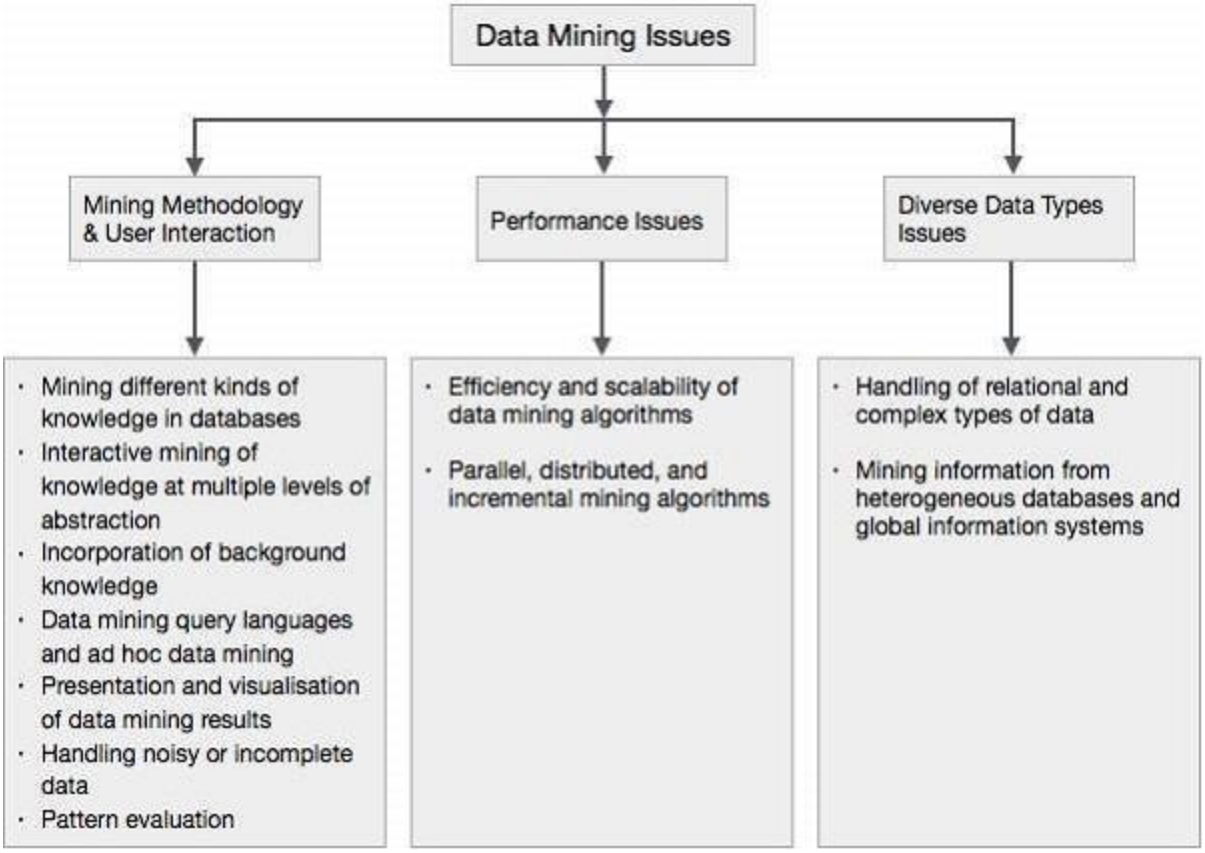
 Digital media: The proliferation of cheap scanners, desktop video cameras and digital cameras is one of the causes of the explosion in digital media repositories. In addition, many radio stations, television channels and film studios are digitizing their audio and video collections to improve the management of their multimedia assets.
 Virtual Worlds: There are many applications making use of three-dimensional virtual spaces. These spaces and the objects they contain are described with special languages such as VRML. Ideally, these virtual spaces are described in such a way that they can share objects and places.
 The World Wide Web repositories: Since the inception of the World Wide Web in 1993, documents of all sorts of formats, content and description have been collected and inter-connected with hyperlinks making it the largest repository of data ever built. Despite its dynamic and unstructured nature, its heterogeneous characteristic, and its very often redundancy and inconsistency, the World Wide Web is the most important data collection regularly used for reference because of the broad variety of topics covered and the infinite contributions of resources and publishers.
 Text reports and memos (e-mail messages): Most of the communications within and between companies or research organizations or even private people, are based on reports and memos in textual forms often exchanged by e-mail. These messages are regularly stored in digital form for future use and reference creating formidable digital libraries.
 CAD and Software engineering data: There are a multitude of Computer Assisted Design (CAD) systems for architects to design buildings or engineers to conceive system components or circuits. These systems are generating a tremendous amount of data. Moreover, software engineering is a source of considerable similar data with code, function libraries, objects, etc., which need powerful tools for management and maintenance.

2.4. ISSUES IN DATA MINING

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create some issues. Here in this tutorial, we will discuss the major issues regarding −

- Mining Methodology and User Interaction
- Performance Issues
- Diverse Data Types Issues

The following diagram describes the major issues.



Mining Methodology and User Interaction Issues

It refers to the following kinds of issues −

- **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore it is necessary for data mining to cover a broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
- **Incorporation of background knowledge** − To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.
- **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.

- **Handling noisy or incomplete data** − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation** − The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

Performance Issues

There can be performance-related issues such as follows −

- **Efficiency and scalability of data mining algorithms** − In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms** − The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

Diverse Data Types Issues

- **Handling of relational and complex types of data** − The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kind of data.
- **Mining information from heterogeneous databases and global information systems** − The data is available at different data sources on LAN or WAN. These data source may be structured, semi structured or unstructured. Therefore mining the knowledge from them adds challenges to data mining.

2.5.STATISTICAL DESCRIPTION OF DATA

Statistical descriptions are usually classified by the features of the sample data that they are trying to describe. The most common ones are *measures of location* or center, which are indicative of the 'center' of the data, while the *measures of variation* are indicative of the variability of the data.

2.5.1.Measures of Location: The mean

Sample mean

The sample mean is obtained by adding all the values in your sample and dividing by the sample size (which is usually denoted by small *n*). In mathematical notation, we have

(1)

$$\bar{x}=\sum xn$$

Notice that the symbol for the sample mean is $\bar{x}$, an x with a bar above, and it is read x bar. The symbol
Σ is the sum sign in mathematics, which means that you should add up all the values in your sample.

Say, for example that you collected the age of 4 students in the class to estimate the average age of the whole class. Then the 4 students are the *sample*, and the whole class the *population*.

The population size is $n=4$.

If the ages you collected are 22, 21, 48, and 21, then the sample values are written as

(2)

$$x_1=22,\ x_2=21,\ x_3=48,\ x_4=21.$$

and for this particular example, the mean is

(3)

$$\bar{x}=\sum_{i=1}^{4} x_i 4=x_1+x_2+x_3+x_4 4=22+21+48+214=1124=28$$

Population mean

If we were to calculate the mean from a population instead of a sample, then we would still proceed in the same way, we would add up all the values in the population (more values to add) and we would divide by the population size (denoted by N).

The symbol for the population mean is the Greek letter mu: μ, so we obtain

(4)

$$\mu=\sum xN$$

In the case of the population of students in the classroom with the following ages:
21 21 25 20 26 22 46 25 58 24 20 20 25 23 27 21 23 22 28, the population mean is 26.2.

In this case, it is understood that the $\sum$ (Sigma)- sign indicates the sum over all the elements of the population (not only the sample), therefore, in this case, the sum indicates $x_1+x_2+x_3+...+x_N$, where in this case N= 20, instead of $x_1+x_2+...+x_n$ in the case of the sample mean, where n=4 in our example above. When we want to make this difference more evident, then we write $\sum_{i=1}^{N}$ for the population, and $\sum_{i=1}^{n}$ for the sample.

Properties of the mean

The mean as a measure of center is probably the most frequently used measure of center, partly because it has the following properties:

1. it can be calculated for any numerical data set.
2. it's value is unambiguous and unique for a given data set
3. it lends itself to further statistical treatment
4. if each value in the sample would be replaced by the mean, $\sum x$ would remain unchanged
5. it takes into account every value in the data set
6. it is relatively reliable (does not fluctuate widely when selecting different samples)

However, **outliers** have a strong effect on the mean, particularly if the sample size is small. Therefore, at times the **trimmed mean** is used instead, in which case the upper and lower 5% is deleted, and the mean is taken without those values.

In our population above, we would *trim* the data set 58 and the data set 20, and then add the values and divide by 18, to obtain 23.3.

2.5.2.Measures of Location: The weighted mean

When taking averages, it is often important to take into account that data have different *weights*; some data points are more "important" than others. For example, if we have the household income per state, we see that Alaska has the 4th highest median household income (64,333) , while New York has the 19th highest (53,514).
If we want to calculate the national average, however, New York's value counts much more than Alaska's because New York's population is much larger than Alaska's (19,490,297 *vs.* 686,293).

Therefore, to get the correct mean, we would have to weight each state's household income with the appropriate weight, which is a measure of relative importance (in this case the population size).

The trimmed mean is given by

(5)

$$x_w=\sum w\cdot x\sum w=\sum_{i=1}^{n} w_i\cdot x_i\sum_{i=1}^{n} w_i=w_1x_1+x_2x_2+\cdots+w_nx_nw_1+w_2+\cdots+w_n$$

For example, suppose that we have a sample with the 2 states named above and Texas ($47,548) and Mississippi ($36,338), with corresponding populations 24,326,974 and 2,938,618 respectively.

Then, the weighted mean is

(6)

$$x_w = 19490297 \cdot 53514 + 686,293 \cdot 64333 + 24326974 \cdot 47548 + 2938618 \cdot 3633819,490,297 + 686,293 + 24,326,974 + 2,938,618$$

which gives a weighted average of $49, 547. The national median household income is $50,740.

Grand mean

A special case of the formula for the weighted average is the **grand mean**, which is the overall mean. In that case, it has a special notation and slightly differenct formula:

(7)

$$\overline{x} = \sum_{i=1}^{k} n_i \cdot \overline{x_i} \sum_{i=1}^{k} n_i = n_1 \overline{x_1} + n_2 \overline{x_2} + \cdots + n_k \overline{x_n} n_1 + n_2 + \cdots + n_k$$

2.5.3. Measures of Location: The Median and other fractiles

The median is a measure of center, like the mean, which is not affected by outliers like the mean is. The symbol used for the sample median is $\tilde{x}$and for the population median $\tilde{\mu}$.

To obtain the median, we first need to re-arrange the data in ascending order (sorted), and then find the middle value, namely,

- *when n is odd, the median is the value in the middle (after sorting)*
- *when n is even, it is the mean of the two items nearest to the middle*

For example, suppose that we select a sample of size 4 from the population of students in the class listed above, obtaining the following ages: 20, 26, 22, 46.

Then, since $n=4$, the median is the average of the two values in the middle (after sorting, hence the average of 22 and 26, or $\tilde{x}=24$.

If we added another value to the sample, say 28, then the sorted sample would be 20, 22, 26, 28, and 46, and hence the median would be $\tilde{x}=26$.

When dealing with a small population, or larger sample, it is sometimes convenient to make a stem-and-leaf plot to find the median. The reason for this is that the stem-and-leaf plot **sorts the data**.

