Recitation 1
XLMiner Tutorial

15.062 Data Mining: Algorithms and Application
Spring 2003 H2

Review of Data Mining Fundamentals

**Data Mining**: Identify patterns and relationships in large data sets

- aka: Machine Learning, Computational Statistics, Knowledge Discover, Artificial Intelligence

Example: Fleet has enormous amounts of data on every individual who has taken out a loan from them

- From the data, can they accurately predict which future clients are like to default?
The Data:

<table>
<thead>
<tr>
<th>ID</th>
<th>default</th>
<th>age</th>
<th>avg_inc</th>
<th>curr_inc</th>
<th>house</th>
<th>score</th>
<th>p-index</th>
</tr>
</thead>
<tbody>
<tr>
<td>977321</td>
<td>0</td>
<td>31</td>
<td>65,000</td>
<td>65,000</td>
<td>1</td>
<td>0.2196803</td>
<td>0.4591304</td>
</tr>
<tr>
<td>977322</td>
<td>1</td>
<td>56</td>
<td>89,000</td>
<td>84,000</td>
<td>1</td>
<td>0.7302223</td>
<td>0.7202815</td>
</tr>
<tr>
<td>977323</td>
<td>1</td>
<td>27</td>
<td>63,000</td>
<td>170,000</td>
<td>0</td>
<td>0.1805354</td>
<td>0.2007951</td>
</tr>
<tr>
<td>977324</td>
<td>0</td>
<td>47</td>
<td>78,000</td>
<td>54,000</td>
<td>0</td>
<td>0.7195863</td>
<td>0.7981367</td>
</tr>
</tbody>
</table>

Columns: variables, features, attributes,....

Rows: observations, data points, cases, records, patterns...

Inputs: independent variables, explanatory variables, predictors...

Output: dependent variables, predicted variable, target, outcome....

Goal: Given these historical records, build a model that can accurately predict defaults (target) given new customers (observations)

Review of Data Mining Fundamentals

\[ f(\cdot) \]

\[ y = f(x_1, x_2, x_3, \ldots) \]

• Supervised Learning – Goal is to predict the value of an output based on inputs
  • Classification, Regression
  • Unsupervised Learning – No output. Determine/describe patterns in the inputs.
    • Clustering, Association

Challenge in supervised learning is to determine what model to use on the data to get an accurate prediction of the output variable
Review of Data Mining Fundamentals

Types of Supervised Learning Problems:

- **Classification Problems**: Output categorical/discrete
  - E.g., default or no default, fraud or not fraud, cancer or no cancer,
  - Models: k-nearest neighbor, naïve Bayes, classification trees,…

- **Regression Problems**: Output continuous
  - E.g., value of house, price of asset, expected payoff
  - Models: linear regression, regression trees,…

![Diagram of data mining model](image)

**Input variables**
(e.g., income, credit score, age
\[x_1, x_2, x_3, \ldots\]

**Data Mining Model**

**Output**
(e.g., default)

\[y = f(x_1, x_2, x_3, \ldots)\]

Classification: discrete
Regression: continuous

Model Building Procedures

- **Partition data into 2 or 3 parts:**
  - Training Set
  - Validation set
  - Testing set

- **Build Model:**
  - Build preliminary Model
  - Test Model
  - Test Final Model

Fine tune parameters
• Download instructions of “News” section of course website
• Online tutorial available with software and on the web:
  http://www.resamplecom/xlminer/help/Index.htm
• For additional information about XLMiner:
  http://www.xlminer.com