Lecture Notes for Chapter 2

Introduction to Data Mining

by

Tan, Steinbach, Kumar

(modified for I211 by P. Radivojac)
What is Data?

- 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 feature, variable, variate
- A collection of attributes describe a data point
  - Data point is also known as object, record, instance, or example

<table>
<thead>
<tr>
<th>Tid</th>
<th>Home Owner</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>
Attribute Values

- Attribute values are numbers or symbols assigned to an attribute

- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value
      - average age is interesting to know, but average ID is meaningless
The way you measure an attribute is something that may not match the attributes properties.
Types of Attributes

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval
    - Examples: calendar dates, temperatures in Celsius or Fahrenheit.
  - Ratio
    - Examples: temperature in Kelvin, length, time, counts
Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: = ≠
  - Order: < >
  - Addition: + -
  - Multiplication: * /

- Nominal attribute: distinctness
- Ordinal attribute: distinctness & order
- Interval attribute: distinctness, order & addition
- Ratio attribute: all 4 properties
<table>
<thead>
<tr>
<th>Attribute Type</th>
<th>Description</th>
<th>Examples</th>
<th>Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)</td>
<td>zip codes, employee ID numbers, eye color, sex: {male, female}</td>
<td>mode, entropy, contingency correlation, $\chi^2$ test</td>
</tr>
<tr>
<td>Ordinal</td>
<td>The values of an ordinal attribute provide enough information to order objects. (&lt;, &gt;)</td>
<td>hardness of minerals, {good, better, best}, grades, street numbers</td>
<td>median, percentiles, rank correlation, run tests, sign tests</td>
</tr>
<tr>
<td>Interval</td>
<td>For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)</td>
<td>calendar dates, temperature in Celsius or Fahrenheit</td>
<td>mean, standard deviation, Pearson's correlation, $t$ and $F$ tests</td>
</tr>
<tr>
<td>Ratio</td>
<td>For ratio variables, both differences and ratios are meaningful. (*, /)</td>
<td>temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current</td>
<td>geometric mean, harmonic mean, percent variation</td>
</tr>
<tr>
<td>Attribute Level</td>
<td>Transformation</td>
<td>Comments</td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>----------------</td>
<td>----------</td>
<td></td>
</tr>
<tr>
<td>Nominal</td>
<td>Any permutation of values</td>
<td>If all employee ID numbers were reassigned, would it make any difference?</td>
<td></td>
</tr>
<tr>
<td>Ordinal</td>
<td>An order preserving change of values, i.e., $new_value = f(old_value)$ where $f$ is a monotonic function.</td>
<td>An attribute encompassing the notion of good, better best can be represented equally well by the values ${1, 2, 3}$ or by ${0.5, 1, 10}$.</td>
<td></td>
</tr>
<tr>
<td>Interval</td>
<td>$new_value = a \times old_value + b$ where $a$ and $b$ are constants</td>
<td>Thus, the Fahrenheit and Celsius temperature scales differ in terms of where their zero value is and the size of a unit (degree).</td>
<td></td>
</tr>
<tr>
<td>Ratio</td>
<td>$new_value = a \times old_value$</td>
<td>Length can be measured in meters or feet.</td>
<td></td>
</tr>
</tbody>
</table>
Discrete and Continuous Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.
  - Note: binary attributes are a special case of discrete attributes

- **Continuous Attribute**
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.
  - Continuous attributes are typically represented as floating-point variables.
Types of data sets

- **Record**
  - Data Matrix
  - Document Data
  - Transaction Data

- **Graph**
  - World Wide Web
  - Molecular Structures

- **Ordered**
  - Spatial Data
  - Temporal Data
  - Sequential Data
  - Genetic Sequence Data
Important Characteristics of Structured Data

