Data Exploration

ZHANG,Pingchuan
1155173674@link.cuhk.edu.hk

October 6, 2024

Data Cleaning

What data types will we encounter?

1. Sequential Data
2. Data Matrix
3. Spatial Data
4. Temporal Data

Data Matrix Review

1. Data that consists of a collection of records, each of which consists of a fixed set of attributes.
2. Data set can be represented by an n by m matrix, where there are n rows, one for each object, and m columns, one for each attribute.

Person	Height (m)	Weight (kg)
P1	1.79	75
P2	1.64	54
P3	1.70	63
P4	1.88	78

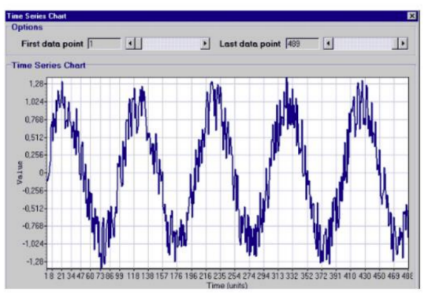
Table 1: 4 by 2 matrix. We have 4 people, each with 2 attributes.

Examples of data quality problems

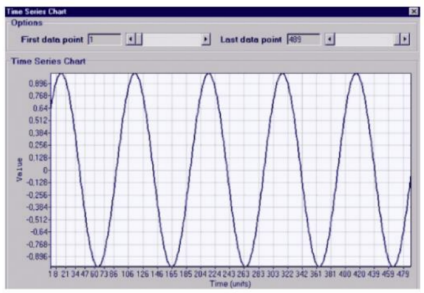
1. Noise and outliers
2. Missing values
3. Duplicate data
4. Unnormalized data
5. Categorical data

1. Noise

Noise refers to modification of original values



A sine wave with noise



The denoised sine wave

Figure 1: Examples of Noise

2. Outlier

Outlier are data objects with characteristics that are considerably different than most of the other data objects in the data set

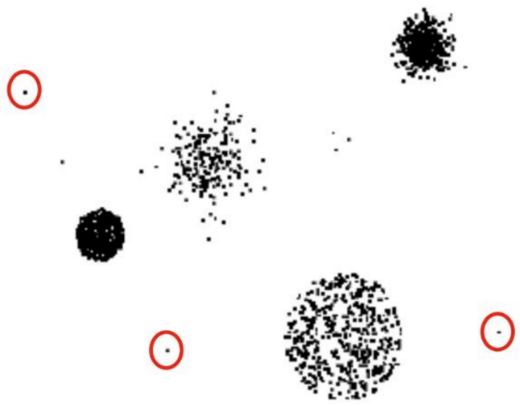


Figure 2: Red values in clustered data that are clearly outside the similar clusters

3. Missing values

Reasons for missing values

1. Information is not collected(e.g., people decline to give their age and weight)
2. Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)

How to handle missing values?

1. Eliminate Data Objects
2. Estimate Missing Values
3. Ignore the Missing Value During Analysis
4. Replace with all possible values (weighted by their probabilities)

4. Duplicate data

Dataset may include data objects that are duplicates, or almost duplicates of one another. Major issue when merging data from heterogeneous sources

Database 1		
Person	**Height (m)**	**Weight (kg)**
P1	1.79	75
P2	1.64	54
P3	1.70	63
P4	1.88	78

Database 2		
Person	**Height (m)**	**Weight (kg)**
P1	1.79	75
P7	1.65	55
P8	1.69	63
P9	1.87	77

Table 2: Two databases with height and weight information.

5. Unnormalized data

Attributes not on the similar level of measurement

Solutions to unnormalized data:

1. Min-max normalization

$$v' = \frac{v - v_{\min}}{v_{\max} - v_{\min}}$$

2. Z-score normalization

$$v' = \frac{v - Mean(v)}{Std(v)}$$

Note: Both normalization methods rely on the concept of measuring the distance of each entry from the expected value. Min-max normalization transforms the data so that all entries fall within a range between 0 and 1.

