

Data Mining:

Concepts and Techniques

— Chapter 2 —

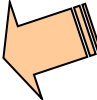
Jiawei Han, Micheline Kamber, and Jian Pei

University of Illinois at Urbana-Champaign

Simon Fraser University

©2013 Han, Kamber, and Pei. All rights reserved.

Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types 
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary

Types of Data Sets

- Record

- Relational records
- Data matrix, e.g., numerical matrix, crosstabs
- Document data: text documents: term-frequency vector
- Transaction data

- Graph and network

- World Wide Web
- Social or information networks
- Molecular Structures

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

- Ordered

- Video data: sequence of images
- Temporal data: time-series
- Sequential Data: transaction sequences
- Genetic sequence data

- Spatial, image and multimedia:

- Spatial data: maps
- Image data:
- Video data:

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Important Characteristics of Structured Data

- Dimensionality
 - Curse of dimensionality
- Sparsity
 - Only presence counts
- Resolution
 - Patterns depend on the scale
- Distribution
 - Centrality and dispersion

Data Objects

- Data sets are made up of data objects.
- A **data object** represents an entity.
- Examples:
 - sales database: customers, store items, sales
 - medical database: patients, treatments
 - university database: students, professors, courses
- Also called *samples* , *examples*, *instances*, *data points*, *objects*, *tuples*.
- Data objects are described by **attributes**.
- Database rows -> data objects; columns -> attributes.

Attributes

- **Attribute (or dimensions, features, variables):** a data field, representing a characteristic or feature of a data object.
 - *E.g., customer_ID, name, address*
- **Types:**
 - Nominal
 - Binary
 - Ordinal
 - Numeric: quantitative
 - Interval-scaled
 - Ratio-scaled

Attribute Types

- **Nominal:**
- Nominal means “relating to names.”
- The values of a **nominal attribute** are symbols or “*names of things*”.
- Each value represents some kind of category, code, or state.
- So nominal attributes are also referred to as **categorical**.
- The values do not have any meaningful order.
 - *Hair_color* = { *black, brown, grey, red, white* }
 - *Occupation* = { *teacher, dentist, programmer, farmer* }
- It is possible to represent the values of as symbols with numbers.
 - With *hair color*, we can assign a code of 0 for *black*, 1 for *brown*, and so on.
 - Another example is *customer ID*, with possible values that are all numeric.
 - In such cases, the numbers are not intended to be used quantitatively.
- Mathematical operations on values of nominal attributes are not meaningful.
- A nominal attribute may have integers as values, it is not considered as a numeric attribute because the integers are not meant to be used quantitatively.

Attribute Types

■ Binary

- Nominal attribute with only 2 states (0 and 1)
- Binary attributes are referred to as Boolean if the two states correspond to *true* and *false*.
- Symmetric binary:
 - its states are equally valuable and carry the same weight
 - There is no preference on which outcome should be coded as 0 or 1.
 - e.g., gender
- Asymmetric binary:
 - The outcomes of the states are not equally important,
 - We code the most important outcome, which is usually the rarest one, by 1 and the other by 0.
 - e.g., medical test (positive vs. negative)
 - Convention: assign 1 to most important outcome (e.g., HIV positive)

Attribute Types

■ Ordinal

- An attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known.
- $Size = \{small, medium, large\}$
- $Grade = (e.g., A^+, A, A^-, B^+, \text{and so on})$
- Ordinal attributes are useful for registering subjective assessments of qualities.
- Cannot be measured objectively.
- Ordinal attributes are often used in surveys for ratings.

■ Nominal, binary, and ordinal attributes are *qualitative*.

■ They *describe* a feature of an object without giving an actual size or quantity.

■ The values of such qualitative attributes are typically words representing categories.

Numeric Attribute Types

- A **numeric attribute** is *quantitative*.
- It is a measurable quantity, represented in integer or real values.
- Numeric attributes can be *interval-scaled* or *ratio-scaled*.
- **Interval-scaled**
 - Measured on a scale of **equal-sized** units.
 - The values have order and can be positive, 0, or negative.
 - provides a ranking of values, Compare and quantify the difference between values.
 - The outdoor temperature value for a number of different days.
 - By ordering the values, we obtain a ranking of the objects with respect to temperature.
 - We can quantify the difference between values.
 - For example, a temperature of 20° C is five degrees higher than a temperature of 15°C.

Numeric Attribute Types

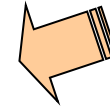
- Calendar dates are another example. For instance, the years 2002 and 2010 are eight years apart.
- Temperatures in Celsius and Fahrenheit do not have a true zero-point, that is, neither 0°C nor 0° indicates “no temperature.”
- **Ratio-scaled**
 - Inherent **zero-point**
 - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K° is twice as high as 5 K°).
 - e.g., *temperature in Kelvin, length, counts, monetary quantities*

Discrete vs. Continuous Attributes

- Classification algorithms developed often talk of attributes as being either *discrete* or *continuous*.
- **Discrete Attribute**
 - Has only a finite or countably infinite set of values
 - E.g., zip codes, profession, or the set of words in a collection of documents
 - Sometimes, represented as integer variables
 - Note: Binary attributes are a special case of discrete attributes
- **Continuous Attribute**
 - Has real numbers as attribute values
 - E.g., temperature, height, or weight
 - Practically, real values can only be measured and represented using a finite number of digits
 - Continuous attributes are typically represented as floating-point variables

