Knowledge Discovery and Data Mining

Spring 2021

Chap 2. Getting to Know Your Data

Jiawei Han, Micheline Kamber and Jian Pei, Data Mining: Concepts and Techniques, 3rd ed., The Morgan Kaufmann Series in Data Management Systems Morgan Kaufmann Publishers, July 2011. ISBN 978-0123814791

Tuong Le, PhD

Outline

1. Data Objects and Attribute Types

2. Basic Statistical Descriptions of Data

3. Data Visualization

4. Measuring Data Similarity and Dissimilarity

5. Summary

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**:
  - **Qualitative**: Nominal, Binary, Ordinal.
  - **Numeric**: quantitative
    - Interval-scaled
    - Ratio-scaled

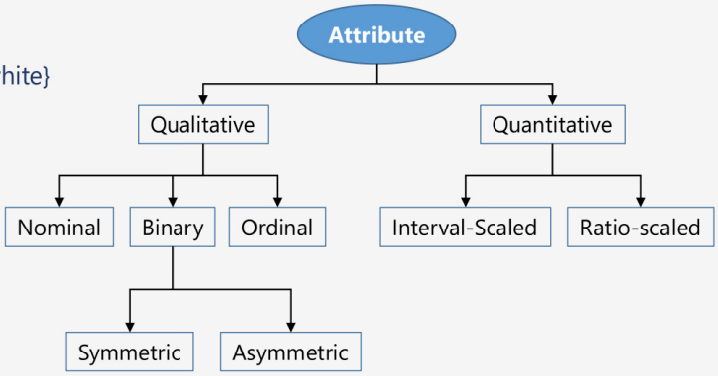
Attributes

Data objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Attribute Types

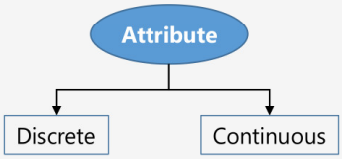
- **Nominal**: categories, states, or "names of things"
  - Hair_color = {auburn, black, blond, brown, grey, red, white}
  - Marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., Medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - Size = {small, medium, large}, grades, army rankings



Numeric (Quantity) Attribute Types

- **Interval-scaled**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., temperature in C˚ or F˚, calendar dates
  - No true zero-point
- **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 (Machine learning)



❑ **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

Outline

1. Data Objects and Attribute Types

2. Basic Statistical Descriptions of Data

3. Data Visualization

4. Measuring Data Similarity and Dissimilarity

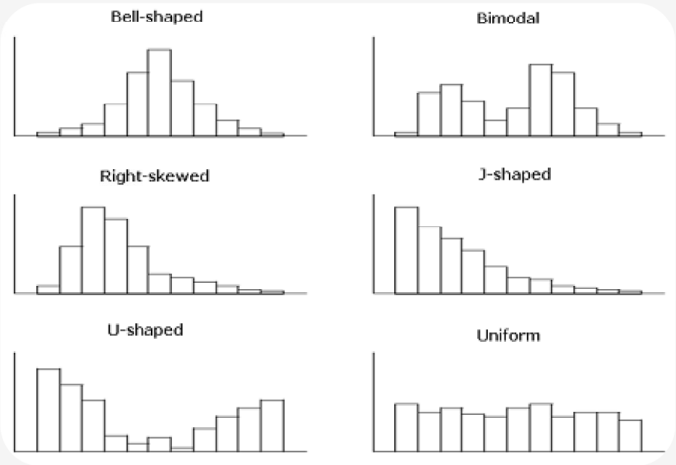
5. Summary

Basic Statistical Descriptions of Data

- **Motivation:** For data preprocessing to be successful, it is essential to have an overall picture of your data. Basic statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.
- **Measures of central tendency:** measure the location of the middle or center of a data distribution.
- **Measuring the Dispersion of Data**: assessing the dispersion or spread of numeric data
  - Range, quantiles, quartiles, percentiles, and the interquartile range
  - The five-number summary
  - Variance and standard deviation
- **Graphic Displays of Basic Statistical Descriptions of Data**
  - Quantile plots, quantile–quantile plots, histograms, and scatter plots

Measuring the Central Tendency

- Measuring the location of the middle or center of a data distribution.
- **Mean** (algebraic measure): mean and weighted arithmetic mean.
- **Median**
- **Mode**
- **Symmetric vs Skewed Data**



Mean (algebraic measure)

- Let $x_1, x_2, \ldots, x_n$ be a set of N values or observations, such as for the numeric attribute X. The mean of this set of values is

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

- Sometimes, each value x_i in a set may be associated with a weight w_i for $i = 1, \ldots, \mathrm{N}$. The weights reflect the significance, importance, or occurrence frequency attached to their respective values. In this case, we can compute the weighted arithmetic mean or the weighted average as follows.

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

Mean: Example

- **Example 1:** 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. Using formula, we have:

$$\bar{x} = \frac{30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12} = 58$$

  - The mean salary is $58,000.

- **Example 2:** According to the journal Chemical engineering, an important property of fiber is its water absorbency. A random sample of 19 pieces of cotton fiber was taken and the absorbency on each piece was measured. The following are the absorbency values: 18.71, 21.41, 20.72, 21.81, 19.29, 22.43, 20.17, 23.71, 19.44, 20.50, 18.92, 20.33, 22.85, 19.25, 21.77, 22.11, 19.77, 18.04, 21.12. Calculate the sample mean and median and median for the above sample values?

$$\bar{x} = \frac{1}{n}\sum_{1}^{n} x_i = \frac{392.35}{19} = 20.65$$

Median

- Middle value if odd number of values, or average of the middle two values otherwise
  - **Example**: 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 median is: $\frac{52+56}{2} = 54$. If the list is 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70. The median is 52.
- When a large number of observations, approximate median

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

- 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.

Median: Example

- Calculate Median for the following data

Age	1-5	6-15	16-20	21-50	51-80	81-110
Frequency	200	450	300	1500	700	44

- **Solution:**

Age	Frequency	$(\sum freq)_l$
1-5	200	450
6-15	450	650
16-20	300	950
21-50	1500	2450
51-80	700	3150
81-110	44	3194
	N = 3194	

The Median is the mean of the 1597th and 1598th , so is in the group 21-50.

