Overview

Ji Zhu
445C West Hall
734-936-2577
jizhu@umich.edu
What is Data Mining?

- Data mining is a multi-disciplinary field of study concerned with the design of algorithms that allow computers to learn from large data repositories.

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data

- There are many other definitions.
Data Mining Examples and Non-Examples

• Data mining
  – Certain names are more prevalent in certain US locations (O’Brien, O’Rurke, O’Reilly... in Boston area)
  – Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com, etc.)

• Not data mining
  – Look up phone number in phone directory
  – Query a web search engine for information about “Amazon”
Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite (NASA)
  - telescopes scanning the skies (SDSS)
  - microarrays generating gene expression data (MEDLINE)
  - scientific simulations generating terabytes of data (GIS)
• Data mining may help scientists
  – in classifying and segmenting data
  – in hypothesis formation
  – etc
Why Mine Data? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data, e-commerce (Google, Yahoo, Amazon, Ebay)
  - Purchases at department and grocery stores (Walmart)
  - Bank/credit card transactions (Bank of America, Visa, Mastercard)
• Competitive pressure is strong
  – Provide better, customized services for an edge
Mining Large Data Sets - Motivation

- There is often information "hidden" in the data that is not readily evident.
- Human analysts may take months to discover useful information.
- Much of the data is never analyzed at all.
Technological Driving Factors

• Larger, cheaper memory

• Faster, cheaper processors
  – the CRAY of 20 years ago is now on your desk

• Success of relational databases and the Web
  – everybody is a “data owner”

• New ideas in machine learning and statistics
Origins of Data Mining

- Draws ideas from machine learning, pattern recognition, statistics and database systems.
• Traditional techniques may be unsuitable due to
  – enormity of data
  – high dimensionality of data
  – heterogeneous, distributed nature of data
Data Mining vs Statistics

- Traditional statistics
  - first hypothesize, then collect data, then analyze
  - often model-oriented

- Data mining
  - few if any a priori hypothesis
  - often algorithm-oriented rather than model-oriented
• Different?
  – Yes, in terms of culture, motivation; however...
  – Statistical ideas are very useful in data mining, e.g., in validating whether discovered knowledge is useful

• Increasing overlap at the boundary of statistics and data mining: use the tools of probability and statistics to provide a mathematical framework for
  – posing data mining problems
  – formulating solutions to those problems
Data Mining vs Machine Learning

- To first-order, very little difference...
  - Data mining relies heavily on ideas from machine learning (and from statistics)

- Some differences between data mining and machine learning
  - More emphasis in data mining on scalability
  - Data mining is somewhat more applications-oriented
Two Types of Data Mining Tasks

- **Prediction methods**: Use some variables to predict unknown or future values of other variables
- **Description methods**: Find human-interpretable patterns that describe the data
Examples of Data Mining Tasks

- Visualization (Descriptive)
- Classification (Predictive)
- Regression (Predictive)
- Association analysis (Descriptive)
- Clustering (Descriptive)
Classification: Definition

- Given a collection of data points (training set)
  - Each data point contains a set of variables, one of the variables is the class

- Find a model for the class variable as a function of the values of other variables

- Goal: previously unseen data points should be assigned a class as accurately as possible
  - A test set is used to determine the accuracy of the model
Classification Example: Customer Scoring

- Example: a bank has a database of 1 million past customers, 10% of whom took out mortgages

- Use data mining to predict whether a new customer will take out a “mortgage or not” based on the customer data

- Customer data
  - Other credit data
  - Demographic data on the customer
Classification Example: Spam Detection

Customize an email spam detection system for individual user. Relative frequencies in a message of most commonly occurring words and punctuation marks.

<table>
<thead>
<tr>
<th></th>
<th>george</th>
<th>you</th>
<th>your</th>
<th>hp</th>
<th>free</th>
<th>re</th>
<th>remove</th>
</tr>
</thead>
<tbody>
<tr>
<td>spam</td>
<td>0.00</td>
<td>2.26</td>
<td>1.38</td>
<td>0.02</td>
<td>0.52</td>
<td>0.13</td>
<td>0.28</td>
</tr>
<tr>
<td>email</td>
<td>1.27</td>
<td>1.27</td>
<td>0.44</td>
<td>0.90</td>
<td>0.07</td>
<td>0.42</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Classification Example: Microarray
Deviation/Anomaly Detection

- Detect significant deviations from normal behavior

- Applications:
  - Credit card fraud detection
  - Network intrusion detection
Fraud Detection

- Credit card fraud detection
  - Credit card losses in the US are over 1 billion $ per year
  - Roughly 1 in 50k transactions are fraudulent

- Fair-Issac’s fraud detection software based on neural networks, led to reported fraud decreases of 30 to 50%

- Issues: false alarm rate vs missed detection – what is the tradeoff?
Regression

• Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency

• Examples
  – Predicting sales amounts of new product based on advertising expenditure
  – Predicting wind velocities as a function of temperature, humidity, air pressure, etc
  – Time series prediction of stock market indices
Clustering: Definition

- Given a set of data points, each having a set of variables, find clusters such that
  - data points in one cluster are more “similar” to one another, and
  - data points in separate clusters are “less similar” to one another.
• Similarity measures
  – Euclidean distance if variables are continuous
  – Other problem-specific measures
Cluster Example: Market Segmentation

- Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.

- Collect different variables of customers based on their geographical and lifestyle related information

- Find clusters of similar customers
Clustering Examples

- **Document clustering**: find groups of documents that are similar to each other based on the important terms appearing in them.

- **Clustering stocks** based on their movements every day.
Association Rule Discovery: Definition

- Given a set of records each of which contains some number of items from a given collection
  - Produce *dependency rules* which will predict occurrence of an item based on occurrences of other items

- Goal is to discover interesting *local* patterns in the data rather than to characterize the data globally
<table>
<thead>
<tr>
<th>ID</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>

Rules discovered

- \{\text{Coke}\} \implies \{\text{Milk}\}
- \{\text{Diaper, Milk}\} \implies \{\text{Beer}\}
Association Rule Discovery Example

- Supermarket shelf management
  - Goal: To identify items that are bought together by sufficiently many customers
  - A classic rule: If a customer buys diaper and milk, then he is very likely to buy beer

- Amazon, Netflix
Data Mining: the Downside

- Hype?

- Data dredging, snooping and fishing
  - Finding spurious structure in data that is not real

- Historically, “data mining” was derogatory term in the statistics community
  - Rhine paradox
  - The Super Bowl fallacy
  - Bangladesh butter prices and the US stock market