- Dimensionality
  - Curse of Dimensionality

- Sparsity
  - Only presence counts

- Resolution
  - Patterns depend on the scale

- Attribute and Class Imbalance
  - small number of non zero elements (related to sparsity)
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>Home Owner</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>
```
Data Matrix

- If data objects have the same fixed set of numeric attributes, then the data objects can be thought of as points in a multi-dimensional space, where each dimension represents a distinct attribute.

- Such data set can be represented by an m-by-n matrix, where there are m rows, one for each object, and n columns, one for each attribute.

<table>
<thead>
<tr>
<th>Projection of x Load</th>
<th>Projection of y load</th>
<th>Distance</th>
<th>Load</th>
<th>Thickness</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.23</td>
<td>5.27</td>
<td>15.22</td>
<td>2.7</td>
<td>1.2</td>
</tr>
<tr>
<td>12.65</td>
<td>6.25</td>
<td>16.22</td>
<td>2.2</td>
<td>1.1</td>
</tr>
</tbody>
</table>
Document Data

- Each document becomes a `term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

<table>
<thead>
<tr>
<th></th>
<th>team</th>
<th>coach</th>
<th>y</th>
<th>pla</th>
<th>ball</th>
<th>score</th>
<th>game</th>
<th>n</th>
<th>wi</th>
<th>lost</th>
<th>timeout</th>
<th>season</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document 1</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document 2</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document 3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example
Matlab Code

% read from pre-prepared files and count some words

dictionary = {'gaza', 'fuel', 'the', 'patriots'};

fid = fopen('1.txt', 'rt');
s = textscan(fid, '%s');
fclose(fid);
s{1} = lower(s{1});

fid = fopen('2.txt', 'rt');
t = textscan(fid, '%s');
fclose(fid);
t{1} = lower(t{1});

for i = 1 : length(dictionary)
    D(1, i) = length(strmatch(dictionary{i}, s{1}));
    D(2, i) = length(strmatch(dictionary{i}, t{1}));
end
Transaction Data

- A special type of record data, where
  - each record (transaction) involves a set of items.
  - For example, consider a grocery store. The set of products purchased by a customer during one shopping trip constitute a transaction, while the individual products that were purchased are the 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>
Machine Learning Repository at UCI

- contains a number of user deposited ML problems

Discussion:
- Pima Indians diabetes example ([link](#))
- Boston housing example ([link](#))
- German credit example ([link](#))
How to load certain data formats

- From web sites
  - use readurl function

- From Excel files
  - use xlsread function

- From text files
  - use textscan and related functions

- From CSV files
  - use csvread function

- Many files are unstructured, parsing is needed
Reading Custom File Types

- Need standard I/O for this
  - use fopen, fclose, fgetl, fget for text files
  - use fread, fwrite, fseek, ftell for binary files

- Examples
  - reading protein sequence data
  - reading and writing binary data
Graph Data

- Examples: Generic graph and HTML Links

Scientific diagrams typically use specific symbols and representations to convey information. The diagram shows a generic graph with nodes and edges labeled with numbers. The text associated with each node is as follows:

- Data Mining
- Graph Partitioning
- Parallel Solution of Sparse Linear System of Equations
- N-Body Computation and Dense Linear System Solvers
Chemical Data

- Benzene Molecule: $\text{C}_6\text{H}_6$
Ordered/Sequential Data

- Sequences of transactions

```
( A B)  (D)  (C E)
( B D)  (C)  (E)
( C D)  (B)  (A E)
```

An element of the sequence
**Ordered/ Sequential Data**

- **Genomic sequence data**

<table>
<thead>
<tr>
<th></th>
<th>Sequence</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>...GGTTCCGCCTTCAGCCCCCCCGCC...</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>...GGTTCCGCCTTCAGCCCCGGCGCC...</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>...GGTTCCGCCTTCAGCCCCCGGCC...</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>...GGTTCCGCCTTCAGCCCCCGGCC...</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>...GGTTCCGCCTTCAGCCCCCGGCC...</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>...GGTTCCGCCTTCAGCCCCCGGCC...</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>...GGTTCCGCCTTCAGCCCCCGGCC...</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>...GGTTCCGCCTTCAGCCCCCGGCC...</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>...GGTTCCGCCTTCAGCCCCCGGCC...</td>
<td>0</td>
</tr>
</tbody>
</table>

© Tan, Steinbach, Kumar

Introduction to Data Mining

4/18/2004
Ordered Data

- Spatio-Temporal Data

Average Monthly Temperature of land and ocean

Aug
Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

- Examples of data quality problems:
  - Noise and outliers
  - missing values
  - duplicate data
Noise

- Noise refers to modification of original values
  - Examples: distortion of a person’s voice when talking on a poor phone and “static” on television screen
Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set.
Missing Values

- Reasons for missing values
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)

- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)
Duplicate Data

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

- Examples:
  - Same person with multiple email addresses

- Data cleaning
  - Process of dealing with noise and duplicate data issues
Understanding data collection process

- Sometimes, the whole population is available for labeling (say, we can provide features and a the class with limited resources for any object in the population). In such a case we’d like to select a sample that is a good representative of the population.

- However, there are situations where we do not have control of the data at hand. It is very important to understand the mechanism how the data was generated!!!

- Example 1: Bank loan data
  - all people eligible to apply for a loan > all people who apply > all people who are accepted > all people who take the loan

  - Banks can only study behavior of the people who take the loan. Banks can only make inferences about a subset of the overall population.

- Example 2: 1936 Presidential elections in the USA (Roosevelt vs. Landon)

- Example 3: 2007 Democratic primary in New Hampshire (Obama vs. Clinton)