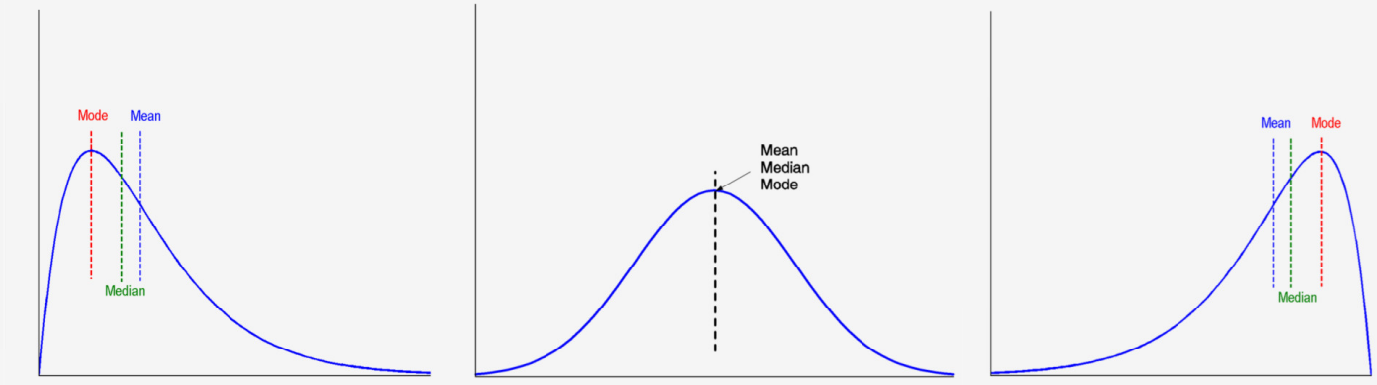
- L_1 =20,5 (the lower class boundary of the group 21-50)
- N = 3194
- $(\sum freq)_l$ = 200+450+300 =950
- $freq_{median}$ = 1500
- Width = 30
- Median = $20,5 + \frac{\frac{3194}{2} - 950}{1500} \times 30 = 33,44$

Mode

- Mode is the value that occurs most frequency in the data set.
- Data sets with one, two, or three modes are respectively called **unimodal**, **bimodal**, and **trimodal**. In general, a data set with two or more modes is **multimodal**.
- At the other extreme, if each data value occurs only once, then there is no mode.
- **Example**: 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 data is bimodal. The two modes are $52,000 and $70,000.

Symmetric vs Skewed Data

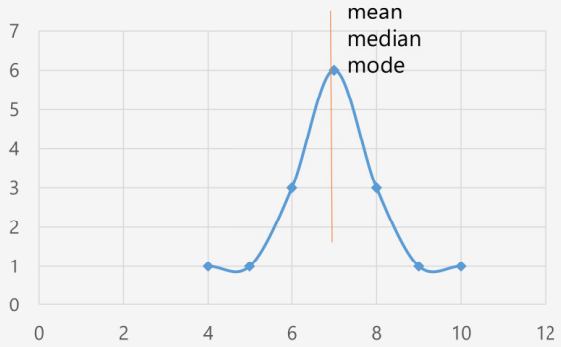
- Median, mean and mode of summetric, positively and negatively skewed data.



Positively skewed Symmetric Negatively skewed

Example: symmetric

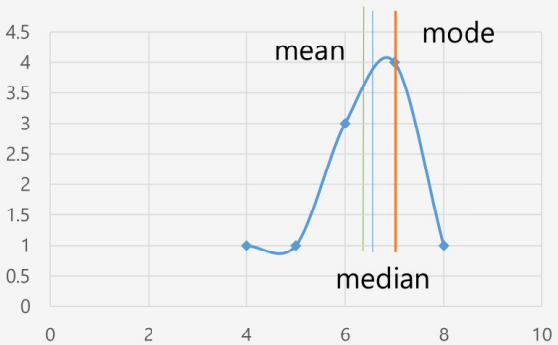
- Consider the following data set: 4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10
- Mean = median = mode = 7



- The chart displays a **symmetrical** distribution of data. A distribution is symmetrical if a vertical line can be drawn at some point in the chart such that the shape to the left and the right of the vertical line are mirror images of each other. The mean, the median, and the mode are each seven for these data.

Example: negatively skewed

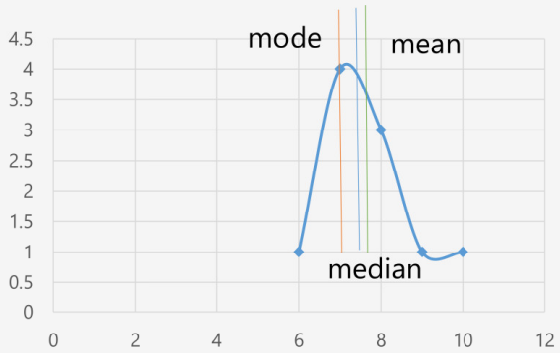
- Consider the following data set: 4; 5; 6; 6; 6; 7; 7; 7; 7; 8 is not symmetrical
  - Mean = (4+5+6+6+6+7+7+7+7+8)/10 =6.3
  - Median = 6.5
  - Mode = 7



- The mean is 6.3, the median is 6.5, and the mode is 7. Notice that the mean is less than the median, and they are both less than the mode. The mean and the median both reflect the skewing, but the mean reflects it more so.