For example, in the case of the GDP per capita of the 20 countries listed in chapter 2, we got the following stem and leaf plot:

```
The decimal point is 1 digit(s) to the right of the |

3 | 99999
4 | 02445669
5 | 5567
6 |
7 | 06
8 | 1
```

Since *n=20*, we need to find the average between the values in the 10th and 11th position, namely 45 and 46 thousand USD. Therefore, the median GDP per capita, for the 20 richest countries (per capita) is 45,500 USD.

Fractiles

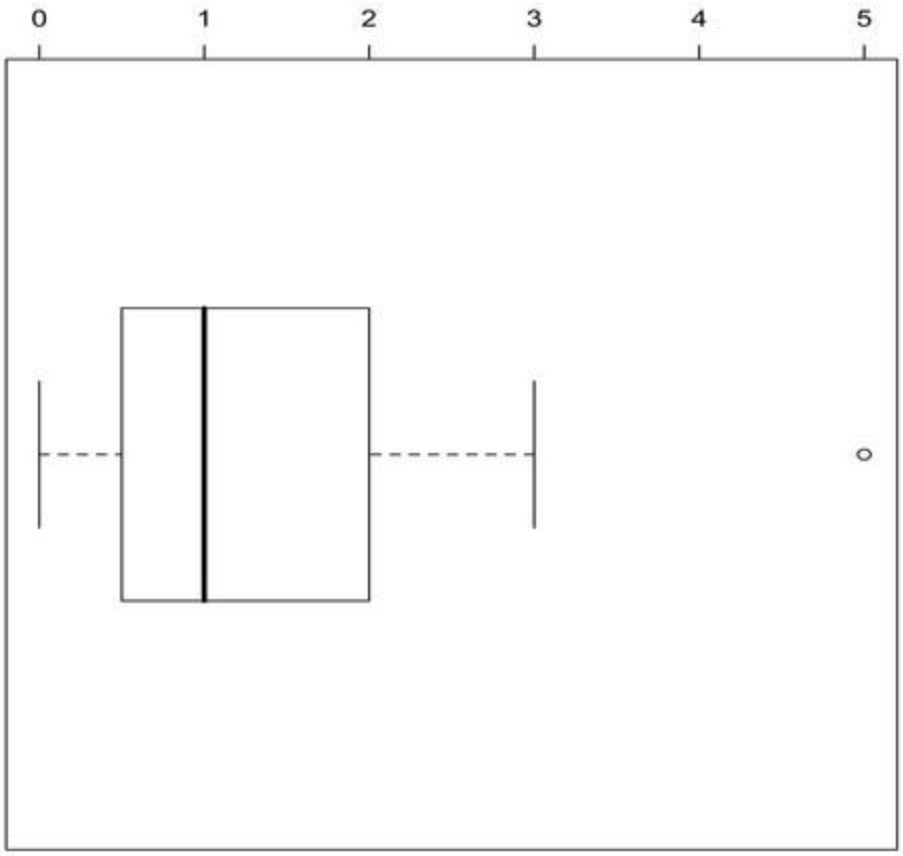
The median is one example of a fractile, a measure (value) that divides the data set into two or more equal parts. Another such example are quartiles, that divide the data into 4 equal parts, obtaining the values *Q1, Q2,* and *Q3*. Q4 is the maximum value, so no need to calculate it. Namely, Q1 is the value such that 25% of the data fall below it, and Q3 is such that 75% of the data fall below it, while Q2 is the median, and hence 50% of the data values fall below.

In the list above, there is 5 countries whose GDP per capita is 39,000 and the next one has a GDP per capita of 40,000, hence, Q1=39,500. Similarly, Q2 = 45,500 = $\tilde{x}$ and Q3 = 55,500.

Boxplot

A box plot is a graph that summarizes the data by representing 5 values, the minimum and maximum value, the Q1, Q2 and Q3.

In the graph below, the 3 central horizontal lines represent Q1, Q2 and Q3, while the point on the extreme represents an outlier value. This is a variation of the boxplot as described in the previous line, since the other two horizontal lines cannot be the minimum and maximum. They are the 5th and 95th percentile.



2.5.4. Measures of Location: The mode

The mode is a measure of center that is usually used for data that is non-numerical. Namely, the mode is the *value that occurs most frequently*, and it is the only value that can be collected for qualitative data.

For example, in example 3.17, they list the size of dresses sold by a store to be 10,7, 14, 9, 9, 14, 18, 9, 11, 12, 16, 14, 9, 14, 14, 11, 9 and 20. In this case, the number 9 and 14 appear most frequently, both showing exactly 5 times.
Therefore, you would say in this case that this data is bimodal (has two modes), namely 9 and 14.

2.5.5. Measures of Variation: The range

The range is a measure of variation or variability of the data. *The range of a data set is the largest value minus the smallest value*.

For example, for the age distribution in the class, which we write here again for convenience:
21 21 25 20 26 22 46 25 58 24 20 20 25 23 27 21 23 22 28,

the range is 58-20 = 38 years.

One disadvantage of the range is that it is highly influenced by outliers. For that reason, we use the following measure of variation more often.

2.5.6. Measures of Variation: The standard deviation

The standard deviation is the most general measure of variation.

To calculate the standard deviation of a population, one first takes the difference of each data point to the mean (the variation), and squares that difference (to insure it's positive). Then, all those squared differences
are added together and divided by *N*, the size of the population. Finally, the square root is taken. This is summarized in the following formula:

Population standard deviation

(8)

$$\sigma=\sqrt{\frac{\sum(x-\mu)^2}{N}}$$

The square of this quantity is called the **population variance**, and even though it is not a measure of variation *per se*, it is a notion that we will widely use during the course.

Population variance

(9)

$$\sigma^2=\frac{\sum(x-\mu)^2}{N}$$

Population example

Suppose that we had 20 students in the classroom with the following ages:
21 21 25 20 26 22 46 25 58 24 20 20 25 23 27 21 23 22 28 23.

The mean μ=26 in that case.

Then, one way to calculate the standard deviation of the population would be to do the following table:

x	x-μ	(x- μ)2
21	21 - 26=-5	25
21	21 - 26=-5	25
25	25 - 26=-1	1
20	20 - 26=-6	36

and so on, until the last two rows:

28	28 - 26=2	4
23	23 - 26=-3	9
Sum	**=**	**1678**

Therefore, the variance of the population is $\sigma 2=1678/20=83.9$, and the standard deviation is the square root of this result, namely $\sigma=9.16$.

Sample standard deviation

The sample standard deviation is calculated almost in the same way as the population standard deviation, but substituting the sample mean x^{-} for the population mean σ and the population size *N* for the **sample size minus 1: *n-1***.

(10)

$$s=\sum(x-x^{-})2n-1----------\sqrt$$

The square of this quantity is called the **sample variance**.

Sample variance

(11)

$$s2=\sum(x-x^{-})2n-1$$

Sample example

Suppose that we take a sample of 4 students out of the population listed above, say the ones with ages 26 22 46 25.

Then the sample standard deviation can be calculated by though the table:

x	**x-** $\bar{x}$	**(x-** $\bar{x}$**)2**
27	27 - 30=-3	9
22	22 - 30=-8	64
46	46 - 30=16	256
25	25 - 30=-5	5
Sum	**=**	**354**

This total we would have to divide by *n-1=3* to obtain the sample variance s^2=118. The sample standard deviation is then *s*=10.86.

"Fast" formula

An alternative but equivalent formula for calculating the sample standard deviation is

(12)

$$s=S_{xx}n-1-----\sqrt{}$$

where

(13)

$$S_{xx}=\sum x2-(\sum x)2n.$$

The advantage of this formula is that you need one less column to calculate the sample standard deviation, which could lead to faster calculations, specially when a lot of data is involved. For the example above, we obtain:

x	**x^2**
27	729

22	484
46	2116
25	625
$\sum x$**=120**	$\sum x$**=3954**

Therefore,

$S_{xx}=3954-(120)^2/4=354$

Dividing this by $n-1=3$, we obtain 118. This is again the value of the sample variance, and therefore we obtain the same value for the sample standard deviation, namely, $s=10.86$.

2.5.7. Application of the standard deviation

Chebyshev's theorem

For any set of data and any constant k greater than one, at least $1-1/k^2$ of the data must lie within k standard deviations on either side of the mean.

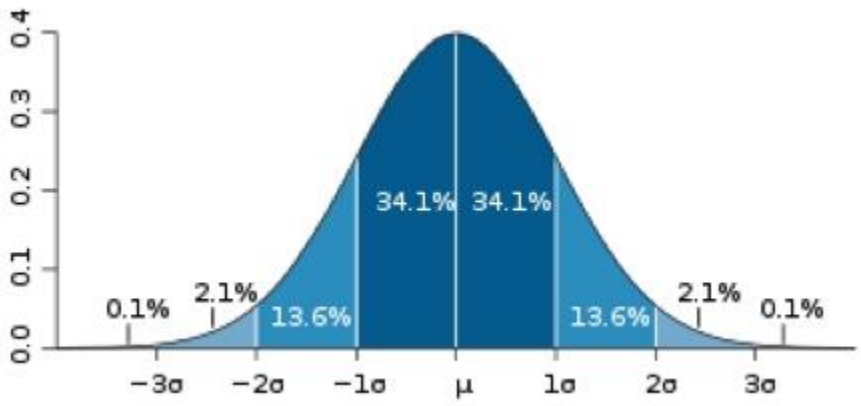
This theorem gives us a broad bound of how much data should be inside the mean plus or minus k standard deviations.

For example, when $k=3$, it is telling us that at least

$1-1/(3^2)=1-1/9=8/9$ or approximately 89%

of the data must lie within 3 standard deviations from the mean.

We will see in this course that many data sets follow a *normal distribution*, which is a bell-shaped distribution like the one in the graph below:



For the normal distribution, Chebyshev's theorem applies, but there is actually a more precise **empirical rule**:

1. About 68% of all the data values lie within 1 standard deviation from the mean
2. About 95% of all the data values lie within 2 standard deviation from the mean
3. About 99.7% of all the data values lie within 3 standard deviation from the mean

For example, if the height of women in the US is normally distributed with a mean of 64 inches and a standard deviation of 2.5 inches, this implies that 95% of all women in the US are between 59 and 69 inches tall.

z-scores

The *z*-score is a measure of relative value which is useful to compare values from different data sets, it is calculated using the following formula in the case of a population:

(14)

$$z=x-\mu\sigma$$

and in the case of the sample by

(15)

$$z=x-\bar{x}s.$$

For example, Rebecca Lobo, a female basketball player, is 76 inches tall, therefore here *z*-score is

(16)

$$z=76-642.5=4.8,$$

using the values for the population mean and standard deviation given above.

In other words, she is 4.8 standard deviations higher than the average US woman. That is extraordinarily tall!

Coefficient of variation

A measure that allows us to compare variation from different data sets is the **coefficient of variation**, given by

$V=\frac{s}{\bar{x}}\cdot 100\%$ or $V=\frac{\sigma}{\mu}\cdot 100\%$

For example, if one statistics class averaged 75 with a standard deviation of 10 points and another one averaged 65 with a standard deviation of 8 points, the coefficient of variation would allow us to find which class has less variability (is more homogeneous).

