Data Preprocessing

UE 141 Spring 2013

Jing Gao
SUNY Buffalo
Outline

• Data

• Data Preprocessing
  – Improve data quality
  – Prepare data for analysis

• Exploring Data
  – Statistics
  – Visualization
Document Data

- Each document is a collection of words
Transaction Data

• A collection of transactions
  – Each transaction involves a set of items

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Diaper, Milk</td>
</tr>
<tr>
<td>4</td>
<td>Beer, Bread, Diaper, Milk</td>
</tr>
<tr>
<td>5</td>
<td>Coke, Diaper, Milk</td>
</tr>
</tbody>
</table>
Graph Data

Aspirin

Yeast protein interaction network

Internet

Co-author network
Sequence Data

DNA Sequence

GGTTCCGCTTTCAAGCCCCCGCGCC
CGCAGGGCCCCGCCCCCGCGCCGTC
GAGAAGGGCCCCTGCGGCGGGGC
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGCGCCAGCGGACAG
GCCAAGTAGAACAACCGCGAAGCGC
TGGGCTGCTGCTGCGACCAGGG
Spatial-Temporal Data

Average Monthly Temperature of land and ocean
Record Data

- Data that consists of a collection of records, each of which consists of a fixed set of attributes

<table>
<thead>
<tr>
<th>Tid</th>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Cheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>
• Collection of data objects and their attributes

• An attribute is a property or characteristic of an object
  – Examples: eye color of a person, temperature, etc.
  – Attribute is also known as variable, field, characteristic, or feature

• A collection of attributes describe an object
  – Object is also known as record, point, case, sample, entity, or instance

<table>
<thead>
<tr>
<th>Tid</th>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Cheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Types of Attribute

• **Categorical Attribute**
  – Has only a finite or countably infinite set of values
  – Examples: zip codes, weather conditions, or the set of words in a collection of documents

• **Numerical Attribute**
  – Has real numbers as attribute values
  – Examples: temperature, height, or weight
Outline

• Data

• Data Preprocessing
  – Improve data quality
  – Prepare data for analysis

• Exploring Data
  – Statistics
  – Visualization
Data Quality Issue

• **Data in the real world is dirty**
  – incomplete: lacking attribute values, lacking certain attributes of interest
    • e.g., occupation=" " (missing data)
  – noisy: containing noise, errors, or outliers
    • e.g., Salary="−10" (an error)
  – inconsistent: containing discrepancies in codes or names, e.g.,
    • Age="42" Birthday="03/07/1997"
    • Was rating “1,2,3”, now rating “A, B, C”
    • discrepancy between duplicate records
Preprocessing

• **Handle missing values**
  – Ignore the records with missing values
  – Estimate missing values

• **Remove outliers**
  – Find and remove those values that are significantly different from the others

• **Resolve conflicts**
  – Merge information from different data sources
  – Find duplicate records and identify the correct information
Prepare Data for Analysis

• Sampling
• Feature selection
• Dimensionality reduction
• Discretization
Sampling

• Goal
  – Extract a subset of records so that the selected records are representative of original data
Attribute Selection

• **Redundant attributes**
  – duplicate much or all of the information contained in one or more other attributes
  – Example: purchase price of a product and the amount of sales tax paid

• **Irrelevant attributes**
  – contain no information that is useful for the data mining task at hand
  – Example: students' ID is often irrelevant to the task of predicting students' GPA
Dimensionality Reduction

• Goal is to reduce the number of attributes by creating a new set of attributes
Discretization

• Binning
  – Convert numerical data into categorical data
  – Divides the range into $N$ intervals

Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into 3 bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
Outline

• Data

• Data Preprocessing
  – Improve data quality
  – Prepare data for analysis

• Exploring Data
  – Statistics
  – Visualization
Center measurement is a summary measure of the overall level of a dataset.

Mean: Summing up all the observation and dividing by number of observations. Mean of 20, 30, 40 is $(20+30+40)/3 = 30$.

Notation: Let $x_1, x_2, \ldots, x_n$ are $n$ observations of a variable $x$. Then the mean of this variable,

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$
Variance: The variance of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of the n observations $x_1, x_2, \ldots, x_n$ is

$$S^2 = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n - 1}$$

Variance of 5, 7, 3? Mean is $(5+7+3)/3 = 5$ and the variance is

$$\frac{(5 - 5)^2 + (3 - 5)^2 + (7 - 5)^2}{3 - 1} = 4$$

Standard Deviation: Square root of the variance. The standard deviation of the above example is 2.
Statistics: Frequency Distribution

Consider a data set of 26 children of ages 1-6 years. Then the frequency distribution of variable ‘age’ can be tabulated as follows:

Frequency Distribution of Age

<table>
<thead>
<tr>
<th>Age</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>5</td>
<td>3</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

Grouped Frequency Distribution of Age:

<table>
<thead>
<tr>
<th>Age Group</th>
<th>1-2</th>
<th>3-4</th>
<th>5-6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>8</td>
<td>12</td>
<td>6</td>
</tr>
</tbody>
</table>
Question

• Data Visualization

– Choose a few creative and fascinating examples of data visualization to show to the class

http://selection.datavvisualization.ch/