Example: positively skewed

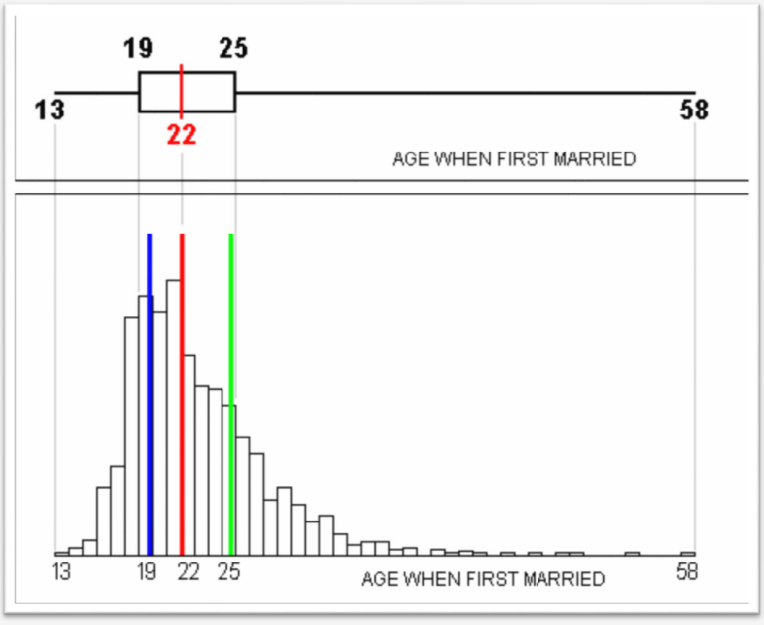
- Consider the following data set: 6; 7; 7; 7; 7; 8; 8; 8; 9; 10, is also not symmetrical
  - Mean = (6+7+7+7+7+8+8+8+9+10)/10 = 7.7
  - Median = 7.5
  - Mode = 7



- The mean is 7.7, the median is 7.5, and the mode is 7. Of the three statistics, the mean is the largest, while the mode is the smallest. Again, the mean reflects the skewing the most.

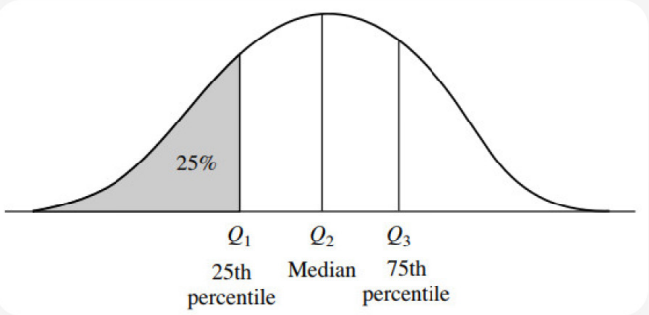
Measuring the Dispersion of Data

- Range, Quartiles, and Interquartile Range
- Five-Number Summary, Boxplots, and Outliers
- Variance and standard deviation



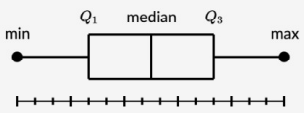
Range, Quartiles, and Interquartile Range

- Let $x_1, x_2, \ldots, x_n$ be a set of observations for some numeric attribute, X. The **range** of the set is the difference between the largest and smallest values.
- **Quantiles** are points taken at regular intervals of a data distribution, dividing it into essentially equal size consecutive sets
  - **The 2-quantile** is the data point dividing the lower and upper halves of the data distribution. It corresponds to the median
  - **The 4-quantiles (quartiles)** are the three data points that split the data distribution into four equal parts; each part represents one-fourth of the data distribution
- **Quartiles**: Q1 (25th percentile), Q3 (75th percentile)
- **Inter-quartile range**: **IQR** = Q3 – Q1

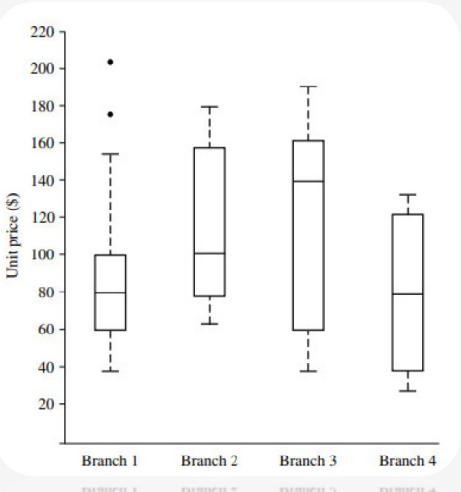


Five-Number Summary, Boxplots, and Outliers

- No single numeric measure of spread (e.g., IQR) is very useful for describing skewed distributions.
- **Outlier**: usually, a value higher/lower than 1.5 x IQR.
- **Five-number summary** of a distribution
  - Minimum, Q1, Median, Q3, Maximum



- **Boxplot**
  - Data is represented with a box
  - The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
  - The median is marked by a line within the box
  - Whiskers: two lines outside the box extended to Minimum and Maximum
  - Outliers: points beyond a specified outlier threshold, plotted individually



Boxplot Analysis: Example

- A sample of 10 boxes of raisins has these weights (in grams): 25, 28, 29, 29, 30, 34, 35, 35, 37, 38. Find the five-number summary and Make a box plot.
  - Our data is already in order.
  - Median = $\frac{30+34}{2}$ = 32
  - The first quartile is the median of the data points to the left of the median.

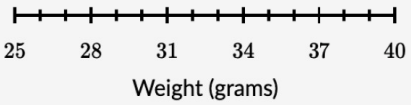
    25, 28, 29, 29, 30. Q1 = 29
  - The third quartile is the median oaf the data points to the right of the median.

    34, 35, 35, 37, 38. Q3 = 35
  - The min is the smallest data point, which is 25
  - The max is the largest data point, which is 38

→ The five-number summary is 25, 29, 32, 35, 38.

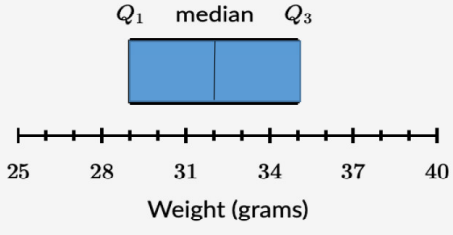
Boxplot Analysis

Step 1: Scale and label an axis that fits the five-number summary



Step 2: Draw a box from Q1 to Q3 with a vertical line through the median.

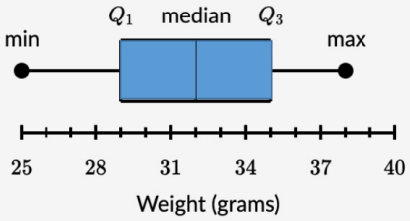
Recall that Q1=29, the median is 32, and Q3=35.