The coefficients of variation are $\frac{10}{75}100\%=13.3\%$ and $\frac{8}{65}100\%=12.3\%$, so that the second class is more homogeneous (or consistent).

Grouped data

When we create or receive a frequency distribution, the data has been grouped already, and therefore we have lost some of the original information.

For example, in the last section we saw the frequency distribution of GDP per capita of the 20 richest countries:

GDP range	Number of countries
80,000 -89,999	1
70,000-79,999	2
60,000-69,999	4
50,000-59,999	0
40,000-49,999	8
30,000-39,999	5
Total	**20**

Even though we have lost some information, we can still calculate and approximate mean and standard deviation. Namely, let $f_1, f_2, \ldots f_k$ be the class frequencies, and let $x_1, x_2, \ldots x_k$ be the midpoints of every class, then we can approximate the mean by

$$\bar{x} = \sum x_k f_k n$$

and the standard deviation using the "fast formula" by

$$s = \sqrt{S_{xx} n - 1} \text{ where } S_{xx} = \sum x^2 f - (\sum x f)^2 n$$

Extending the table above, we obtain

GDP range	f	x	xf	x^2f
80,000 -89,999	1	85K	85	7225
70,000-79,999	2	75K	150	11250
60,000-69,999	4	65K	260	16900
50,000-59,999	0	55K	0	0
40,000-49,999	8	45K	350	16200
30,000-39,999	5	35K	175	6125
Total	**20**		**1030**	**57700**

Therefore,

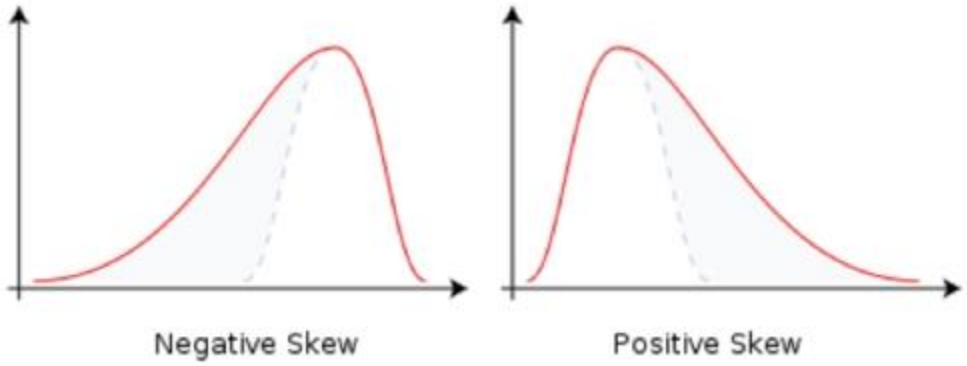
$$\bar{x} = 103020 = \$51.5K$$
$$S_{xx} = 57{,}700 - (1030)^2/20 = 4655$$

Therefore the variance from grouped data is 4655/19=245.52 and the standard deviation is 15.67.

Skewness

When the data in a distribution is tilted to the left of center or to the right of center, then you say that it is *skewed*, namely *skewed to the left (positive skew)* or *skewed to the right*

(negative skew), respectively, as can be seen in the image below.



If the data is not skewed either way, then we call it symmetric, like in the case of the normal distribution depicted above

2.6. MEASURING SIMILARITY AND DISSIMILARITY

Similarity measure

- is a numerical measure of how alike two data objects are.
- higher when objects are more alike.
- often falls in the range [0,1]

Similarity might be used to identify

- duplicate data that may have differences due to typos.
- equivalent instances from different data sets. E.g. names and/or addresses that are the same but have misspellings.
- groups of data that are very close (clusters)

Dissimilarity measure

- is a numerical measure of how different two data objects are
- lower when objects are more alike
- minimum dissimilarity is often 0 while the upper limit varies

Dissimilarity might be used to identify

- outliers
- interesting exceptions, e.g. credit card fraud
- boundaries to clusters

Proximity refers to either a similarity or dissimilarity

Single attribute sim/dissim measures

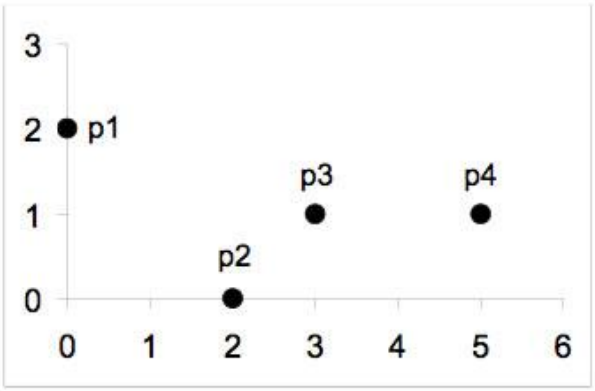
Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \lvert x - y \rvert / (n-1)$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d$, $s = \frac{1}{1+d}$, $s = e^{-d}$, $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Euclidean distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

where *n* is the number of dimensions (attributes) and x_k and y_k are, respectively, the *k-th* attributes (components) or data objects **x** and **y**

Standardization/normalization may be necessary to ensure an attribute does not skew the distances due to different scales.



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

2.6.1. Minkowski Distance

is a generalization of Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

where *r* is a parameter, *n* is the number of dimensions (attributes) and x_k and y_k are, respectively, the *k-th* attributes (components) or data objects **x** and **y.**

Examples

r = 1. City block (Manhattan, taxicab, L_1 norm) distance.

- A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors

r = 2. Euclidean distance (L_2 norm)

$r = \infty$. "supremum" (L_{max} norm, L_∞ norm) distance. This is the maximum difference between any component of the vectors

Do not confuse r with n, i.e., all these distance measures are defined for all numbers of dimensions.

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

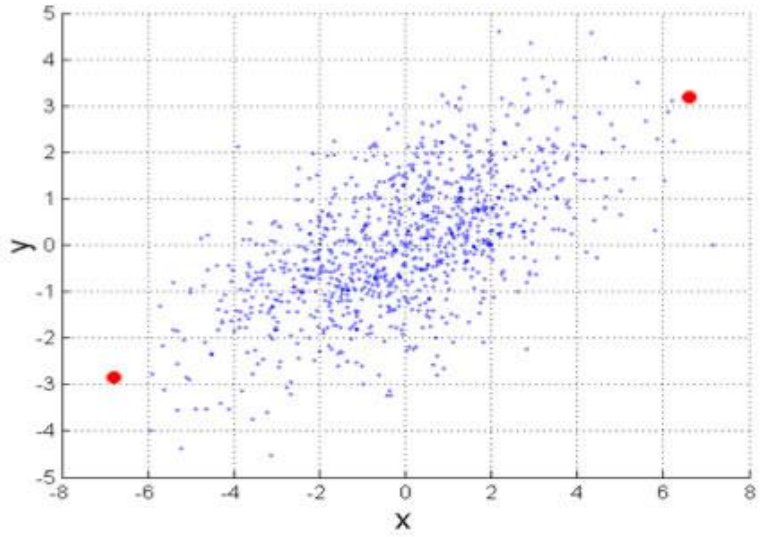
L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distance Matrix

2.6.2. Mahalanobis Distance

Essentially this measures distance from the centroid of the cluster and there's significant correlation among the attributes.

$$\mathbf{mahalanobis}(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T \Sigma^{-1} (\mathbf{x}-\mathbf{y})$$



Σ is the covariance matrix

For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

Common Properties of Distance

Distances, such as the Euclidean distance, have some well known properties.

1. **Positivity:** d(x, y) ≥ 0 for all x and **y**, and d(x, y) = 0 only if **x** = **y**.
2. **Symmetry**: d(x, y) = d(y, x) for all x and **y**.
3. **Triangle Inequality:** d(x, z) ≤ d(x, y) + d(y, z) for all points x, y, and z.

where d(x, y) is the distance (dissimilarity) between points (data objects), x and y.

A distance that satisfies these properties is a **metric**.

Similarity Properties

Similarities, also have some well known properties.

1. s(x, y) = 1 (or maximum similarity) only if x = y (0 ≤ s ≤ 1)
2. s(x, y) = s(y, x) for all x and y. (Symmetry)

where s(x, y) is the similarity between points (data objects), x and y.

Similarity Between Binary Vectors

A common situation is that objects, p and q, have only binary attributes

Compute similarities using the following quantities

f_{01} = the number of attributes where p was 0 and q was 1
f_{10} = the number of attributes where p was 1 and q was 0
f_{00} = the number of attributes where p was 0 and q was 0
f_{11} = the number of attributes where p was 1 and q was 1

Simple Matching and Jaccard Coefficients

SMC = number of matches / number of attributes = $(f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$

J = number of 11 matches / number of non-zero attributes = $(f_{11}) / (f_{01} + f_{10} + f_{11})$

Example SMC v Jaccard

x = 1 0 0 0 0 0 0 0 0 0
y = 0 0 0 0 0 0 1 0 0 1

$f_{01} = 2$ (the number of attributes where p was 0 and q was 1)
$f_{10} = 1$ (the number of attributes where p was 1 and q was 0)
$f_{00} = 7$ (the number of attributes where p was 0 and q was 0)
$f_{11} = 0$ (the number of attributes where p was 1 and q was 1)

$$\text{SMC} = (f_{11} + f_{00}) / (f_{01} + f_{10} + f_{11} + f_{00})$$
$$= (0+7) / (2+1+0+7) = 0.7$$

$$J = (f_{11}) / (f_{01} + f_{10} + f_{11}) = 0 / (2 + 1 + 0) = 0$$

Cosine Similarity

If $\mathbf{d_1}$ and $\mathbf{d_2}$ are two document vectors, then cos($\mathbf{d_1}$, $\mathbf{d_2}$) = <$\mathbf{d_1}$,$\mathbf{d_2}$> / ||$\mathbf{d_1}$|| ||$\mathbf{d_2}$|| , where <$\mathbf{d_1}$,$\mathbf{d_2}$> indicates inner product or vector dot product of vectors, $\mathbf{d_1}$ and $\mathbf{d_2}$, and || d || is the length of vector d.

