The Art and Science of Data Mining

By Marianna Dizik,
Williams-Sonoma, Inc.

What is Data Mining?

- Data Mining is the process of exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules

What Can Data Mining Do?

- Directed Data Mining:
  - Classification
  - Estimation
  - Prediction
- The goal:
  - To build a model that describes/predict one particular variable of interest in terms of the rest of available data

- Undirected Data Mining:
  - Affinity grouping or association rules
  - Clustering
  - Description or visualization
- The goal:
  - To establish some relationship among all the variables

The Business Context for Data Mining

- DM as a Research Tool
  - Pharmaceutical, bioinformatics
- DM for Process Improvement
  - Manufacturing process control
- DM for Marketing
- DM for Customer Relationship Management

Technical Context for Data Mining

- DM and Machine Learning:
  - AI, pattern recognition
- DM and Statistics
- DM and Decision Support
  - Data Warehouse:
  - OLAP, Data Marts, and Multidimensional Databases
    - Single view of the data to increase speed and responsiveness
- DM and Computer Technology

The Societal Context for Data Mining

- Privacy

Classical Techniques: Statistics

- Descriptive Statistics and Visualization
- Statistic for Prediction:
  - Linear Regression
    - Power Transformation
  - WLS Regression
  - Logistic Regression: $\ln(p/(1-p))=a+Bx$
    - $p/(1-p)$ – odds ratio
    - $\ln(p/(1-p))$ - logit

Classic Techniques: Neighborhoods

- Clustering: grouping similar objects
  - Hierarchical and Non-Hierarchical
    - Clustering for clarity (data reduction)
    - Clustering for action (policies, rules, promotions)
    - Clustering for outliers (fraud detection)
- Nearest Neighbor Prediction
    - Recommendation Engines
    - Response Prediction
    - Stock Market Predictions
    - Building Taxonomies
    - Text Mining, Spam Identification

Next Generation Techniques: Trees

- Decision Trees: Segmentation with a Purpose
  - Exploration
  - Preprocessing
  - Prediction
    - CHAID (Chi-Square Interaction Detector)
    - CART (Gini metrics)

Next Generation Techniques: Neural Networks

- Neural Networks: Black Box Approach
  - Supervised Learning
    - Predictions
  - Unsupervised Learning
    - Clustering
      - Kohonen Network
    - Outlier Analysis
    - Feature Extraction
- NN Models
  - Input Nodes
  - Hidden Nodes
  - Output Nodes

Next Generation Techniques: Rule Induction

- Rule Induction: Knowledge Discovery in unsupervised learning system
  - "If pickles are purchased, then ketchup is purchased"
- Rule +Accuracy + Coverage
- Target
  - The Antecedent
  - The Consequent
    - Based on Accuracy
    - Based on Coverage
    - Based on "Interestingness"
- Rules do not Imply Causality

Conclusion: Next Generation Business Intelligence and Data Mining

- Information Mining:
  - Uncover associations, patterns and trends
  - Detect deviations
  - Group and classify information
  - Develop predictive models
- Knowledge Management (KM)
  - Text Mining
- Data “Archeology”:
  - Semantic Computing Tools
  - Artificial Perception Systems