Data Mining: Overview

What is Data Mining?

- Recently* coined term for confluence of ideas from statistics and computer science (machine learning and database methods) applied to large databases in science, engineering and business.
- In a state of flux, many definitions, lot of debate about what it is and what it is not. Terminology not standard e.g. bias, classification, prediction, feature = independent variable, target = dependent variable, case = exemplar = row.

* First International workshop on Knowledge Discovery and Data Mining was in 1995

Broad and Narrow Definitions

- Broad Definition includes traditional statistical methods, Narrow Definition emphasizes automated and heuristic methods
- Data mining, data dredging, fishing expeditions
- Knowledge Discovery in Databases (KDD)

My Favorite

- "Statistics at scale and speed"
  Darryl Pregibon

- My extension:
  - " . . . And simplicity"

Gartner Group

- “Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques.”

Drivers

- Market: From focus on product/service to focus on customer
- IT: From focus on up-to-date balances to focus on patterns in transactions - Data Warehouses - OLAP
- Dramatic drop in storage costs : Huge databases
  - e.g Walmart: 20 million transactions/day, 10 terabyte database, Blockbuster: 36 million households
- Automatic Data Capture of Transactions
  - e.g. Bar Codes , POS devices, Mouse clicks, Location data (GPS, cell phones)
- Internet: Personalized interactions, longitudinal data

Core Disciplines

- Statistics (adapted for 21st century data sizes and speed requirements). Examples:
  - Descriptive: Visualization
  - Models (DMD): Regression, Cluster Analysis
- Machine Learning: e.g. Neural Nets
- Data Base Retrieval: e.g. Association Rules
- Parallel developments: e.g. Tree methods, k Nearest Neighbors, OLAP-EDA

Process

1. Develop understanding of application, goals
2. Create dataset for study (often from Data Warehouse)
3. Data Cleaning and Preprocessing
4. Data Reduction and projection
5. Choose Data Mining task
6. Choose Data Mining algorithms
7. Use algorithms to perform task
8. Interpret and iterate thru 1-7 if necessary
9. Deploy: integrate into operational systems.

Data Mining

SEMMA Methodology (SAS)

- **S**ample from data sets, Partition into Training, Validation and Test datasets
- **E**xplore data set statistically and graphically
- **M**odify:Transform variables, Impute missing values
- **M**odel: fit models e.g. regression, classfication tree, neural net
- **A**ssess: Compare models using Partition, Test datasets

Illustrative Applications

- Customer Relationship Management
- Finance
- E-commerce and Internet

Customer Relationship Management

- Target Marketing
- Attrition Prediction/Churn Analysis
- Fraud Detection
- Credit Scoring

Target marketing

- Business problem: Use list of prospects for direct mailing campaign
- Solution: Use Data Mining to identify most promising respondents combining demographic and geographic data with data on past purchase behavior
- Benefit: Better response rate, savings in campaign cost

Example: Fleet Financial Group

- Redesign of customer service infrastructure, including $38 million investment in data warehouse and marketing automation
- Used logistic regression to predict response probabilities to home-equity product for sample of 20,000 customer profiles from 15 million customer base
- Used CART to predict profitable customers and customers who would be unprofitable even if they respond

Churn Analysis: Telcos

- Business Problem: Prevent loss of customers, avoid adding churn-prone customers
- Solution: Use neural nets, time series analysis to identify typical patterns of telephone usage of likely-to-defect and likely-to-churn customers
- Benefit: Retention of customers, more effective promotions

Example: France Telecom

- CHURN/Customer Profiling System implemented as part of major custom data warehouse solution
- Preventive CPS based on customer characteristics and known cases of churning and non-churning customers identify significant characteristics for churn
- Early detection CPS based on usage pattern matching with known cases of churn customers.

Fraud Detection

- Business problem: Fraud increases costs or reduces revenue
- Solution: Use logistic regression, neural nets to identify characteristics of fraudulent cases to prevent in future or prosecute more vigorously
- Benefit: Increased profits by reducing undesirable customers

Example: Automobile Insurance Bureau of Massachusetts

- Past reports on claims adjustors scrutinized by experts to identify cases of fraud
- Several characteristics (over 60) of claimant, type of accident, type of injury/treatment coded into database
- Dimension Reduction methods used to obtain weighted variables. Multiple Regression Step-wise Subset selection methods used to identify characteristics strong correlated with fraud

Risk Analysis

- Business problem: Reduce risk of loans to delinquent customers
- Solution: Use credit scoring models using discriminant analysis to create score functions that separate out risky customers
- Benefit: Decrease in cost of bad debts

Finance

- Business problem: Pricing of corporate bonds depends on several factors, risk profile of company , seniority of debt, dividends, prior history, etc.
- Solution Approach: Through DM, develop more accurate models of predicting prices.

