Data Mining

2.3 Data Cleaning

Fall 2008

Instructor: Dr. Masoud Yaghini

Outline

- Introduction
- Missing Values
- Outliers
- Inconsistent and Duplicate Data
- References

Introduction

Data Cleaning

- Real-world data tend to be incomplete, noisy, and inconsistent.
- Data cleaning tasks
  - Handle missing values
  - Detect and remove outliers
  - Correct inconsistent data
  - Remove duplicate data
- In this section, you will study basic methods for data cleaning.

Data Cleaning

- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

Missing Values

Missing Values

- Data is not always available
  - E.g., many instances have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to:
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred.

How to Handle Missing Data?

- **Ignore the instance**
  - usually done when class label is missing (assuming the tasks in classification)
  - not effective when the percentage of missing values per attribute varies considerably.
- **Fill in the missing value manually**
  - this approach is time-consuming and may not be feasible given a large data set with many missing values.

How to Handle Missing Data?

- **Use a global constant to fill in the missing value**
  - Replace all missing attribute values by the same constant, such as a label like "**Unknown**"
  - the mining program may mistakenly think that they form an interesting concept, since they all have a value in common—that of "Unknown."
- **Use the attribute mean to fill in the missing value**
  - For example, suppose that the average income of *AllElectronics* customers is $56,000.
  - Use this value to replace the missing value for *income.*

How to Handle Missing Data?

- **Use the attribute mean for all samples belonging to the same class**
  - 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.
- **Use the most probable value to fill in the missing value**
  - This may be determined with **regression**, inference-based tools using a **Bayesian formalism**, or **decision tree**.
  - For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for *income..*

How to Handle Missing Data?

- Use the most probable value to fill in the missing value is a popular strategy.
- In comparison to the other methods, it uses the most information from the present data to predict missing values.

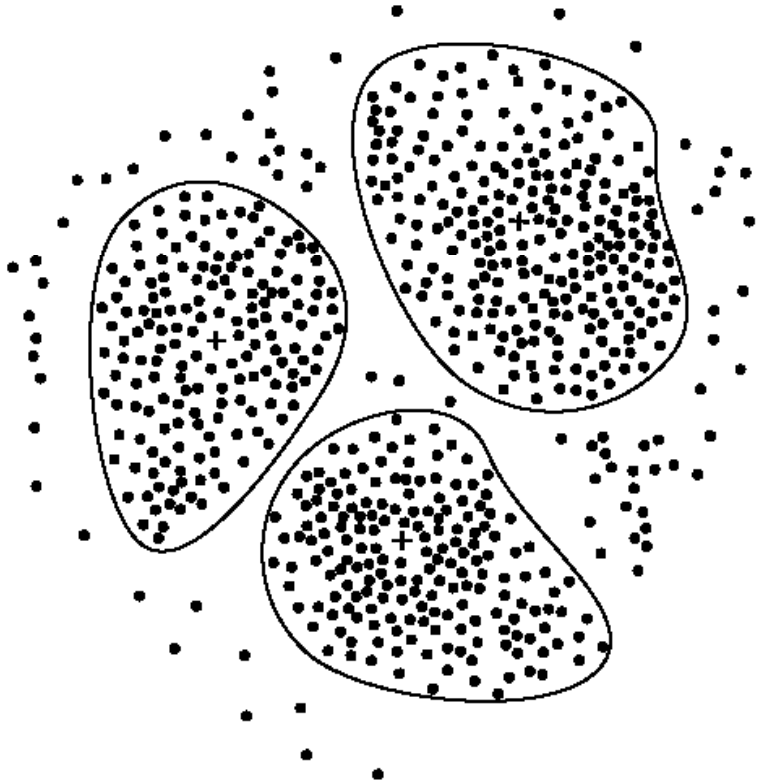
Outliers

Outliers

- Outliers
  - Outliers are data instances with **characteristics that are considerably different** than most of the other data instances in the data set

Outliers

- Outliers may be detected by clustering, where similar values are organized into groups, or “clusters.”



Outliers

- A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters.
- Each cluster centroid is marked with a "+", representing the average point in space for that cluster.
- Outliers may be detected as values that fall outside of the sets of clusters.

Inconsistent and Duplicate Data

Inconsistent

- Inconsistent: containing discrepancies in codes or names
  - e.g., Age="42" Birthday="03/07/1997"
  - e.g., Was rating "1,2,3", now rating "A, B, C"
  - e.g., discrepancy between duplicate records

Duplicate Data

Duplicate Data

- Data set may include data instances that are duplicates, or almost duplicates of one another
  - Major issue when merging data from different sources

- Data cleaning
  - Process of dealing with duplicate data issues

References

References

- J. Han, M. Kamber, **Data Mining: Concepts and Techniques**, Elsevier Inc. (2006). (Chapter 2)

The end