Data Mining

Part 2. Data Understanding and Preparation

2.3 Data Cleaning

Spring 2010

Instructor: Dr. Masoud Yaghini

Outline

- Introduction
- Handling missing values
- Detecting and removing outliers
- Correcting inconsistent data
- Schema integration
- Handling redundancy
- References

Introduction

Data Cleaning

- Real-world data tend to be incomplete, noisy, and inconsistent.
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation

Data Cleaning

- The first step in data cleaning as a process is **discrepancy detection**.
- Use **metadata**
  - what are the domain and data type of each attribute?
  - What are the acceptable values for each attribute?
  - What is the range of the length of values?
  - Do all values fall within the expected range?
  - Are there any known dependencies between attributes?

Data Cleaning

- Use **descriptive data summaries**
  - for grasping data trends and identifying anomalies.
  - e.g. values that are more than two standard deviations away from the mean for a given attribute may be flagged as potential outliers.

Data Cleaning

- The data should also be examined regarding:
  - A **unique rule**
    - says that each value of the given attribute must be different from all other values for that attribute.
  - A **consecutive rule**
    - says that there can be **no missing** values between the lowest and highest values for the attribute, and that all values must also be **unique** (e.g., as in check numbers).
  - A **null rule**
    - specifies the use of blanks, question marks, special characters, or other strings that may indicate the null condition (e.g., where a value for a given attribute is not available), and how such values should be handled.

Data Cleaning Tasks

- **Data cleaning** tasks
  - Handling missing values
  - Detecting and removing outliers
  - Correcting inconsistent data
  - Removing duplicate data
  - Schema integration
  - Handling redundancy

Handling Missing Values

Handling 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

Handling Missing Values

- **Handling Missing Values**
  - Ignore the instance
  - Fill in the missing value manually
  - Use a global constant to fill in the missing value
  - Use the attribute mean to fill in the missing value
  - Use the attribute mean for all samples belonging to the same class
  - Use the most probable value to fill in the missing value

Handling Missing Values

- **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.

Handling Missing Values

- **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*.

Handling Missing Values

- **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*..

Handling Missing Values

- 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.

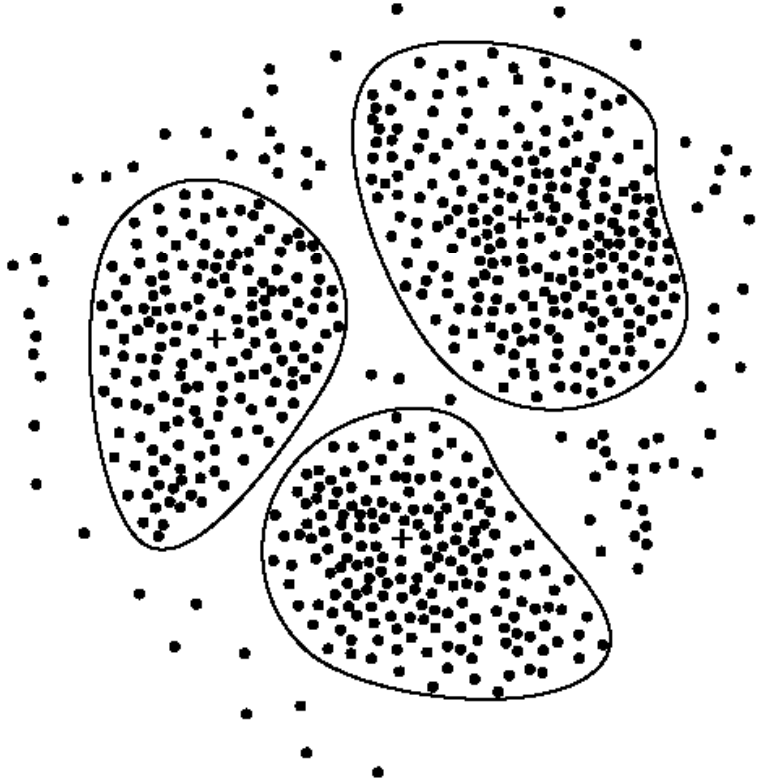
Detecting and Removing Outliers

Detecting and Removing Outliers

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

Detecting and Removing Outliers

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



Detecting and Removing 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.

Correcting Inconsistent Data

Correcting Inconsistent Data

- **Inconsistent**: containing discrepancies in codes or names
- Examples:
  - the data codes for ***pay_type*** in one database may be *"H"* and *"S"*, and *1* and *2* in another.
  - a ***weight*** attribute may be stored in metric units in one system and British imperial units in another.
  - For a hotel chain, the ***price*** of rooms in different cities may involve not only different currencies but also different services (such as free breakfast) and taxes.