6. Categorical data

Person	Height(m)	Weight(kg)	Gender
P1	0.625	0.875	Male
P2	0	0	Female
P3	0.25	0.375	Female
P4	1	1	Male

Person	Height(m)	Weight(kg)	Male	Female
P1	0.625	0.875	1	0
P2	0	0	0	1
P3	0.25	0.375	0	1
P4	1	1	1	0

Computers are better on handling numbers
For categorical data, we can use one-hot encoding

Figure 3: Categorical Data

Data Exploration

1. Summary statistics

Summary statistics are numbers that summarize properties of the data

Summarized properties include frequency,location and spread

Most summary statistics can be calculated in a single pass through the data

2. Measures of location: mean and median

1. The mean is the most common measure of the location of a set of points

$$\text{mean}(x) = \frac{1}{m}\sum_{i=1}^{m} x_i$$

Note:the mean is very sensitive to outliers

2. The median or a trimmed mean is thus also commonly used

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r+1 \\ \frac{1}{2}\left(x_{(r)} + x_{(r+1)}\right) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

3. Measures of spread: range and variance

1. Range is the difference between the max and min
2. The variance or standard deviation is the most common measure of the spread of a set of points

$$\text{variance}(x) = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \text{mean}(x))^2$$

Note: Sensitive to outlier

3. Other measures: Median absolute deviation (MAD):

$$\text{median}\left(|x_1 - \text{mean}(x)|, \ldots, |x_m - \text{mean}(x)|\right)$$

Interquartile range:

$$x_{75\%} - x_{25\%}$$

4. Percentiles

Given an ordinal or continuous attribute x and a number p between 0 and 100, the p-th percentile is a value of x such that $p\%$ of the observed values of x are less than x_p.

$$p = 50 \Rightarrow \quad x_p \text{ is close to the median value}$$

5. Frequency and mode

1. The frequencyof an attribute value is the percentage of time the value occurs in data set
2. The mode of an attribute is the most frequent attribute value
3. The notions of frequency and mode are typically used with categorical data

Exploratory visualization

Definition: **Visualization** is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analysed or reported.

Visualization of data is one of the most powerful and appealing techniques for data exploration

1. Humans have a well-developed ability to analyse large amounts of information that is presented visually
2. Can detect general patterns and trends
3. Can detect outliers and unusual patterns

1. Histograms

Usually shows the distribution of values of a single variable

1. Divide the values into bins, show a bar plot of the number of objects in each bin
2. The height of each bar indicates the number of objects
3. Shape of histogram depends on the number of bins