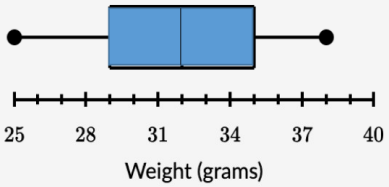
Boxplot Analysis

Step 3: Draw a whisker from Q1 to the min and from Q3 to the max.

Recall that the min is 25 and the max is 38.



We don't need the labels on the final product:



Variance and standard deviation

- Variance and standard deviation are measures of data dispersion. They indicate how spread out a data distribution is.
- **Variance**: (algebraic, scalable computation)

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

- **Standard deviation** σ is the square root of variance.
- A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.

Variance and standard deviation: example

- 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.
- We have $\bar{x} = 58{,}000\$$, N=12

$$\sigma^2 = \frac{1}{12}\left(30^2 + 36^2 + ... + 110^2\right) - 58^2 = 379.17$$
$$\sigma = \sqrt{379.17} = 19.47$$

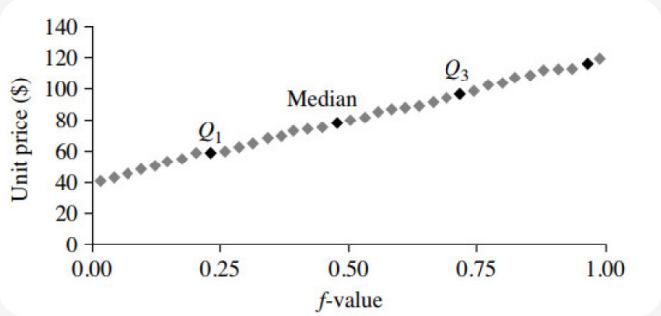
- The basic properties of the standard deviation as a measure of spread are as follows:
  - σ measures spread about the mean and should be considered only when the mean is chosen as the measure of center.
  - $\sigma = 0$ only when there is no spread, that is, all observations have the same value. Otherwise, $\sigma > 0$.

Graphic Displays of Basic Statistical Descriptions

- The graphic displays of basic statistical descriptions: **quantile plots**, **quantile–quantile plots**, **histograms**, and **scatter plots**. Such graphs are helpful for the visual inspection of data, which is useful for data preprocessing.
- **Quantile plot**: show univariate distributions (i.e., data for one attribute)
- **Quantile-quantile (q-q) plot**: univariate distributions
- **Histogram**: univariate distributions
- **Scatter plot**: bivariate distributions (i.e., involving two attributes)

Quantile Plot

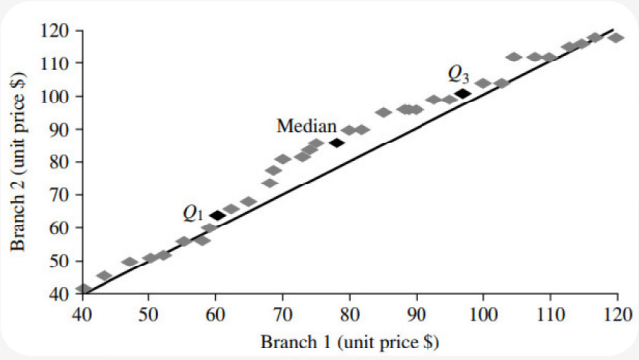
- Let x_i, for $i = 1$ to N, be the data sorted in increasing order so that x_1 is the smallest observation and x_N is the largest for some ordinal or numeric attribute X.
- Each observation, x_i, is paired with a percentage, f, which indicates that approximately $f_i \times 100\%$ of the data are below the value, x_i.



- Note that the 0.25 percentile corresponds to quartile Q_1, the 0.50 percentile is the median, and the 0.75 percentile is Q_3.
- This graph allows us to compare different distributions based on their quantiles. In the example, we can compare their Q1, median, Q3, and other f_i values at a glance.

Quantile-Quantile Plot

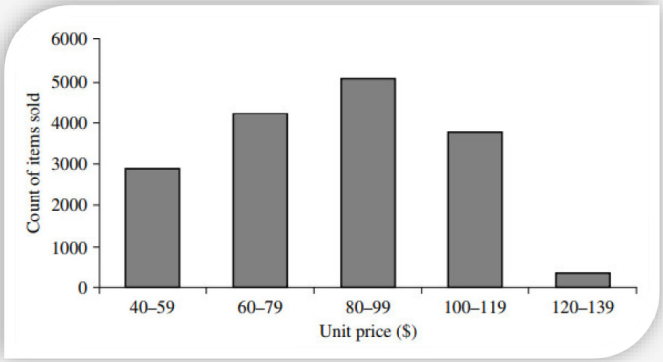
- Graphs the quantiles of one univariate distribution against the corresponding quantiles of another. It allows the user to view whether there is a shift in going from one distribution to another.
- Suppose that we have two sets of observations for the attribute or variable unit price, taken from two different branch locations. Let $x_1, x_2, \ldots, x_N$ be the data from the first branch, and $y_1, y_2, \ldots, y_M$ be the data from the second, where each data set is sorted in increasing order.



  - If M = N: plot y_i against x_i, where y_i and x_i are both $(i - 0.5)$/N quantiles of their respective data sets.
  - If M < N, there can be only M points on the q-q plot. Here, y_i is the $(i - 0.5)$/M quantile of the y data, which is plotted against the $(i - 0.5)$/M quantile of the x data.

Histogram Analysis

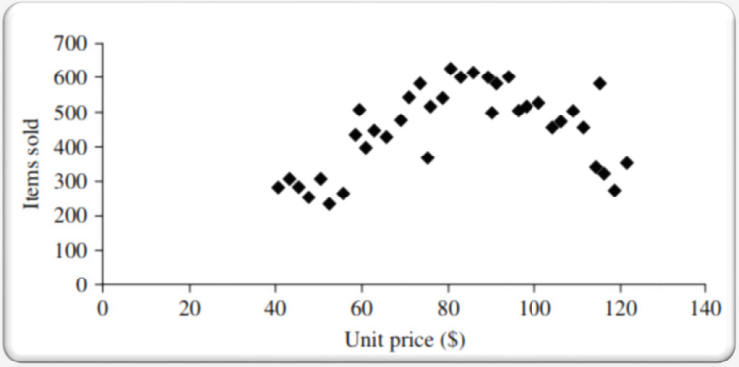
- Histogram: a graphical method for summarizing the distribution of a given attribute, X
- If X is nominal, a pole or vertical bar is drawn for each known value of X. The height of the bar indicates the frequency (i.e., count) of that X value. The resulting graph is more commonly known as a bar chart.