E-commerce and Internet

- Collaborative Filtering
- From Clicks to Customers

Recommendation systems

- Business opportunity: Users rate items (Amazon.com, CDNOW.com, MovieFinder.com) on the web. How to use information from other users to infer ratings for a particular user?
- Solution: Use of a technique known as collaborative filtering
- Benefit: Increase revenues by cross selling, up selling

Clicks to Customers

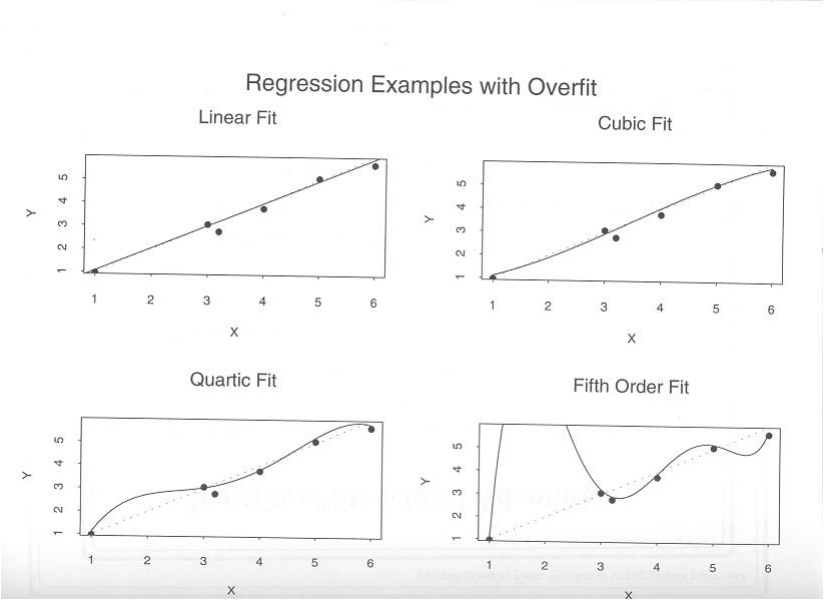
- Business problem: 50% of Dell's clients order their computer through the web. However, the retention rate is 0.5%, i.e. of visitors of Dell's web page become customers.
- Solution Approach: Through the sequence of their clicks, cluster customers and design website, interventions to maximize the number of customers who eventually buy.
- Benefit: Increase revenues

Emerging Major Data Mining applications

- Spam
- Bioinformatics/Genomics
- Medical History Data – Insurance Claims
- Personalization of services in e-commerce
- RF Tags : Gillette
- Security :
  - Container Shipments
  - Network Intrusion Detection

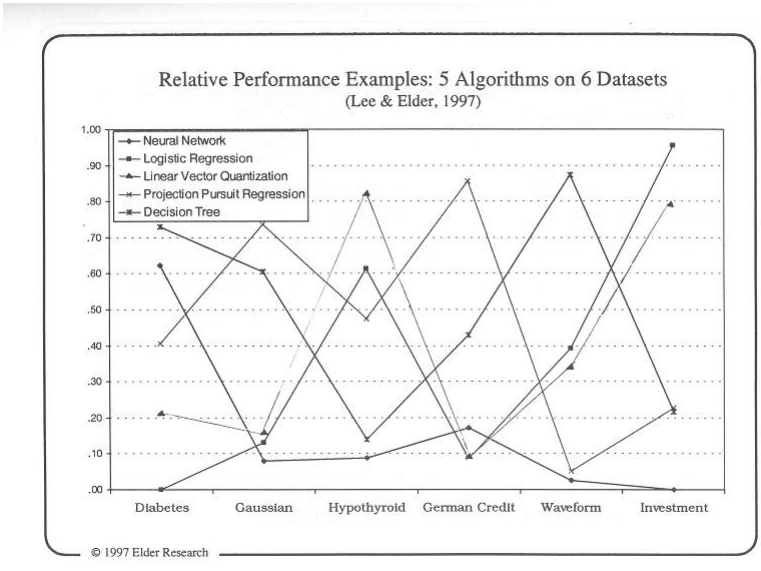
Core Concepts

- Types of Data:
  - Numeric
    - Continuous – ratio and interval
    - Discrete
    - Need for Binning
  - Categorical – order and unordered
  - Binary
- Overfitting and Generalization
- Regularization: Penalty for model complexity
- Distance
- Curse of Dimensionality
- Random and stratified sampling, resampling
- Loss Functions



Typical characteristics of mining data

- "Standard" format is spreadsheet:
  - Row=observation unit, Column=variable
- Many rows, many columns
- Many rows moderate number of columns (e.g. tel. calls)
- Many columns, moderate number of rows (e.g. genomics)
- Opportunistic (often by-product of transactions)
  - Not from designed experiments
  - Often has outliers, missing data



Course Topics

- Supervised Techniques
  - Classification:
    - k-Nearest Neighbors, Naïve Bayes, Classification Trees
    - Discriminant Analysis, Logistic Regression, Neural Nets
  - Prediction (Estimation):
    - Regression, Regression Trees, k-Nearest Neighbors
- Unsupervised Techniques
  - Cluster Analysis, Principal Components
  - Association Rules, Collaborative Filtering