Example:

$\mathbf{d_1}$ = 3 2 0 5 0 0 0 2 0 0
$\mathbf{d_2}$ = 1 0 0 0 0 0 0 1 0 2

<$\mathbf{d_1}$, $\mathbf{d_2}$> = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5
|| $\mathbf{d_1}$ || = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)$^{1/2}$ = (42) $^{1/2}$ = 6.481
|| $\mathbf{d_2}$ || = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2) $^{1/2}$ = (6)$^{1/2}$ = 2.449

cos($\mathbf{d_1}$, $\mathbf{d_2}$) = 0.3150

Extended Jaccard Coefficient (Tanimoto)

Variation of Jaccard for continuous or count attributes

- Reduces to Jaccard for binary attributes

$$EJ(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x}\cdot\mathbf{y}}$$

Correlation

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x \; s_y}, \quad (2.11$$

where we are using the following standard statistical notation and definitions

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y}) \quad (2.12$$

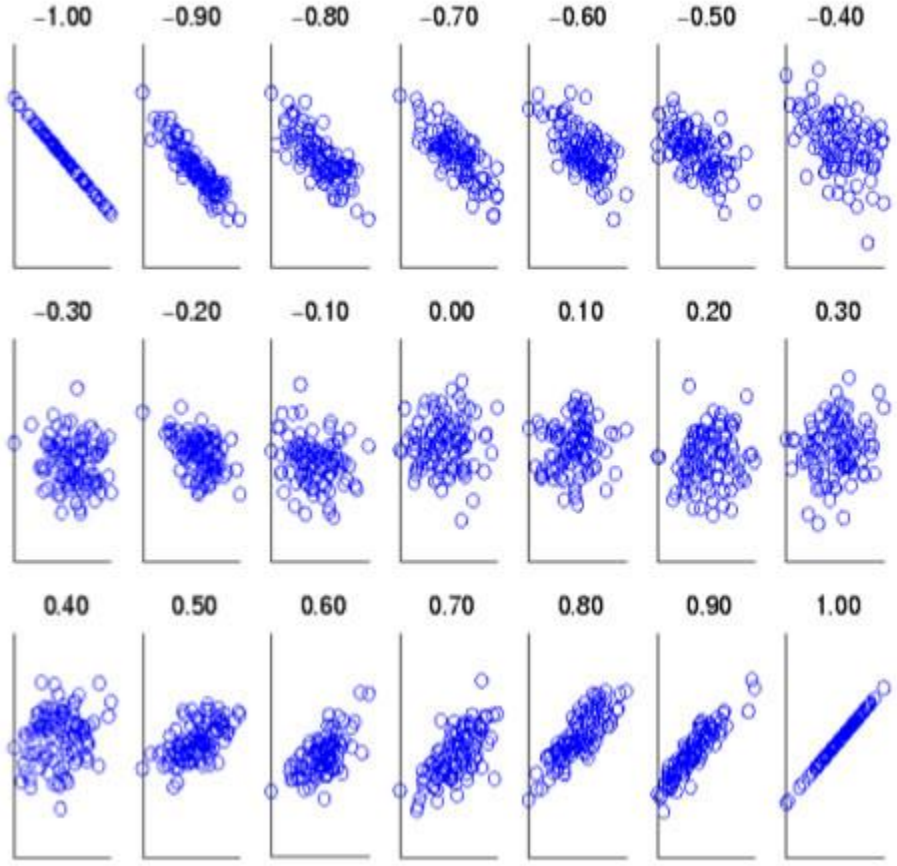
$$\text{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}$$

$$\text{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \overline{y})^2}$$

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k \text{ is the mean of } \mathbf{x}$$

$$\overline{y} = \frac{1}{n} \sum_{k=1}^{n} y_k \text{ is the mean of } \mathbf{y}$$

Visually evaluating correlation



Scatter plots showing the similarity from –1 to 1.

Counterexample of correlation computation

x = (-3, -2, -1, 0, 1, 2, 3)
y = (9, 4, 1, 0, 1, 4, 9)

In this case, notice that $\mathbf{y_i = x_i^2}$ so there's clearly a functional relationship between x and y

$\mu(x) = 0$, $\mu(y) = 4$
$\sigma(x) = 2.16$, $\sigma(y) = 3.74$

But the correlation = (-3)(5)+(-2)(0)+(-1)(-3)+(0)(-4)+(1)(-3)+(2)(0)+3(5) / (6 * 2.16 * 3.74) = 0

Comparison of Proximity Measures

How to choose among the proximity measures?

Domain of application often drives choice

- Similarity measures tend to depend on the type of attribute and data
- Record data, images, graphs, sequences, 3D-protein structure, etc. will use different measures

However, one can talk about various properties that you would like a proximity measure to have

- Symmetry is a common one
- Tolerance to noise and outliers is another
- Ability to find more types of patterns?
- Many others possible

The measure must be applicable to the data and produce results that agree with domain knowledge

---

Information Based Measures

Information theory is a well-developed and fundamental disciple with broad applications

Information relates to possible outcomes of an event

E,g, transmission of a message, flip of a coin, or measurement of a piece of data

The more certain an outcome, the less information that it contains and vice-versa

- For example, if a coin has two heads, then an outcome of heads provides no information

More quantitatively, the information is [inversely] related the probability of an outcome

- The smaller the probability of an outcome, the more information it provides and vice-versa

2.6.3.Entropy

is the commonly used measure for information

For a variable (event), X, with n possible values (outcomes), $\mathbf{x_1, x_2 \ldots, x_n}$
and each outcome having probability, $\mathbf{p_1, p_2 \ldots, p_n}$

the entropy of X , ***H(X)***, is given by

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

$\boldsymbol{0 \le H(X) \le \log_2 n}$ and is measured in **bits**.

Thus, entropy is a measure of how many bits it takes to represent an observation of X on average. Not likely an integer!

Examples

For a coin with probability p for heads and probability q = 1 – p for tails

$\boldsymbol{H = - p \log_2 p - q \log_2 q}$

For a fair coin, p= 0.5, q = 0.5 **H = 1**
For a weighted coin, p = 1 or q = 1, **H = 0**

A more realistic example

Hair Color	**Count**	**P**	**$-p \log_2 p$**
Black	75	0.75	0.3113
Brown	15	0.15	0.4105
Blond	5	0.05	0.2161
Red	0	0.00	0
Other	5	0.05	0.2161
Total	**100**	**1.0**	**1.1540**

Maximum entropy is $\log_2 5 = 2.3219$, that is, we need more than 2 bits for 5 unique values, but not more than 3

In general, suppose we have a number of observations (**m**) of some attribute, X, e.g., the hair color of students in the class,

where there are **n** different possible values.

And the number of observations in the **ith** category is m_i , thus the probability is m_i / m

Then, for this sample, the entropy is

$$H(X) = -\sum_{i=1}^{n} \frac{m_i}{m} \log_2 \frac{m_i}{m}$$

For continuous data, the calculation is harder.

2.6.4.Density

Measures the degree to which data objects are close to each other in a specified area

The notion of density is closely related to that of proximity

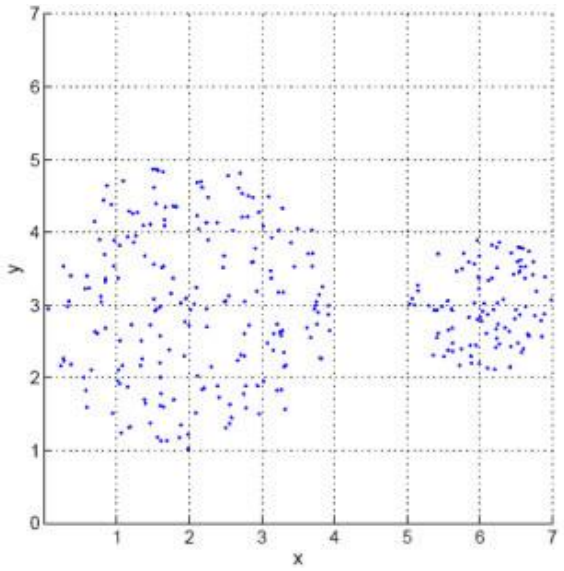
Concept of density is typically used for clustering and anomaly detection

Examples:

- **Euclidean density** = number of points per unit volume
- **Probability density**: Estimate what the distribution of the data looks like
- **Graph-based density**: Connectivity

Euclidean Density: Grid-based Approach

Simplest approach is to divide region into a number of rectangular cells of equal volume,
then define density as # of points the cell contains.



0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

Euclidean density is the number of points within a specified radius of the point

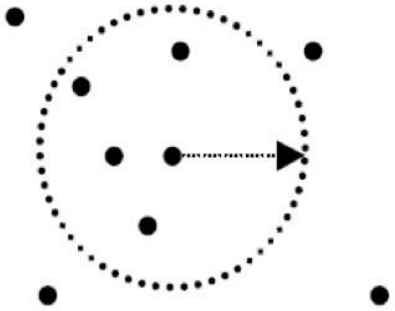


Illustration of center-based density.

2.7. Data Preprocessing

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

Need for Data Preprocessing.

 Data in the real world is dirty. It can be in incomplete, noisy and inconsistent from. These data needs to be preprocessed in order to help improve the quality of the data, and quality of the mining results.

 If no quality data, then no quality mining results. The quality decision is always based on the quality data.

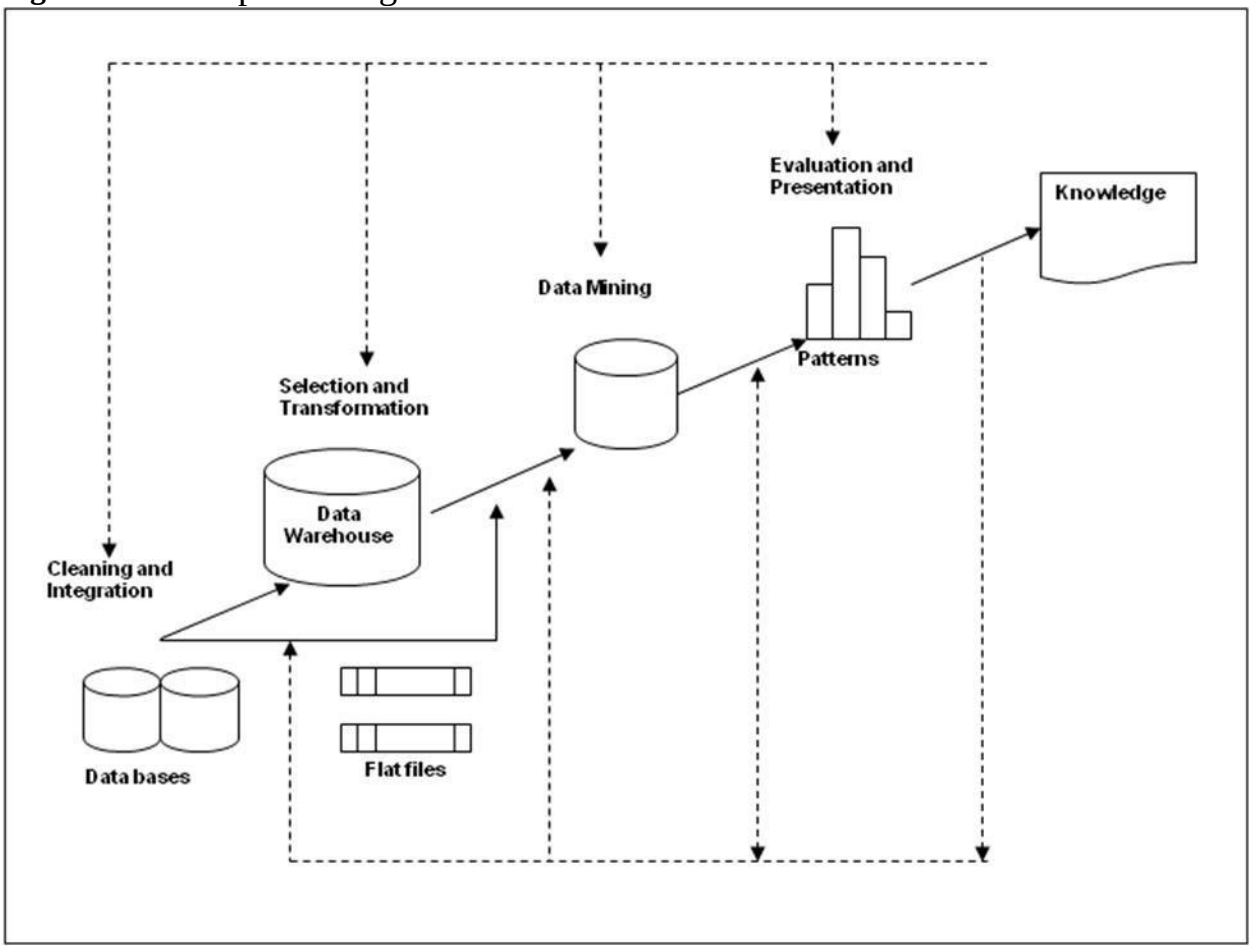
 If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult

 Incomplete data: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. e.g., occupation=“ ”.
Noisy data: containing errors or outliers data. e.g., Salary=“-10”
Inconsistent data: containing discrepancies in codes or names. e.g., Age=“42” Birthday=“03/07/1997”
 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 by 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
 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 transformation
 Normalization and aggregation
 Data reduction
 Obtains reduced representation in volume but produces the same or similar analytical results
 Data discretization
 Part of data reduction but with particular importance, especially for numerical data

2.7.1. Data cleaning

Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

Figure. Data Preprocessing



Various methods for handling this problem:

The various methods for handling the problem of missing values in data tuples include:

(a) Ignoring the tuple: This is usually done when the class label is missing (assuming the mining task involves classification or description). This method is not very effective unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.