- If X is numeric, the term histogram is preferred. The range of values for X is partitioned into disjoint consecutive subranges. The subranges, referred to as buckets or bins, are disjoint subsets of the data distribution for X. The range of a bucket is known as the width. Typically, the buckets are of equal width.
- Histograms and partitioning rules are further discussed in **Chapter 3** on data reduction

Scatter Plot and Data Correlation

- A scatter plot is one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numeric attributes.
- To construct a scatter plot, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane.



Scatter Plot and Data Correlation

- If the plotted points pattern slopes from lower left to upper right, this means that the values of X increase as the values of Y increase, suggesting a positive correlation.
- If the pattern of plotted points slopes from upper left to lower right, the values of X increase as the values of Y decrease, suggesting a negative correlation.
- Three cases for which there is no correlation relationship between the two attributes

Outline

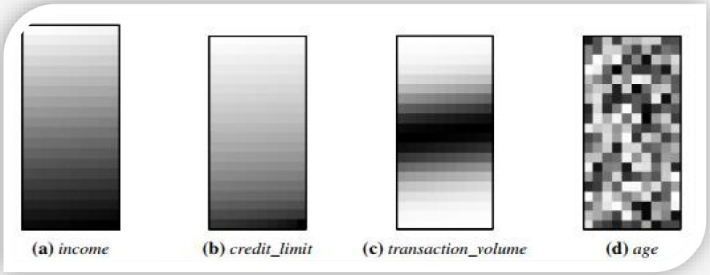
1. Data Objects and Attribute Types
2. Basic Statistical Descriptions of Data
3. **Data Visualization**
4. Measuring Data Similarity and Dissimilarity
5. Summary

Data Visualization

- Data visualization aims to communicate data clearly and effectively through graphical representation.
- More popularly, we can take advantage of visualization techniques to discover data relationships that are otherwise not easily observable by looking at the raw data.
- Categorization of visualization methods:
  - Pixel-oriented visualization techniques
  - Geometric projection visualization techniques

Pixel-Oriented Visualization Techniques

- A simple way to visualize the value of a dimension is to use a pixel where the color of the pixel reflects the dimension's value.
- For a data set of *m* dimensions, **pixel-oriented techniques** create *m* windows on the screen, one for each dimension.
- The *m* dimension values of a record are mapped to m pixels at the corresponding positions in the windows
- The colors of the pixels reflect the corresponding values.

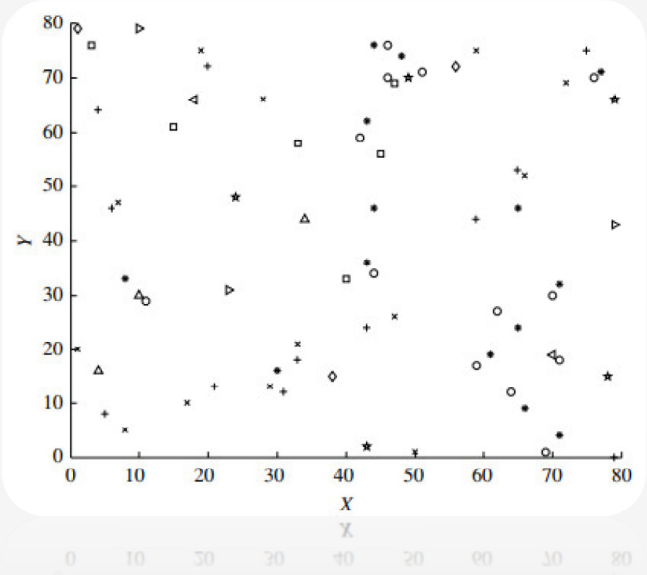


Pixel-oriented visualization of four attributes by sorting all customers in income ascending order.

- **Example**: We can sort all customers in income-ascending order, and use this order to lay out the customer data in the four visualization windows. The pixel colors are chosen so that the smaller the value, the lighter the shading. Using pixel-based visualization, we can easily observe the following: credit limit increases as income increases; customers whose income is in the middle range are more likely to purchase more; there is no clear correlation between income and age.

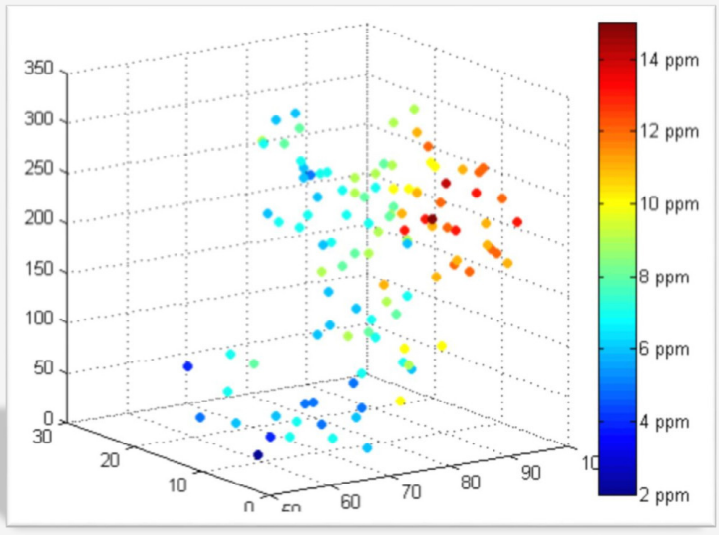
Geometric Projection Visualization Techniques

- **Geometric projection techniques** help users find interesting projections of multidimensional data sets. The central challenge the geometric projection techniques try to address is how to visualize a high-dimensional space on a 2-D display.
- A **scatter plot** displays 2-D data points using Cartesian coordinates. A third dimension can be added using different colors or shapes to represent different data points.
- In **the example**, X and Y are two spatial attributes and the third dimension is represented by different shapes. Through this visualization, we can see that points of types "+" and "×" tend to be co-located.