Schema Integration

Schema Integration

- Schema Integration
  - Integrate metadata from different sources
  - The same attribute or object may have different names in different databases
  - e.g. *customer_id* in one database and *cust_number* in another
- The metadata include:
  - the name, meaning, data type, and range of values permitted for the attribute, and etc.

Handling Redundancy

Handling Redundancy

- Redundancy
  - Redundant data occur often when integration of multiple databases
- Type of redundancies:
  - **Redundant instances**
    - Data set may include data instances that are duplicates, or almost duplicates of one another
  - **Redundant attributes**
    - **Object identification**: The same attribute or object may have different names in different databases
    - **Derivable data**: One attribute may be a "derived" attribute in another table, e.g., annual revenue

Handling Redundancy

- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

- Redundant attributes may be able to be detected by **correlation analysis**
  - **Correlation coefficient** for numerical attributes
  - **Chi-square test** for categorical (discrete) data

Correlation coefficient

- Correlation coefficient (also called Pearson's product moment coefficient)

$$r_{A,B} = \frac{\sum_{i=1}^{N}(a_i - \overline{A})(b_i - \overline{B})}{N\sigma_A\sigma_B} = \frac{\sum_{i=1}^{N}(a_i b_i) - N\,\overline{A}\,\overline{B}}{N\sigma_A\sigma_B}$$

- where
  - N is the number of instances
  - $\overline{A}$ and $\overline{B}$ are the respective means of A and B
  - σ_A and σ_B are the respective standard deviation of A and B
  - $\Sigma(a_i\, b_i)$ is the sum of the AB cross-product

Correlation coefficient

- If:
  - $r_{A,B} > 0$: A and B are positively correlated (A's values increase as B's). The higher, the stronger correlation.
  - $r_{A,B} = 0$: independent
  - $r_{A,B} < 0$: negatively correlated

- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Chi-square Test

- A correlation relationship between two categorical (discrete) attributes, *A* and *B*, can be discovered by a X^2 (**chi-square**) test.

Chi-square Test

- Suppose:
  - A has c distinct values, namely a_1, a_2, …, a_c.
  - B has r distinct values, namely b_1, b_2, …, b_r
  - The data instances described by A and B can be shown as a contingency table, with
    - the c values of A making up the columns and
    - the r values of B making up the rows.
  - Let (A_i, B_j) denote the event that attribute A takes on value *ai* and attribute B *takes on value bj, that is, where* $(A = a_i, B = b_j)$.
  - Each and every possible (A_i, B_j) joint event has its own cell (or slot) in the table.

Chi-square Test

- The X^2 value (also known as the *Pearson* X^2 statistic) is computed as:

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

  - where o_{ij} is the observed frequency (i.e., actual count) of the joint event (A_i, B_j) *and*
  - e_{ij} is the expected frequency of (A_i, B_j)

Chi-square Test

- e_{ij} is the expected frequency of (A_i, B_j), which can be computed as

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{N}$$

- where
  - N is the number of data instances, $count(A=a_i)$ is the number of instances having value a_i *for* A
  - $count(B = b_j)$ is the number of instances having value b_j for B.

Chi-square Test

- The larger the X^2 value, the more likely the variables are related
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count

Chi-Square Calculation: An Example

- Suppose that a group of 1,500 people was surveyed.
- The observed frequency (or count) of each possible joint event is summarized in the contingency table shown

	Play chess	Not play chess	Sum (row)
Like science fiction	250 (90)	200 (360)	450
Not like science fiction	50 (210)	1000 (840)	1050
Sum(col.)	300	1200	1500

- The numbers in parentheses are the expected frequencies (calculated based on the data distribution for both attributes using Equation e_{ij}).
- Are ***like_science_fiction*** and ***play_chess*** correlated?

Chi-Square Calculation: An Example

- For example, the expected frequency for the cell (play_chess, fiction) is

$$e_{11} = \frac{count(play\_chess) * count(like\_science\_fiction)}{N} = \frac{300 * 450}{1500} = 90$$

- Notice that
  - the sum of the expected frequencies must equal the total observed frequency for that row, and
  - the sum of the expected frequencies in any column must also equal the total observed frequency for that column.

Chi-Square Calculation: An Example

- We can get X^2 by:

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$
$$= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.$$

- For this 2 x 2 table, the degrees of freedom are (2-1)(2-1) = 1.
- For 1 degree of freedom, the X^2 value needed to reject the hypothesis at the 0.001 significance level is **10.828** (taken from the table of upper percentage points of the X^2 distribution, typically available from any textbook on statistics).

Chi-square Test

- Since our computed value is above this, we can reject the hypothesis that ***play chess*** and ***preferred reading*** are **independent** and conclude that the two attributes are (strongly) correlated for the given group of people.

References

References

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

The end