(b) Manually filling in the missing value: In general, this approach is time-consuming and may not be a reasonable task for large data sets with many missing values, especially when the value to be filled in is not easily determined.

(c) Using a global constant to fill in the missing value: Replace all missing attribute values by the same constant, such as a label like "Unknown," or $-\infty$. If missing values are replaced by, say, "Unknown," then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common — that of "Unknown." Hence, although this method is simple, it is not recommended.

(d) Using the attribute mean for quantitative (numeric) values or attribute mode for categorical (nominal) values, for all samples belonging to the same class as the given tuple: For example, if classifying customers according to credit risk, replace the missing value with the average income value for customers in the same credit risk category as that of the given tuple.

(e) Using the most probable value to fill in the missing value: This may be determined with regression, inference-based tools using Bayesian formalism, or decision tree induction. For

example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for income.

Noisy data

Noise is a random error or variance in a measured variable. Data smoothing tech is used for removing such noisy data.

Several Data smoothing techniques:

Binning methods: Binning methods smooth a sorted data value by consulting the neighborhood",
or values around it. The sorted values are distributed into a number of 'buckets', or bins. Because binning methods consult the neighborhood of values, they perform local smoothing.

In this technique,

1. The data for first sorted
2. Then the sorted list partitioned into equi-depth of bins.
3. Then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

Smoothing by bin means: Each value in the bin is replaced by the mean value of the bin.

Smoothing by bin medians: Each value in the bin is replaced by the bin median.

Smoothing by boundaries: The min and max values of a bin are identified as the bin boundaries.

Each bin value is replaced by the closest boundary value.

Example: 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(equi depth of 3 since each bin contains three values):

- **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

In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.
For example, the mean of the values 4, 8, and 15 in Bin 1 is 9. Therefore, each original value in this bin is replaced by the value 9. Similarly, smoothing by bin medians can be employed, in which each bin value is replaced by the bin median. In 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.

Suppose that the data for analysis include the attribute age. The age values for the data tuples are (in increasing order): 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

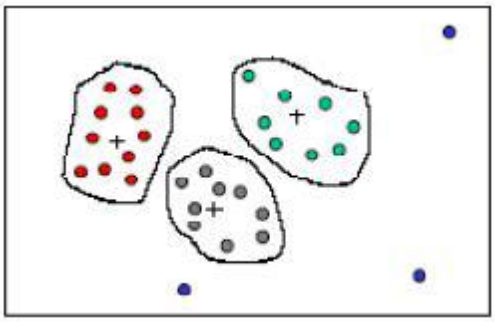
Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data.

The following steps are required to smooth the above data using smoothing by bin means with a bin depth of 3.

• Step 1: Sort the data. (This step is not required here as the data are already sorted.)
• Step 2: Partition the data into equidepth bins of depth 3.
Bin 1: 13, 15, 16 Bin 2: 16, 19, 20 Bin 3: 20, 21, 22
Bin 4: 22, 25, 25 Bin 5: 25, 25, 30 Bin 6: 33, 33, 35
Bin 7: 35, 35, 35 Bin 8: 36, 40, 45 Bin 9: 46, 52, 70
• Step 3: Calculate the arithmetic mean of each bin.
• Step 4: Replace each of the values in each bin by the arithmetic mean calculated for the bin.
Bin 1: 14, 14, 14 Bin 2: 18, 18, 18 Bin 3: 21, 21, 21
Bin 4: 24, 24, 24 Bin 5: 26, 26, 26 Bin 6: 33, 33, 33
Bin 7: 35, 35, 35 Bin 8: 40, 40, 40 Bin 9: 56, 56, 56

Clustering: Outliers in the data may be detected by clustering, where similar values are organized into groups, or 'clusters'. Values that fall outside of the set of clusters may be considered outliers.

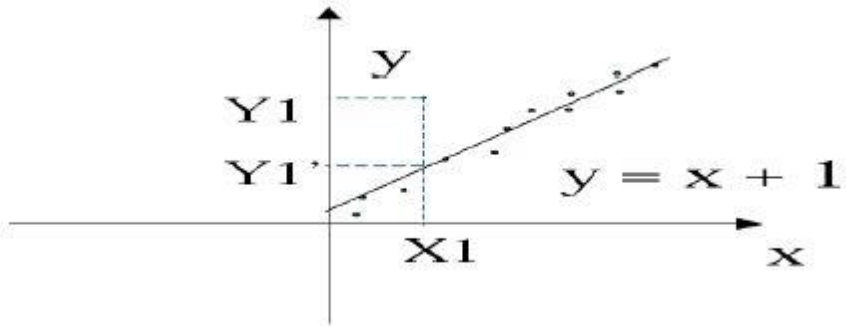
Figure 2.8. Clustering



Regression : smooth by fitting the data into regression functions.

 Linear regression involves finding the best of line to fit two variables, so that one variable can be used to predict the other.

Figure 2.9. Regression



 Multiple linear regression is an extension of linear regression, where more than two variables are involved and the data are fit to a multidimensional surface.

Using regression to find a mathematical equation to fit the data helps smooth out the noise.

Field overloading: is a kind of source of errors that typically occurs when developers compress new attribute definitions into unused portions of already defined attributes.

Unique rule is a rule says that each value of the given attribute must be different from all other values of that attribute

Consecutive rule is a rule says that there can be no missing values between the lowest and highest values of the attribute and that all values must also be unique.

Null rule specifies the use of blanks, question marks, special characters or other strings that may indicate the null condition and how such values should be handled.

2.7.2. Data Integration

It combines data from multiple sources into a coherent store. There are number of issues to consider during data integration.

Issues:

 Schema integration: refers integration of metadata from different sources.

 Entity identification problem: Identifying entity in one data source similar to entity in another table. For example, customer_id in one db and customer_no in another db refer to the same entity

 Detecting and resolving data value conflicts: Attribute values from different sources can be different due to different representations, different scales. E.g. metric vs. British units.

 Redundancy: is another issue while performing data integration. Redundancy can occur due to the following reasons:

 Object identification: The same attribute may have different names in different db

 Derived Data: one attribute may be derived from another attribute.

Handling redundant data in data integration

Correlation analysis

For numeric data Some redundancy can be identified by correlation analysis. The correlation between two variables A and B can be measured by

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

$\overline{A}$, $\overline{B}$ are respective mean values of A and B

σ_A, σ_B are respective standard deviation of A and B

n is the number of tubles

 The result of the equation is > 0, then A and B are positively correlated, which means the value of A increases as the values of B increases. The higher value may indicate redundancy that may be removed.

 The result of the equation is = 0, then A and B are independent and there is no correlation between them.

 If the resulting value is < 0, then A and B are negatively correlated where the values of one attribute increase as the value of one attribute decrease which means each attribute may discourages each other.

-also called Pearson's product moment coefficient

For categorical data

- X^2 (chi-square) test

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

- The larger the X^2 value, the more likely the variables are related
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

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

- X^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

2.7.3.Data Transformation

Data transformation can involve the following:

 Smoothing: which works to remove noise from the data

 Aggregation: where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute weekly and annual total scores.

Generalization of the data

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.

Normalization

Attribute data are scaled so as to fall within a small specified range, such as −1.0 to 1.0, or 0.0 to 1.0.

Attribute construction (feature construction)

This is where new attributes are constructed and added from the given set of attributes to help the mining process.

Normalization

In which data are scaled to fall within a small, specified range, useful for classification

algorithms involving neural networks, distance measurements such as nearest neighbor classification and clustering. There are 3 methods for data normalization. They are:

- min-max normalization
- z-score normalization
- normalization by decimal scaling

Min-max normalization

Performs linear transformation on the original data values. It can be defined as,

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

v is the value to be normalized minA,maxA are minimum and maximum values of an attribute A new_ maxA, new_ minA are the normalization range.

Z-score normalization / zero-mean normalization

The values of an attribute A are normalized based on the mean and standard deviation of A.

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

This method is useful when min and max value of attribute A are unknown or when outliers that are dominate min-max normalization.

Normalization by decimal scaling: normalizes by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of A. A value v of A is normalized to v' by computing

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

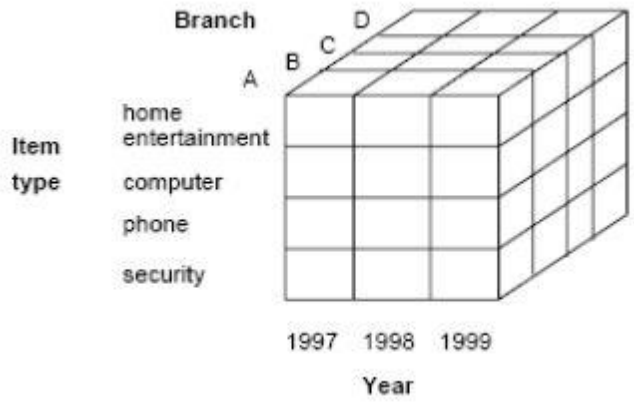
2.7.4. Data Reduction techniques

These techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. Data reduction includes,

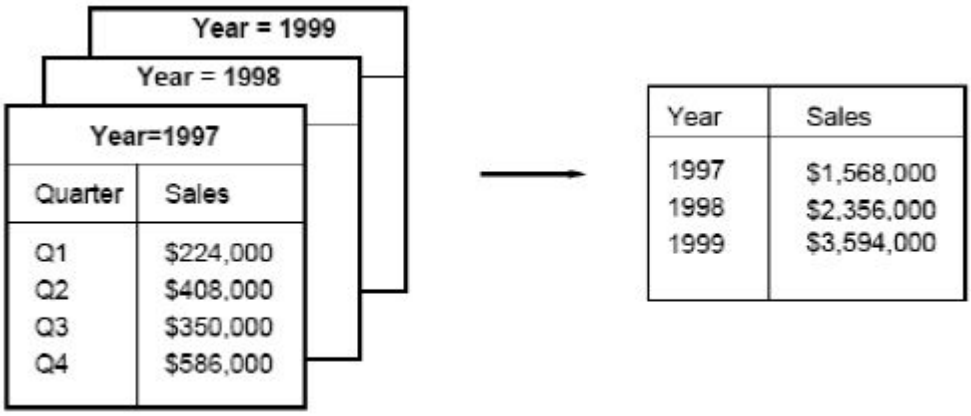
- Data cube aggregation
- Dimension reduction
- Data compression
- Discretization
- Numerosity reduction
  - Regression
  - Histograms
  - Clustering
  - Sampling

Data cube aggregation, where aggregation operations are applied to the data in the construction of a data cube.