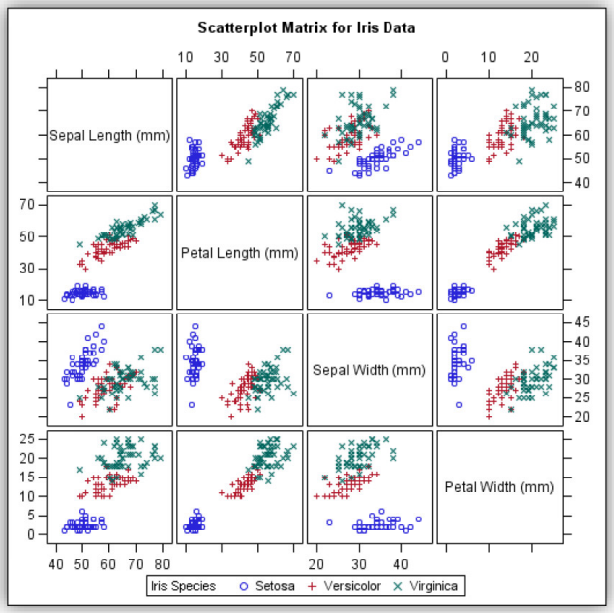
Geometric Projection Visualization Techniques (2)

- A 3-D scatter plot uses three axes in a Cartesian coordinate system. If it also uses color, it can display up to 4-D data points.
- For data sets with more than four dimensions, scatter plots are usually ineffective.

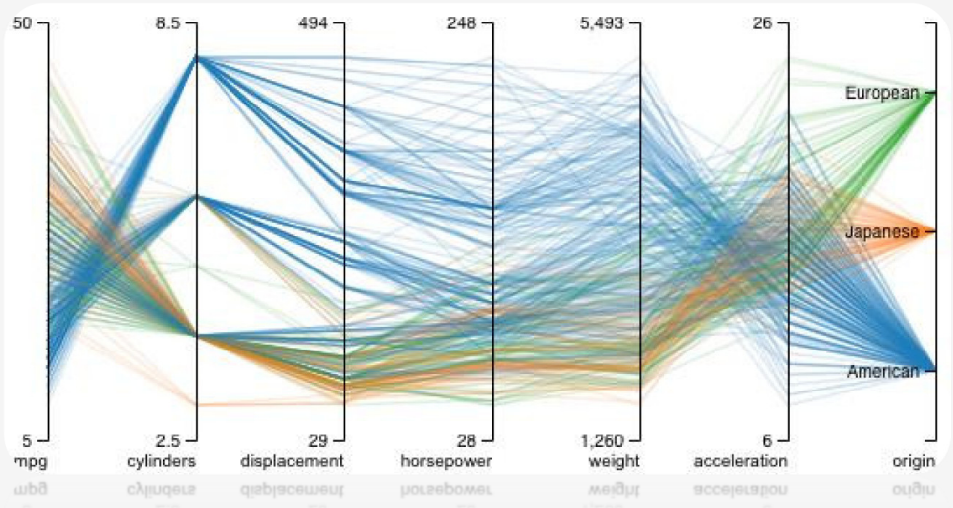


Geometric Projection Visualization Techniques (3)

- For data sets with more than four dimensions, scatter plots are usually ineffective. The scatter-plot matrix technique is a useful extension to the scatter plot.
- For an n-dimensional data set, a scatter-plot matrix is an n × n grid of 2-D scatter plots that provides a visualization of each dimension with every other dimension.
- The scatter-plot matrix becomes less effective as the dimensionality increases.



Geometric Projection Visualization Techniques (4)



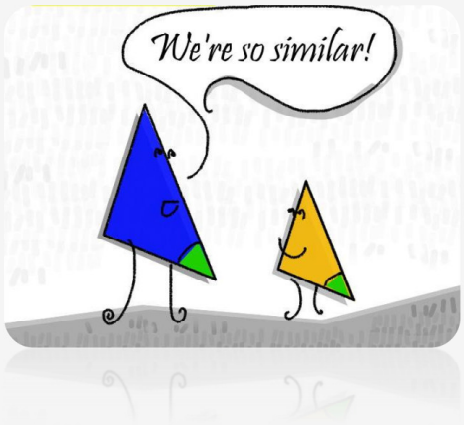
- To visualize n-dimensional data points, the parallel coordinates technique draws n equally spaced axes, one for each dimension, parallel to one of the display axes.
- A data record is represented by a polygonal line that intersects each axis at the point corresponding to the associated dimension value.
- A major limitation of the parallel coordinates technique is that it cannot effectively show a data set of many records

Outline

1. Data Objects and Attribute Types
2. Basic Statistical Descriptions of Data
3. Data Visualization
4. **Measuring Data Similarity and Dissimilarity**
5. Summary

Data Similarity and Dissimilarity

- In data mining applications, such as clustering, outlier analysis, and nearest-neighbor classification, we need ways to assess how alike or unalike objects are in comparison to one another.
- For example, a store may want to search for clusters of customer objects, resulting in groups of customers with similar characteristics (e.g., similar income, area of residence, and age). A cluster is a collection of data objects such that the objects within a cluster are similar to one another and dissimilar to the objects in other clusters.



- Outlier analysis also employs clustering-based techniques to identify potential outliers as objects that are highly dissimilar to others

Data Similarity and Dissimilarity

- **Similarity**
  - Numerical measure of how alike two data objects are
  - Value is higher when objects are more alike
  - Often falls in the range [0,1]
- **Dissimilarity**
  - Numerical measure of how different two data objects are
  - Lower when objects are more alike
  - Minimum dissimilarity is often 0
  - Upper limit varies
- **Proximity** refers to a similarity or dissimilarity

Measuring Data Similarity and Dissimilarity

- Data Matrix versus Dissimilarity Matrix
- Proximity Measures for Nominal Attributes
- Proximity Measures for Binary Attributes
- Dissimilarity of Numeric Data
- Proximity Measures for Ordinal Attributes
- Cosine Similarity for very long and sparse data vectors

