Data Mining
Getting to know your data

Hamid Beigy

Sharif University of Technology

Fall 1395
Table of contents

1 Introduction

2 Getting to know your data

3 Statistical description of data

4 Data visualization

5 Reading
1. Introduction

2. Getting to know your data

3. Statistical description of data

4. Data visualization

5. Reading
A typical knowledge discovery process is
Getting to know your data

- Real-world data are typically **noisy, enormous in volume, and may originate from heterogenous sources**.

- The **first step** of data mining is to know the data. We need to know
  - What are the **type of attributes or fields** that make up the data?
  - What **kind of values** does each attribute have?
  - Which attributes are **discrete** and which are **continuous-valued**?
  - How are the **values distributed**?
  - Are the ways we can **visualize the data** to get a better sense of it?
  - Can we spot any **outlier**?
  - Can we **measure the similarity** of some **data objects** with respect to others?
Attribute types

Nominal attributes  The values of nominal attributes are symbols or name of things. Each value represents some kind of category, code, or state, and nominal attributes are also referred as categorical. The values does not have any meaningful order.

Binary attributes  A binary attribute is a nominal attribute with only two categories.
- A binary attribute is symmetric if both of its states are equally valuable and carry the same weight.
- A binary attribute is asymmetric if the outcomes of the states are not equally important.

Ordinal attributes  An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them, but the magnitude between successive values is not known.

Numerical attributes  A numeric attribute is quantitative; that is, it is a measurable quantity, represented in integer or real values. Numeric attributes can be interval-scaled or ratio-scaled.
- Interval-scaled attributes are measured on a scale of equal-size units. Their values have order. We can compare and quantify the difference between values. Temperatures in Celsius(Fahrenheit) do not have a true zero-point.
- A ratio-scaled attribute is a numeric attribute with an inherent zero-point. Their values have order and allow us to compare and quantify the difference between values. Temperatures in the Kelvin has a true zero-point.
Outline

1 Introduction

2 Getting to know your data

3 Statistical description of data

4 Data visualization

5 Reading
Statistical description of data

- 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.
- Three basic statistical descriptions are measures of central tendency. This measures the location of the middle or center of a data distribution, such as mean, median, and mode.

Measuring the data dispersion. This measures how are the data spread out. The most common data dispersion measures are range, quartile, interquartile range (IQR), five-numbers summary, box plots, variance, and standard deviation.
Box plots are a popular way of visualizing a distribution. A box plot incorporates the five-numbers summary (min, max, $Q_1$, $Q_3$, median).
In **Graphic displays of basic statistical description of data**, graphs are helpful for visual inspection of data. These includes:

- Quantile plots
- Quantile-quantile plots
- Histograms

![Histogram of items sold](image)

Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.

![Scatter plots](image)
Plotting histograms is a graphical method for summarizing the distribution of a given attribute $X$. If $X$ is nominal, then a plot is drawn for each value of $X$. If $X$ is numeric, the range of values for $X$ is partitioned into disjoint consecutive subranges (buckets) or bins. The value of bucket (height of a bar) indicates the frequency of that $X$ value. The resulting graph is more commonly known as a bar chart.
A scatter plot is a graphical method for determining if there appears to be a relationship between two numeric attributes (if any).
1 Introduction

2 Getting to know your data

3 Statistical description of data

4 Data visualization

5 Reading
Data visualization aims to communicate data clearly and effectively through graphical representation.

We can take advantage of visualization techniques to discover data relationships that exist but are not easily observable by looking at the raw data.

Consider the visualization of a data set using scatter plots

Some visualization techniques

- pixe-oriented techniques
- geometric projection techniques
- Icon-based techniques
- Hierarchical techniques
- Graph based techniques
Outline

1. Introduction
2. Getting to know your data
3. Statistical description of data
4. Data visualization
5. Reading
Read chapter 2 of the following book
J. Han, M. Kamber, and Jian Pei, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2012.