Figure 2.10. Data cube



Attribute subset selection, where irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed.
Dimensionality reduction, where encoding mechanisms are used to reduce the data set size. Examples: Wavelet Transforms Principal Components Analysis
Numerosity reduction, where the data are replaced or estimated by alternative, smaller data representations such as parametric models (which need store only the model parameters instead of the actual data) or nonparametric methods such as clustering, sampling, and the use of histograms.
Discretization and concept hierarchy generation, where raw data values for attributes are replaced by ranges or higher conceptual levels. Data discretization is a form of numerosity reduction that is very useful for the automatic generation of concept hierarchies.
Reduce the data to the concept level needed in the analysis. Queries regarding aggregated information should be answered using data cube when possible. Data cubes store multidimensional aggregated information. The following figure shows a data cube for multidimensional analysis of sales data with respect to annual sales per item type for each branch.
Each cell holds an aggregate data value, corresponding to the data point in multidimensional space. Data cubes provide fast access to precomputed, summarized data, thereby benefiting online
analytical processing as well as data mining.
The cube created at the lowest level of abstraction is referred to as the base cuboid. A cube for the highest level of abstraction is the apex cuboid. The lowest level of a data cube (base cuboid). Data cubes created for varying levels of abstraction are sometimes referred to as cuboids, so that a “data cube" may instead refer to a lattice of cuboids. Each higher level of abstraction further reduces the resulting data size.
The following database consists of sales per quarter for the years 1997-1999.
Figure 2.11. Sales per quarter for the years 1997-1999



Suppose, the analyzer interested in the annual sales rather than sales per quarter, the above data can be aggregated so that the resulting data summarizes the total sales per year instead of per quarter. The resulting data in smaller in volume, without loss of information necessary for the analysis task

Dimensionality Reduction

It reduces the data set size by removing irrelevant attributes. This is a method of attribute subset selection are applied.

Attribute sub selection / Feature selection

Feature selection is a must for any data mining product. That is because, when you build a data mining model, the dataset frequently contains more information than is needed to build the model. For example, a dataset may contain 500 columns that describe characteristics of customers, but perhaps only 50 of those columns are used to build a particular model. If you keep the unneeded columns while building the model, more CPU and memory are required during the training process, and more storage space is required for the completed model. In which select a minimum set of features 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

1. Step-wise forward selection: The procedure starts with an empty set of attributes. The best of the original attributes is determined and added to the set. At each subsequent iteration or step, the best of the remaining original attributes is added to the set.

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

3. Combination forward selection and backward elimination: The step-wise forward selection and backward elimination methods can be combined, where at each step one selects the best attribute and removes the worst from among the remaining attributes.

4. Decision tree induction: Decision tree induction constructs a flow-chart-like structure where each internal (non-leaf) 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. The set of attributes appearing in the tree form the reduced subset of attributes.

Forward Selection

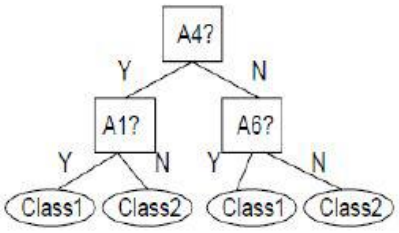
Initial attribute set:
{A1, A2, A3, A4, A5, A6}
Initial reduced set:
{}
-> {A1}
--> {A1, A4}
---> Reduced attribute set:
{A1, A4, A6}

Backward Elimination

Initial attribute set:
{A1, A2, A3, A4, A5, A6}
-> {A1, A3, A4, A5, A6}
--> {A1, A4, A5, A6}
---> Reduced attribute set:
{A1, A4, A6}

Decision Tree Induction

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



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

Greedy (heuristic) methods for attribute subset selection.

Wrapper approach/Filter approach

The mining algorithm itself is used to determine the attribute sub set, then it is called wrapper approach or filter approach. Wrapper approach leads to greater accuracy since it optimizes the evaluation measure of the algorithm while removing attributes.

Data compression

In data compression, data encoding or transformations are applied so as to obtain a reduced or “compressed" representation of the original data. If the original data can be reconstructed from the compressed data without any loss of information, the data compression technique used is called lossless. If, instead, we can reconstruct only an approximation of the original data, then the data compression technique is called lossy. Effective methods of lossy data compression:

 Wavelet transforms

 Principal components analysis.

Wavelet compression is a form of data compression well suited for image compression.

The discrete wavelet transform (DWT) is a linear signal processing technique that, when applied to a data vector D, transforms it to a numerically different vector, D0, of wavelet coefficients.

The general algorithm for a discrete wavelet transform is as follows.

 The length, L, of the input data vector must be an integer power of two. This condition can be met by padding the data vector with zeros, as necessary.

 Each transform involves applying two functions:

 data smoothing

 calculating weighted difference

 The two functions are applied to pairs of the input data, resulting in two sets of data of length L/2.

 The two functions are recursively applied to the sets of data obtained in the previous loop, until the resulting data sets obtained are of desired length.

 A selection of values from the data sets obtained in the above iterations are designated the wavelet coefficients of the transformed data.

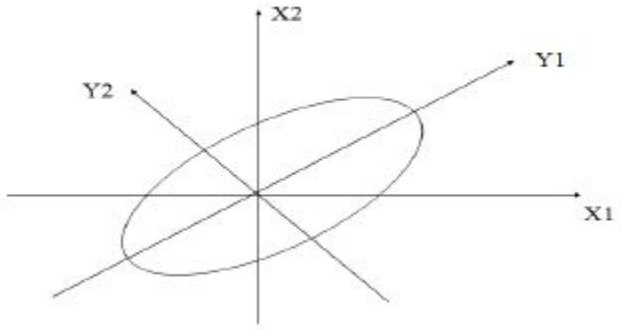
If wavelet coefficients are larger than some user-specified threshold then it can be retained. The remaining coefficients are set to 0. Haar2 and Daubechie4 are two popular wavelet transforms.

Figure 2.12. Wavelet compression

Principal Component Analysis (PCA) is also called as Karhunen-Loeve (K-L) method
Procedure
• Given *N* data vectors from *k*-dimensions, find *c* <= *k* orthogonal vectors that can be best used to represent data
– The original data set is reduced (projected) to one consisting of N data vectors on c principal components (reduced dimensions)
• Each data vector is a linear combination of the *c* principal component vectors
• Works for ordered and unordered attributes
• Used when the number of dimensions is large
The principal components (new set of axes) give important information about variance.

Using the strongest components one can reconstruct a good approximation of the original signal.

Figure 2.13. Principal Component Analysis



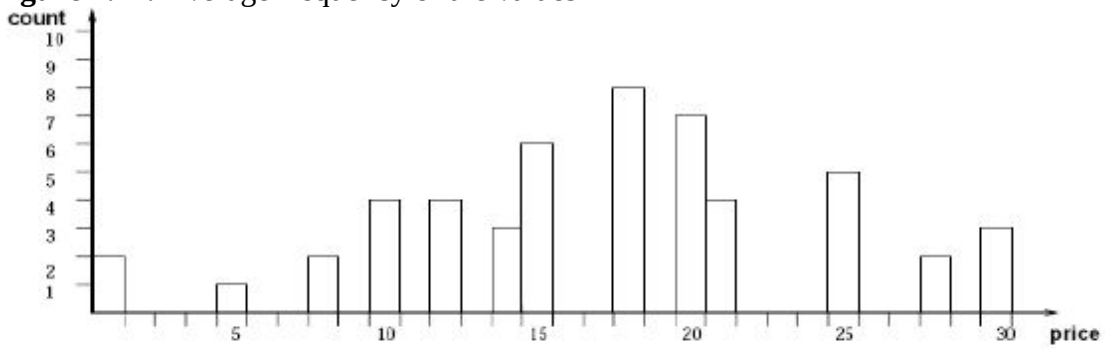
Numerosity Reduction
Data volume can be reduced by choosing alternative smaller forms of data. This tech. can be
 Parametric method
 Non parametric method
Parametric: Assume the data fits some model, then estimate model parameters, and store only the parameters, instead of actual data.
Non parametric: In which histogram, clustering and sampling is used to store reduced form of data.
Numerosity reduction techniques
Regression and log linear models:
 Can be used to approximate the given data
 In linear regression, the data are modeled to fit a straight line using
Y = α + β X, where α, β are coefficients
 Multiple regression: *Y = b0 + b1 X1 + b2 X2.*
– Many nonlinear functions can be transformed into the above.
Log-linear model: The multi-way table of joint probabilities is approximated by a product of

lower-order tables.
Probability: *p(a, b, c, d) = ab ac ad bcd*

Histogram

 Divide data into buckets and store average (sum) for each bucket
 A bucket represents an attribute-value/frequency pair
 It can be constructed optimally in one dimension using dynamic programming
 It divides up the range of possible values in a data set into classes or groups. For each group, a rectangle (bucket) is constructed with a base length equal to the range of values in that specific group, and an area proportional to the number of observations falling into that group.
 The buckets are displayed in a horizontal axis while height of a bucket represents the average frequency of the values.