Data Matrix and Dissimilarity Matrix

- Suppose that we have *n* objects (e.g., persons, items, or courses) described by *p* attributes (also called measurements or features, such as age, height, weight, or gender). The objects are $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$, where x_{ij} is the value for object x_i of the j-th attribute.
- **Data matrix:** This structure stores the *n* data objects in the form of a relational table, or *n*-by-*p* matrix (*n* objects × *p* attributes).
- **Dissimilarity matrix:** a collection of proximities that are available for all pairs of n objects. It is often represented by an n-by-n table. Where $d(i,j)$ is the measured dissimilarity or "difference" between objects i and j. In general, $d(i,j)$ is a non-negative number that is close to 0 when objects i and j are highly similar or "near" each other, and becomes larger the more they differ.
- Measures of similarity can often be expressed as a function of measures of dissimilarity. For example, for nominal data,

$$sim(i,j) = 1 - d(i,j)$$

Attributes

Data objects

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Data objects

Data objects

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Proximity Measure for Nominal Attributes

- Nominal attributes can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute).
- **Method 1**: Simple matching

$$d(i,j) = \frac{p - m}{p}$$

  - *m*: number of matches, *p*: total number of variables
  - Weights can be assigned to increase the effect of m or to assign greater weight to the matches in attributes having a larger number of states.
- **Method 2**: Use a large number of binary attributes
  - Creating a new binary attribute for each of the *M* nominal states
  - For example, to encode the nominal attribute map color, a binary attribute can be created for each of the five colors previously listed. For an object having the color yellow, the yellow attribute is set to 1, while the remaining four attributes are set to 0

Proximity Measure for Nominal Attributes – Example

- Since here we have **one nominal attribute** (test-1), we set p=1 so that $d(i,j)$ evaluates to 0 if objects i and j match, and 1 if the objects differ. Thus, we get

$$\begin{matrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{matrix}$$

- From this, we see that all objects are dissimilar except objects 1 and 4 i.e., $d(4,1) = 0$.

A Sample Data Table Containing Attributes of Mixed Type

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

Proximity Measure for Binary Attributes

- We have the 2 × 2 contingency table, where ***q*** is the number of attributes that equal 1 for both objects i and j, ***r*** is the number of attributes that equal 1 for object i but equal 0 for object j, ***s*** is the number of attributes that equal 0 for object i but equal 1 for object j, and ***t*** is the number of attributes that equal 0 for both objects i and j. The total number of attributes is ***p***, where $p = q + r + s + t$.

Object *i* \ Object *j*	1	0	Sum
1	q	r	q+r
0	s	t	s+t
sum	q+s	r+t	p

- Distance measure for symmetric binary variables:

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

- Distance measure for asymmetric binary variables:

$$d(i,j) = \frac{r+s}{q+r+s}$$

- Jaccard coefficient (similarity measure for asymmetric binary variables):

$$sim(i,j) = \frac{q}{q+r+s} = 1 - d(i,j)$$

Proximity Measure for Binary Attributes: Example

❑ Relational Table Where Patients Are Described by Binary Attributes

Name	gender	fever	Cough	Test – 1	Test - 2	Test – 3	Test – 4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

❑ **Solution:**

- Gender is a symmetric attribute
- The remaining attributes are asymmetric binary
- Let the values *Y* and *P* be 1, and the value *N* be 0

Proximity Measure for Binary Attributes: Example

Mary

Jack		1	0	*sum*
	1	2	0	2
	0	1	3	4
	sum	3	3	6

$$d(Jack, Mary) = \frac{0+1}{2+0+1} = 0.33$$

Jim

Jack		1	0	*sum*
	1	1	1	2
	0	1	3	4
	sum	2	4	6

$$d(Jack, Jim) = \frac{1+1}{1+1+1} = 0.67$$

Mary

Jim		1	0	*sum*
	1	1	1	2
	0	2	2	4
	sum	3	3	6

$$d(Jim, Mary) = \frac{1+2}{1+1+2} = 0.75$$

Dissimilarity of Numeric Data

- Distance measures that are commonly used for computing the dissimilarity of objects described by numeric attributes: **Euclidean**, **Manhattan**, and **Minkowski distances**.
- The data are normalized before applying distance calculations. This involves transforming the data to fall within a smaller or common range, such as [−1,1] or [0.0, 1.0]. Methods for normalizing data are discussed in detail in Chapter 3 on data preprocessing.
- The most popular distance measure is **Euclidean distance**. Let $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ be two objects described by p numeric attributes. The Euclidean distance between objects i and j is defined as

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$

- If each attribute is assigned a weight according to its perceived importance, **the weighted Euclidean distance**:

$$d(i,j) = \sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + \cdots + w_p(x_{ip} - x_{jp})^2}$$

Dissimilarity of Numeric Data (2)

- Another well-known measure is the **Manhattan distance**. It is defined as

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{ip} - x_{jp}|$$

- Both the **Euclidean and the Manhattan distances** satisfy the following mathematical properties (known as metric):
  - **Non-negativity**: $d(i,j) \geq 0$: Distance is a non-negative number.
  - **Identity of indiscernibles**: $d(i,j) = 0$: The distance of an object to itself is 0.
  - **Symmetry**: $d(i,j) = d(j,i)$: Distance is a symmetric function.
  - **Triangle inequality**: $d(i,j) \leq d(i,k) + d(k,j)$: Going directly from object i to object j in space is no more than making a detour over any other object k.

Dissimilarity of Numeric Data (3)

- **Minkowski distance** is a generalization of the Euclidean and Manhattan distances. It is defined as

$$d(i,j) = \sqrt[h]{|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \cdots + |x_{ip} - x_{jp}|^h}$$

where h is a real number such that h ≥ 1. It represents the Manhattan distance when h=1 (i.e., L1 norm) and Euclidean distance when h=2 (i.e., L2 norm).