Chapter 2: Getting to Know Your Data

- Data Objects and Attribute Types
- Basic Statistical Descriptions of Data
- Data Visualization
- Measuring Data Similarity and Dissimilarity
- Summary



Basic Statistical Descriptions of Data

- Motivation
 - To better understand the data: central tendency, variation and spread
- Data dispersion characteristics
 - median, max, min, quantiles, outliers, variance, etc.
- Numerical dimensions correspond to sorted intervals
 - Data dispersion: analyzed with multiple granularities of precision
 - Boxplot or quantile analysis on sorted intervals
- Dispersion analysis on computed measures
 - Folding measures into numerical dimensions
 - Boxplot or quantile analysis on the transformed cube

Measuring the Central Tendency

- Various ways to measure the central tendency of data.
- We have some attribute X , like *salary*, which has been recorded for a set of objects.
- Let x_1, x_2, \dots, x_N be the set of N observed values or *observations* for X .
- These values may also be referred to as the data set.
- If we were to plot the observations for *salary*, where would most of the values fall?
- This gives us an idea of the central tendency of the data.
- Measures of central tendency include the ***mean, median, mode, and midrange.***

MEAN

- The most common and effective numeric measure of the “center” of a set of data is the *(arithmetic) mean*.
- Let x_1, x_2, \dots, x_N be a set of N values or *observations*, such as for some numeric attribute X , like *salary*.
- The **mean** of this set of values is

$$\bar{x} = \frac{1}{N} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_N}{N}$$

MEAN

- Sometimes, each value x_i in a set may be associated with a weight w_i for $i = 1, \dots, N$.
- The weights reflect the significance, importance, or occurrence frequency attached to their respective values.
- In this case, we can compute

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i} = \frac{w_1 x_1 + w_2 x_2 + \dots + w_N x_N}{w_1 + w_2 + \dots + w_N}$$

- This is called the **weighted arithmetic mean** or the **weighted average**.

MEAN

- A major problem with the mean is its sensitivity to extreme (e.g., outlier) values.
- Even a small number of extreme values can corrupt the mean.
 - For example, the mean salary at a company may be substantially pushed up by that of a few highly paid managers.
 - Similarly, the mean score of a class in an exam could be pulled down quite a bit by a few very low scores.
- To offset the effect caused by a small number of extreme values, we can instead use the **trimmed mean**.
- which is the mean obtained after chopping off values at the high and low extremes.
 - For example, we can sort the values observed for *salary* and remove the top and bottom 2% before computing the mean.
 - We should avoid trimming too large a portion (such as 20%) at both ends, as this can result in the loss of valuable information.

MEDIAN

- The data are already sorted in increasing order.
- If there is an even number of observations (i.e., 12); the median is not unique.
- Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.
- It can be any value within the two middlemost values of 52 and 56.
- By convention, we assign the average of the two middlemost values as the median; that is, $(52+56) / 2 = 54$.
- The median is \$54,000.
- Suppose that we had only the first 11 values in the list. Given an odd number of values, the median is the middlemost value. This is the sixth value in this list, which has a value of \$52,000.
- The median is expensive to compute when we have a large number of observations.
- For numeric attributes, however, we can easily *approximate* the value.

MEDIAN

- If that data are grouped in intervals according to their x_i data values and that the frequency of each interval is known.
 - For example, employees may be grouped according to their annual salary in intervals such as \$10–20,000, \$20–30,000, and so on.
 - Let the interval that contains the median frequency be the *median interval*.
- We can approximate the median of the entire data set (e.g., the median salary) by interpolation using the formula

$$median = L_1 + \left(\frac{n/2 - (\sum freq)_l}{freq_{median}} \right) width$$

- where L_1 is the lower boundary of the median interval.
- N is the number of values in the entire data set.
- $(\sum freq)_l$ is the sum of the frequencies of all of the intervals that are lower than the median interval.
- $freq_{median}$ is the frequency of the median interval.
- $width$ is the width of the median interval.

MODE

- The *mode* is another measure of central tendency.
- The **mode** for a set of data is the value that occurs most frequently in the set.
- Therefore, it can be determined for qualitative and quantitative attributes.
- It is possible for the greatest frequency to correspond to several different values, which results in more than one mode.
- Data sets with one, two, or three modes are respectively called **unimodal**, **bimodal**, and **trimodal**.
- A data set with two or more modes is **multimodal**.
- If each data value occurs only once, then there is no mode.
- Suppose we have the following values for *salary* (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.
- The two modes are \$52,000 and \$70,000.

MIDRANGE

- The **midrange** can also be used to assess the central tendency of a numeric data set.
- It is the average of the largest and smallest values in the set.
- This measure is easy to compute using the SQL aggregate functions, `max()` and `min()`.
- The midrange of the data of Example is $(30,000 + 110,000) / 2 = \$70,000$.
- In a unimodal frequency curve with perfect **symmetric** data distribution, the mean, median, and mode are all at the same center value.
- Data in most real applications are not symmetric.
- They may instead be either **positively skewed**, where the mode occurs at a value that is smaller than the median or **negatively skewed**, where the mode occurs at a value greater than the median.

Symmetric vs. Skewed Data

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