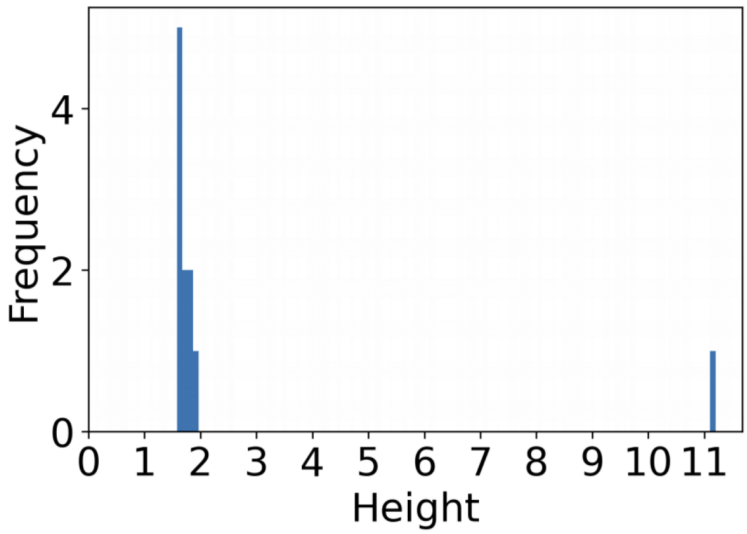


Figure 4: Explore the data very quickly and know the outlier of the data

Two-dimensional histograms show the joint distribution of values of two variable

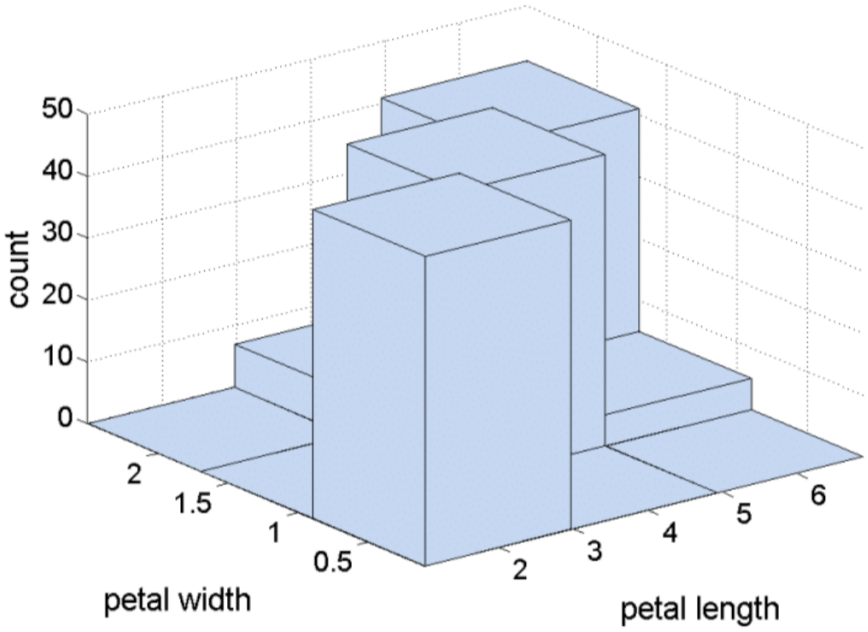


Figure 5: The figure shows that with petal length, petal width increases as well.

2. Box-plot

Box Plots clearly defines the median, the 75thand 25th percentiles and the min and max of a particular attribute as illustrated below. It helps us compare along the different attributes of a data matrix and helps determine the skew and interquartile range

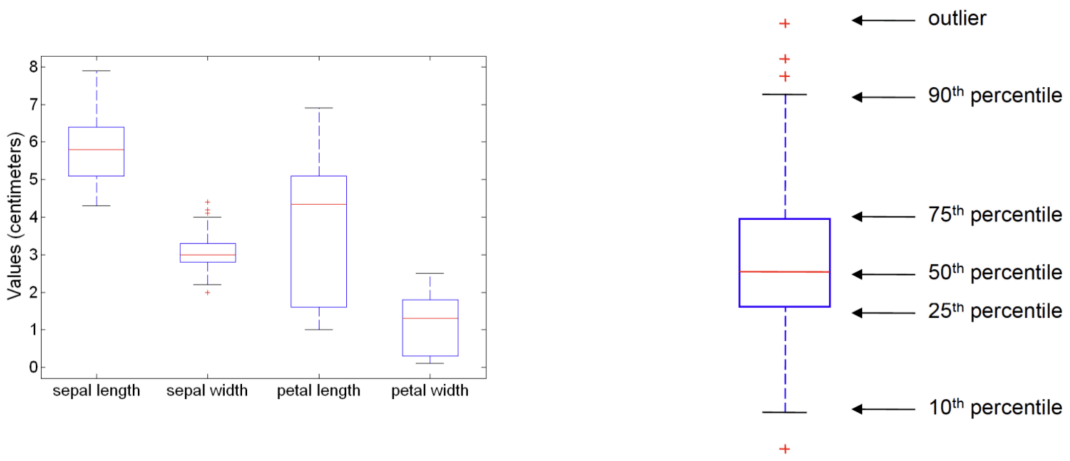


Figure 6: Box-plot