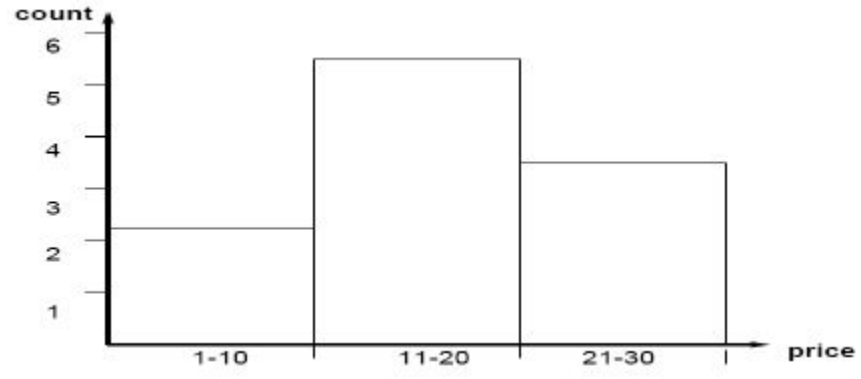
Figure 2.14. Average frequency of the values



Example:

The following data are a list of prices of commonly sold items. The numbers have been sorted.
1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.
Draw histogram plot for price where each bucket should have equi width of 10

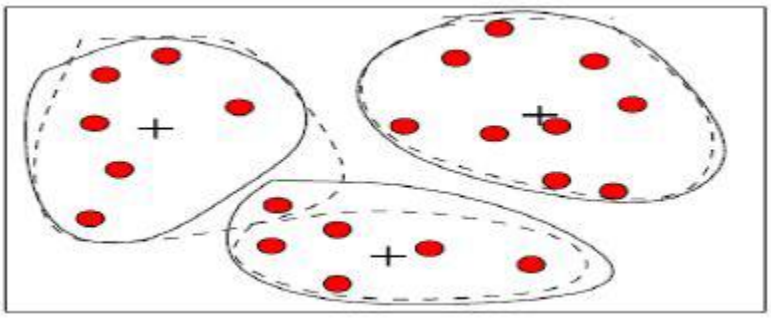
Figure 2.15. Histogram plot for price where each bucket should have equi width of 10



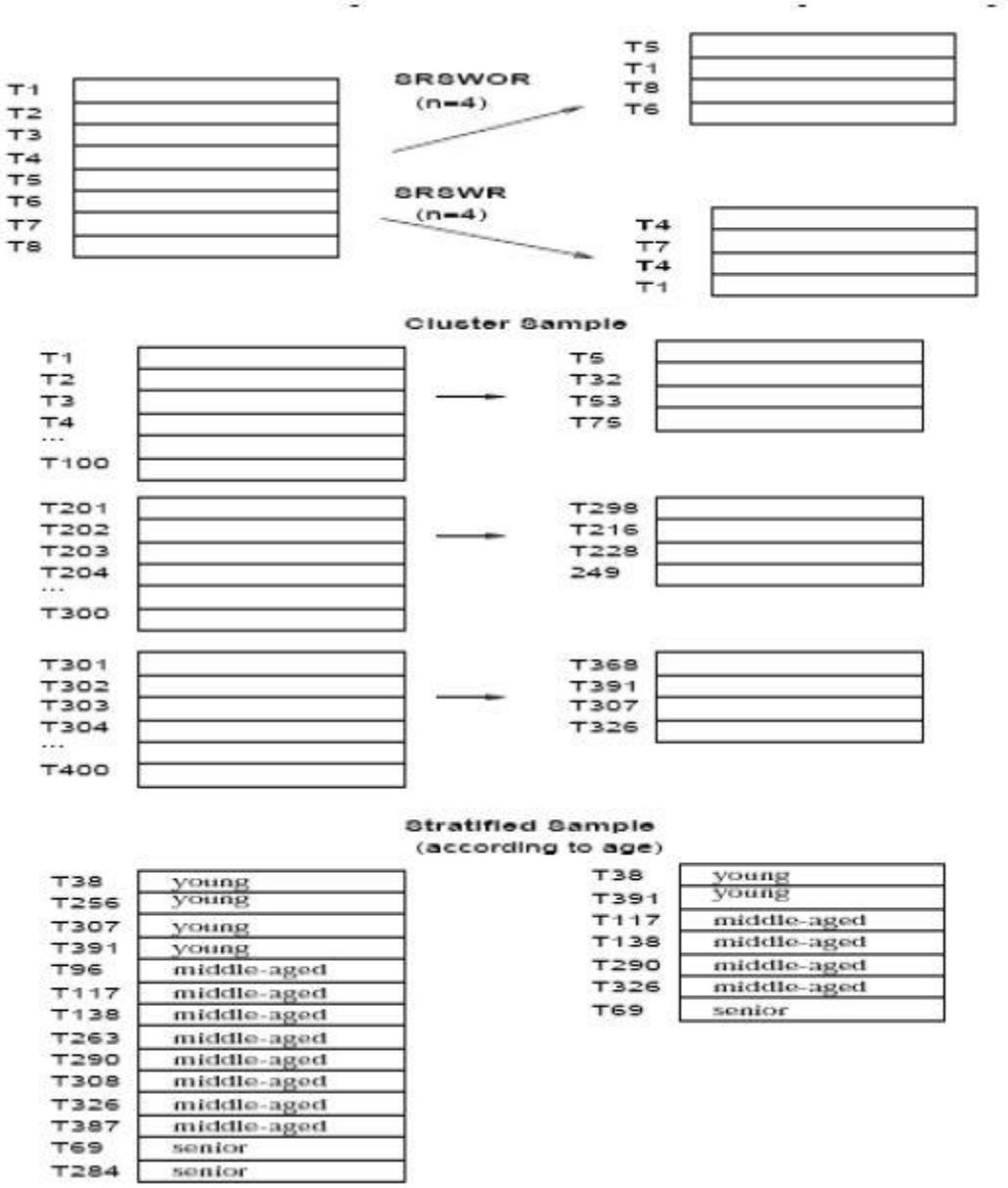
The buckets can be determined based on the following partitioning rules, including the following.

 Equi-width: histogram with bars having the same width
 Equi-depth: histogram with bars having the same height
 V-Optimal: histogram with least variance (countb*valueb)
 MaxDiff: bucket boundaries defined by user specified threshold

V-Optimal and MaxDiff histograms tend to be the most accurate and practical.
Histograms are highly effective at approximating both sparse and dense data, as well as highly skewed, and uniform data.
Clustering techniques consider data tuples as objects. They partition the objects into groups or clusters, so that objects within a cluster are “similar" to one another and “dissimilar" to objects in other clusters. Similarity is commonly defined in terms of how “close" the objects are in space, based on a distance function.
Figure 2.16. Clustering techniques



Quality of clusters measured by their diameter (max distance between any two objects in the cluster) or centroid distance (avg. distance of each cluster object from its centroid)
Sampling
Sampling can be used as a data reduction technique since it allows a large data set to be represented by a much smaller random sample (or subset) of the data. Suppose that a large data set, D, contains N tuples. Let's have a look at some possible samples for D.
Figure 2.17. Sampling



1. Simple random sample without replacement (SRSWOR) of size n: This is created by drawing n of the N tuples from D (n < N), where the probably of drawing any tuple in D is 1=N, i.e., all tuples are equally likely.

2. Simple random sample with replacement (SRSWR) of size n:

This is similar to SRSWOR, except that each time a tuple is drawn from D, it is recorded and then replaced. That is, after a tuple is drawn, it is placed back in D so that it may be drawn again.

3. Cluster sample:

If the tuples in D are grouped into M mutually disjoint “clusters", then a SRS of m clusters can be obtained, where m < M. For example, tuples in a database are usually retrieved a page at a time, so that each page can be considered a cluster. A reduced data representation can be obtained by applying, say, SRSWOR to the pages, resulting in a cluster sample of the tuples.

4. Stratified sample

If D is divided into mutually disjoint parts called “strata", a stratified sample of D is

generated by obtaining a SRS at each stratum. This helps to ensure a representative sample, especially when the data are skewed. For example, a stratified sample may be obtained from customer data, where stratum is created for each customer age group. In this way, the age group having the smallest number of customers will be sure to be represented.

Advantages of sampling

1. An advantage of sampling for data reduction is that the cost of obtaining a sample is proportional to the size of the sample, n, as opposed to N, the data set size. Hence, sampling complexity is potentially sub-linear to the size of the data.
2. When applied to data reduction, sampling is most commonly used to estimate the answer to an aggregate query.

2.8. Discretization and concept hierarchies

Discretization techniques can be used to 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.

Concept Hierarchy

A concept hierarchy for a given numeric attribute defines a discretization of the attribute. Concept hierarchies can be used to reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute age) by higher level concepts (such as young, middle-aged, or senior).

Discretization and Concept hierarchy for numerical data

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
There are five methods for numeric concept hierarchy generation. These include:

- binning,
- histogram analysis,
- clustering analysis,
- entropy-based discretization

Procedure:

- 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

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

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

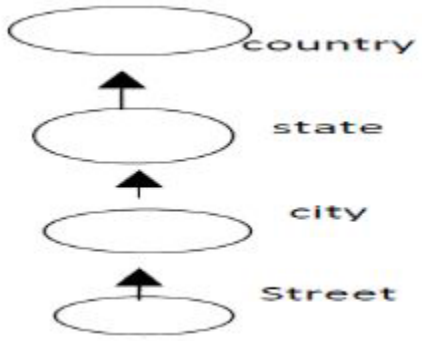
$$Entropy(S_1) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

where p_i is the probability of class i in S_1

- 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
- Such a boundary may reduce data size and improve classification accuracy
  - A simply 3-4-5 rule can be used to segment numeric data into relatively uniform, "natural" intervals.
    - If an interval covers 3, 6, 7 or 9 distinct values at the most significant digit, partition the range into 3 equi-width intervals
    - If it covers 2, 4, or 8 distinct values at the most significant digit, partition the range into 4 intervals
    - If it covers 1, 5, or 10 distinct values at the most significant digit, partition the range into 5 intervals

Concept hierarchy generation for category data

Figure 2.18. A concept hierarchy



A concept hierarchy defines a sequence of mappings from set of low-level concepts to higher-level, more general concepts. It organizes the values of attributes or dimension into gradual levels of abstraction. They are useful in mining at multiple levels of abstraction

Commercial tools available to find data discrepancy detection:

Data scrubbing tools use simple domain knowledge to detect errors and make corrections in the data

Data auditing tools find discrepancies by analyzing the data to discover rules and relationships and detecting data that violate such conditions

Data migration tools allow simple transformation to be specified such as replace the string "gender" by "sex"

ETL(Extraction/Transformation/Loading tools: allow users to specify transforms through a graphical user interface.

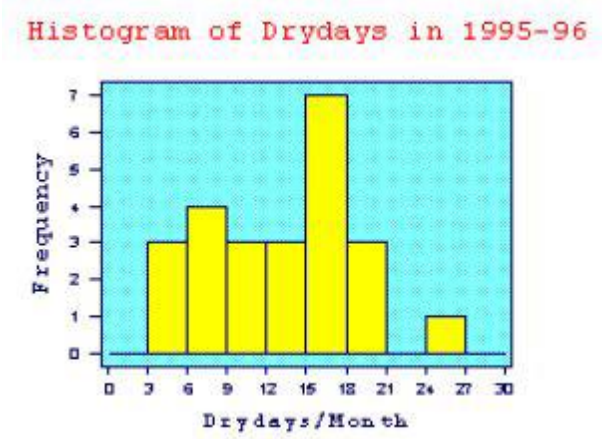
Graphic Displays of Basic Descriptive Data Summaries

2.8.1.Histogram

A histogram is a way of summarizing data that are measured on an interval scale (either discrete or continuous). It is often used in exploratory data analysis to illustrate the major features of the distribution of the data in a convenient form. It divides up the range of possible values in a data set into classes or groups. For each group, a rectangle is constructed with a base length equal to the range of values in that specific group, and an area proportional to the number of observations falling into that group. This means that the rectangles might be drawn of non

uniform height. The histogram is only appropriate for variables whose values are numerical and measured on an interval scale. It is generally used when dealing with large data sets (>100 observations) A histogram can also help detect any unusual observations (outliers), or any gaps in the data set.