- **The supremum distance** is a generalization of the Minkowski distance for $h \to \infty$. This difference is the supremum distance, defined more formally as:

$$d(i,j) = \lim_{h\to\infty} \left(\sum_{f=1}^{p} |x_{if} - x_{jf}|^h\right)^{\frac{1}{h}} = \max |x_{if} - x_{jf}|$$

Dissimilarity of Numeric Data: Example 1

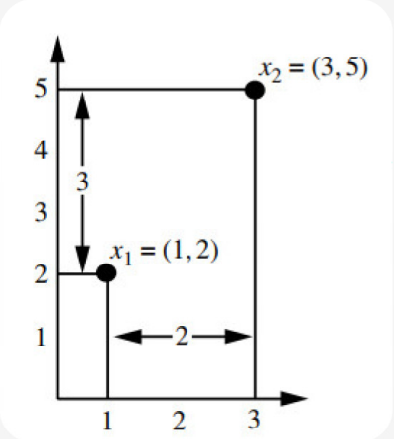
- Let $x_1 = (1,2)$ and $x_2 = (3,5)$ represent two objects
- The Euclidean distance between the two is

$$d(x_1, x_2)_{Euclidean} = \sqrt{2^2 + 3^2} = 3.61$$

- The Manhattan distance between the two is

$$d(x_1, x_2)_{Manhattan} = 2 + 3 = 5$$

- The supremum distance: Because the second attribute gives the greatest difference between values for the objects, the supremum distance between both objects is $5 - 2 = 3$.



Dissimilarity of Numeric Data: Example 2

Distances - Data Matrix and Dissimilarity Matrix

Data Matrix

Point	Attribute 1	Attribute 2
x_1	1	2
x_2	3	5
x_3	2	0
x_4	4	5

Dissimilarity Matrix (with Manhattan (L_1) Distance)

L	x_1	x_2	x_3	x_4
x_1	0			
x_2	5	0		
x_3	3	6	0	
x_4	6	1	7	0

Dissimilarity Matrix (with Euclidean (L2) Distance)

L	x_1	x_2	x_3	x_4
x_1	0			
x_2	3.61	0		
x_3	2.24	5.1	0	
x_4	4.24	1	5.39	0

Dissimilarity Matrix (with Supremum Distance)

L	x_1	x_2	x_3	x_4
x_1	0			
x_2	3	0		
x_3	2	5	0	
x_4	3	1	5	0

Proximity Measures for Ordinal Attributes

- An ordinal variable can be discrete or continuous. Order is important, e.g., rank. An example includes the sequence small, medium, large for a size attribute.
- The dissimilarity computation with respect to f involves the following steps:
  - Replace each x_{if} by its corresponding rank $r_{if} \in \{1, \dots, M_f\}$
  - Map the range of each variable onto [0, 1] by replacing i-th object in the f-th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - Compute the dissimilarity using any of the distance measures for numeric attributes, using z_{if} to represent the f value for the ith object.

Proximity Measures for Ordinal Attributes: Example

- At this time, we only use the *object-identifier* and the *test-2(ordinal)* attribute. Since, there are three states for *test-2*: *fair*, *good*, and *excellent*. Thus $M_f = 3$.
- Step 1, we replace each value for test-2 by its rank.
- Step 2, normalizes the ranking by $z_{if} = \frac{r_{if} - 1}{M_f - 1}$
  - Fair = $\frac{1-1}{3-1} = 0.0$
  - Good = $\frac{2-1}{3-1} = 0.5$
  - Excellent = $\frac{3-1}{3-1} = 1$

Then map rank 1 to 0.0, rank 2 to 0.5 and rank 3 to 1.0.

- Step 3. Compute the dissimilarity using any of the distance measures for numeric attributes

A Sample Data Table Containing Attributes of Mixed Type

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

Object Identifier	Test-2 (old)	Test-2 (new) r_f	z_f
1	Excellent	3	1
2	Fair	1	0
3	Good	2	0.5
4	Excellent	3	1

Cosine Similarity

- A **document** can be represented by thousands of attributes, each recording the *frequency* of a particular word (such as keywords) or phrase in the document.
- Other vector objects: gene features in micro-arrays, ...

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

- Applications: information retrieval, biologic taxonomy, gene feature mapping, ...
- Cosine similarity: Let x and y be two vectors for comparison

$$sim(x, y) = \frac{x \bullet y}{\|x\| \|y\|}$$

where: • indicates vector dot product, $\|d\|$: is the Euclidean norm of vector $x = (x_1, x_2, \ldots, x_p)$, defined as $\sqrt{x_1^2 + x_p^2 + \cdots + x_p^2}$.

- A cosine value of 0 means that the two vectors are at 90 degrees to each other (orthogonal) and have no match. The closer the cosine value to 1, the smaller the angle and the greater the match between vectors.

Cosine Similarity: Example

❑ Find the **Cosine similarity** between documents 1 and 2: d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0), d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)

❑ **Solution:**

- $d_1 \bullet d_2 = 5*3+0*0+3*2+0*0+2*1+0*1+0*1+2*1+0*0+0*1 = 25$
- $||d_1|| = \sqrt{5*5+0*0+3*3+0*0+2*2+0*0+0*0+2*2+0*0+0*0} = \sqrt{42} = 6.481$
- $||d_2|| = \sqrt{3*3+0*0+2*2+0*0+1*1+1*1+0*0+1*1+0*0+1*1} = \sqrt{17} = 4.12$
- $sim(x, y) = \frac{x \bullet y}{||x|| ||y||} = \frac{25}{\sqrt{42}\sqrt{17}} = 0.94$
- Therefore, if we were using the cosine similarity measure to compare these documents, they would be considered quite similar.

Outline

1. Data Objects and Attribute Types

2. Basic Statistical Descriptions of Data

3. Data Visualization

4. Measuring Data Similarity and Dissimilarity

5. Summary

Summary

- **Data attribute** types: **nominal**, **binary**, **ordinal**, **interval-scaled**, **ratio-scaled.**
- Many types of data sets, e.g., **numerical**, **text**, **graph**, **Web**, **image**.
- Basic statistical descriptions provide the analytical foundation for data preprocessing: **mean**, **weighted mean**, **median**, and **mode** for measuring **the central tendency of data**; and **range**, **quantiles**, **quartiles**, **interquartile range**, **variance**, and **standard deviation** for measuring **the dispersion of data.**
- **Data visualization techniques** may be **pixel-oriented**, **geometric-based**, **icon-based**, or **hierarchical**.
- Measures of **object similarity and dissimilarity** are used in data mining applications such as clustering, outlier analysis, and nearest-neighbor classification.