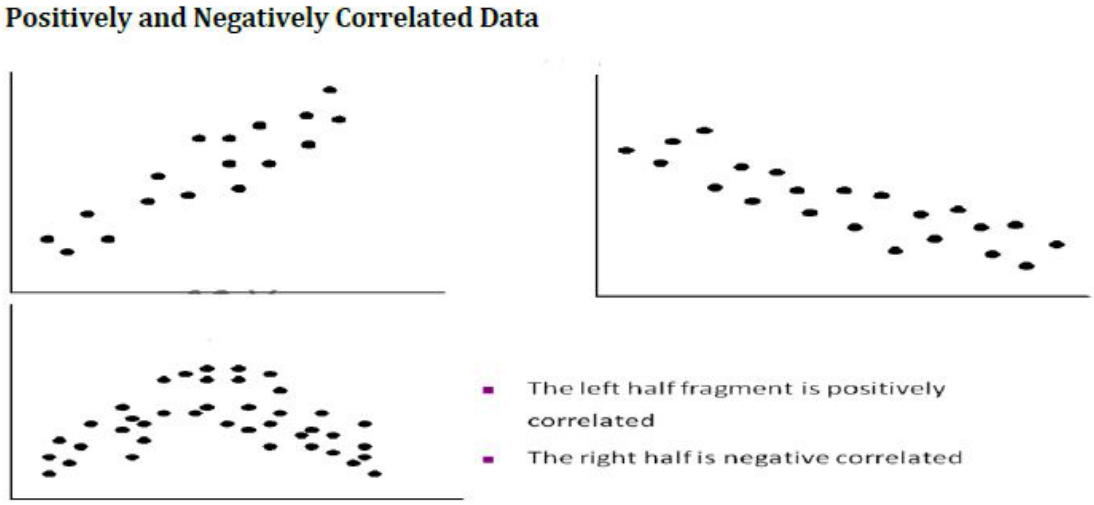
Figure 2.19. Histogram



Scatter Plot

A scatter plot is a useful summary of a set of bivariate data (two variables), usually drawn before working out a linear correlation coefficient or fitting a regression line. It gives a good visual picture of the relationship between the two variables, and aids the interpretation of the correlation coefficient or regression model. Each unit contributes one point to the scatter plot, on which points are plotted but not joined. The resulting pattern indicates the type and strength of the relationship between the two variables.

Figure 2.20. Scatter Plot



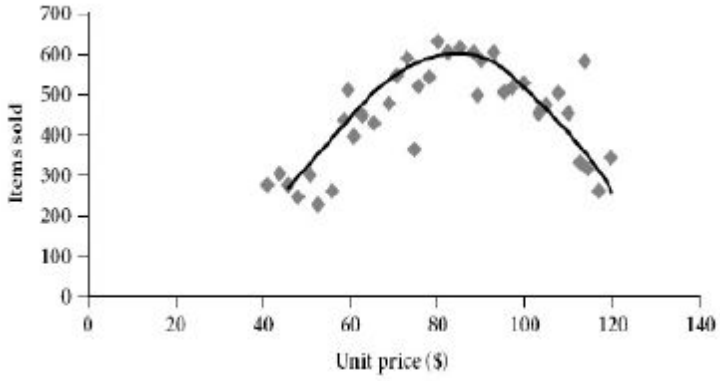
Positively and Negatively Correlated Data

A scatter plot will also show up a non-linear relationship between the two variables and whether or not there exist any outliers in the data.

Loess curve

It is another important exploratory graphic aid that adds a smooth curve to a scatter plot in order to provide better perception of the pattern of dependence. The word loess is short for "local regression."

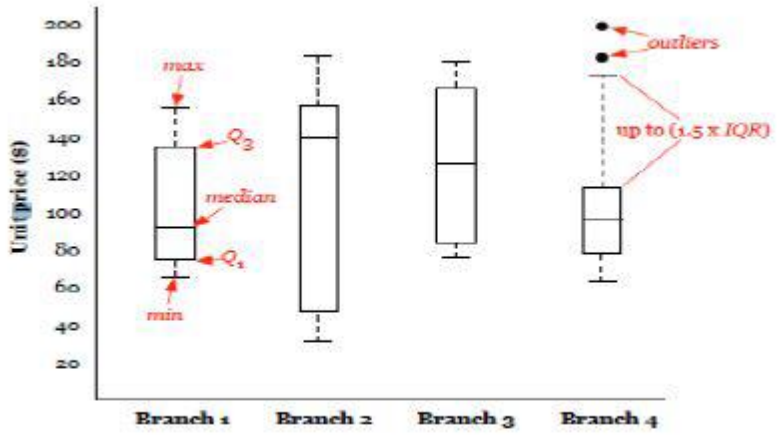
Figure 2.21. Loess curve



Box plot

The picture produced consists of the most extreme values in the data set (maximum and minimum values), the lower and upper quartiles, and the median.

Figure 2.22. Box plot



Quantile plot

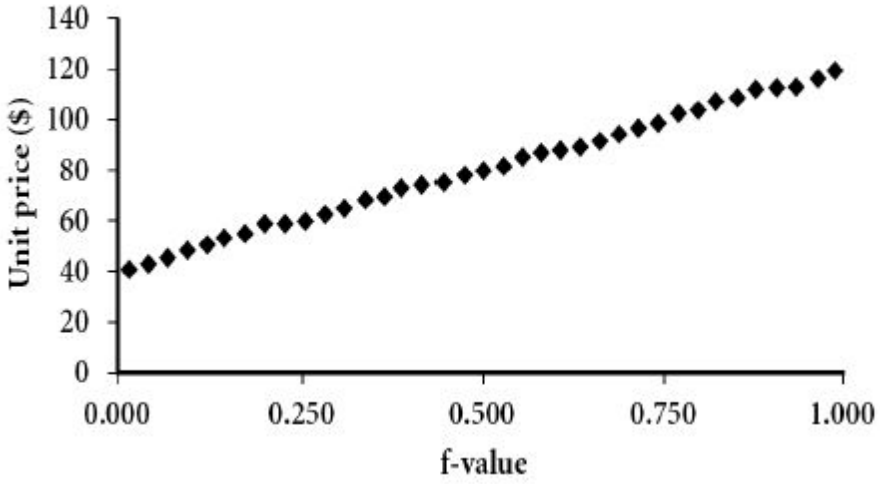
 Displays all of the data (allowing the user to assess both the overall behavior and unusual occurrences)

 Plots quantile information

 For a data *xi* data sorted in increasing order, *fi* indicates that approximately 100 *fi*% of the data are below or equal to the value *xi*

The *f* quantile is the data value below which approximately a decimal fraction *f* of the data is found. That data value is denoted q(f). Each data point can be assigned an *f*-value. Let a time series *x* of length *n* be sorted from smallest to largest values, such that the sorted values have rank. The *f*-value for each observation is computed as . 1,2,..., n . The *f*-value for each observation is computed as,

Figure 2.23. Quantile plot



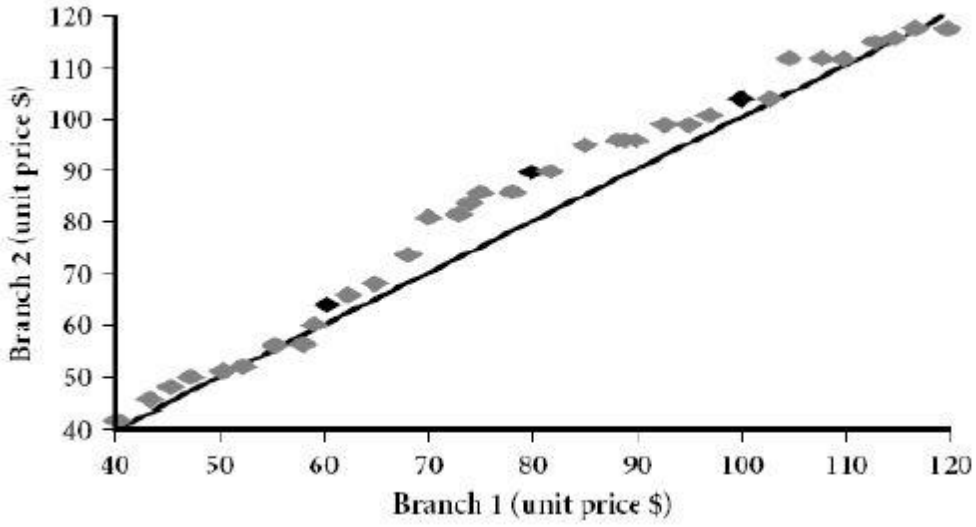
Quantile-Quantile plots (Q-Q plot)

Quantile-quantile plots allow us to compare the quantiles of two sets of numbers.This kind of comparison is much more detailed than a simple comparison of means or medians. A normal distribution is often a reasonable model for the data. Without inspecting the data, however, it is risky to assume a normal distribution. There are a number of graphs that can be used to check the deviations of the data from the normal distribution. The most useful tool for assessing normality is a quantile quantile or QQ plot. This is a scatterplot with the quantiles of the scores on the horizontal axis and the expected normal scores on the vertical axis.
It is a graph that shows the quantiles of one univariate distribution against the corresponding quantiles of another. It is a powerful visualization tool in that it allows the user to view whether there is a shift in going from one distribution to another.

The steps in constructing a QQ plot are as follows:

First, we sort the data from smallest to largest. A plot of these scores against the expected normal scores should reveal a straight line. The expected normal scores are calculated by taking the z-scores of (I - ½)/n where I is the rank in increasing order.
Curvature of the points indicates departures of normality. This plot is also useful for detecting outliers.

Figure 2.24. QQ plot



The outliers appear as points that are far away from the overall pattern op points
How is a quantile-quantile plot different from a quantile plot?
A quantile plot is a graphical method used to show the approximate percentage of values below or equal to the independent variable in a univariate distribution. Thus, it displays quantile information for all the data, where the values measured for the independent variable are plotted against their corresponding quantile.
A quantile-quantile plot however, graphs the quantiles of one univariate distribution against the corresponding quantiles of another univariate distribution. Both axes display the range of values measured for their corresponding distribution, and points are plotted that correspond to the quantile values of the two distributions. A line (y = x) can be added to the graph along with points representing where the first, second and third quantiles lie, in order to

increase the graph's informational value. Points that lie above such a line indicate a correspondingly higher value for the distribution plotted on the y-axis, than for the distribution plotted on the x-axis at the same quantile. The opposite effect is true for points lying below this line.

2.9. QUESTION BANK

PART A

1. Define data preprocessing. (April 2012)
2. What is meant by concept description?
3. Define data cleaning.
4. Define data reduction.
5. Define data transformation.
6. Define data integration.
7. What is binning? (Nov 2013)
8. How can correlation between two attributes be measured?
9. Define data aggregation.
10. What is meant by dimensionality reduction?
11. What are the primitives that define a task?
12. What is boxplot analysis?
13. What is DMQL? (Nov 2012)
14. List the different coupling schemes used in a data mining system.
15. What is entropy of a attribute?
16. Give the formula for mean, median and mode.
17. What is a loess curve?
18. What is a scatter plot?
19. What is a q-q plot?

PART B

1. Explain briefly about data cleaning. (April 2011)
2. Write short notes on data integration and transformation. (Nov 2013)
3. Briefly explain data cube aggregation and dimensionality reduction.
4. How is a class comparison performed? Describe the procedure with example.
5. Describe why concept hierarchies are useful in data mining. (Nov 2012)
6. List and describe the five primitives for specifying data mining task. (Nov 2010)

2.10. REFERENCE.

1. dais.cs.uiuc.edu/education/cs312.html