Data Mining

A. Ronald Gallant
arg@duke.edu

Reading Assignment

Thomas H. Davenport, "Competing on Analytics," *Harvard Business Review,* January 2006.

Berry and Linoff (2000)

- Pages 7–11. What is data mining, what can it do.
- Pages 16–18, 48–53. Data storage, sources, preparation.
- Pages 61–64. Assumptions of data mining.

Definition

Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules.

Berry and Linoff, 1997

Strategic Weapon

- Data mining has been a core competency at a large number of firms for some time.

- It is beginning to be used as a strategic weapon that has allowed some firms to achieve industry dominance.

- Examples are Amazon and Harrahs.

Thomas H. Davenport, "Competing on Analytics," *Harvard Business Review,* January 2006.

Course Objective

To bring you to a level of proficiency in data mining so that you can advise clients as a consultant, manage data mining projects, or participate in data mining projects.

Upon completion of the course you should understand the capabilities of data mining, know which tools are available, be able to assess the quality of the work of data mining specialists, and be able to use data mining tools yourself.

Synonyms

- Data Mining
- Machine Learning
- Statistical Learning
- Knowledge Discovery
- Artificial Intelligence (subset of)

Main Subdivisions of the Subject

- Supervised Learning
  - The goal is to predict the value of an outcome measure based on a number of input measures.
- Unsupervised Learning
  - No output measure; the goal is to describe associations and patterns in input measures.

Some Examples

- Predict offer response from a transactions database.
- Predict default of a consumer loan from creditor data.
- Detect fraud from billing records.
- Predict churning from a transactions database.
- Detect network attacks from traffic data.
- Detect spam from word patterns in e-mail headers and body.

Focus of the Course

This course will cover the primary tools of supervised learning — regression, decision trees, neural nets, nearest neighbors, discriminant analysis, logistic regression — and unsupervised learning — principal components, hierarchical clustering, K-means, association rules — within a case context.

We shall presume that data sets are large which allows the use of the typical machine learning paradigm of splitting the data into a learning sample and one or more validation samples and also compels the use of methods that use machine time and space efficiently.

Organization of Lectures

The course is organized around cases: Ideas and tools are introduced and compared within cases.

The first two cases are simulations because verification of the validity of basic data mining principles requires comparison to the truth.

The next four cases are real data mining applications.

Topics

- Case 1. Credit Scoring
  - Topic 1. Overview of the Course
    - Least Squares Regression
    - Nearest Neighbors
    - Decision Trees
    - Neural Nets
- Case 2. Single Target Single Feature
  - Topic 2. The Bias-Variance Trade-Off
    - Decision Trees

Topics (continued)

- Case 3. Donor Recapture
  - Topic 3. Linear Regression
  - Topic 4. Neural Networks
  - Topic 5. Decision Trees
  - Topic 6. Interactions
- Case 4. Value at Risk
  - Topic 7. Cluster Analysis
    - Principal Components
    - Hierarchical Clustering
    - K-Means
- Case 5. Intrusion Detection
  - Topic 8. Classification
    - Linear Discriminant Analysis
    - Quadratic Discriminant Analysis
    - Logistic Regression

Topics (continued)

- Case 6. Web Viewership Associations
  - Topic 9. Association Rules
- Case 3. Donor Recapture
  - Topic 10. Classification, Oversampling, and Lift

Texts

Berry, Michael J. A., and Gordon Linoff (2000), *Mastering Data Mining, The Art and Science of Customer Relationship Management*, John Wiley & Sons, New York, ISBN 0-471-33123-6.

These slides
http://www.faculty.fuqua.duke.edu/courses/2008/fall1/mgrecon491
or
http://www.econ.duke.edu/∼arg/datamine

Software

XLMiner http://www.resample.com/xlminer

Course Style

Case studies undertaken by the instructor (these slides).

In-class discussion.

PowerPoint presentations of homework by students.

Assigned readings from the text.

Course Performance Evaluation

- Homework 40%.
  - Homework assignments are posted on the website.
  - You may form teams of size two.

- Final 30%.
  - The final exam is posted on the course website.
  - The final is an individual effort.
  - Work on the final as the course progresses.
  - Turn the final in on the last day of class.

- Participation 30%.
  - Homework and final exam presentations.
  - Comments and questions during lecture.

Commercial Data Mining Software

- Teradata, NCR
  www.teradata.com/main
- Darwin, Oracle
  www.oracle.com/ip/analyze/warehouse/datamining
- Enterprise Miner, SAS
  www.sas.com/products/miner/index.html
- Clementine, SPSS
  www.spss.com/spssbi/clementine
- Insightful Miner, Insightful
  www.insightful.com/products
- R, The R Foundation for Statistical Computing (GNU affiliate)
  www.r-project.org

Dominant Players: R & SAS

- Consulting experience.
- Informal survey.
- Reports from students of this course.

R

- R is a public domain statistical package that originated at Bell Labs as S.
- Its advantage, aside from being free, is that it is the research language of statistics and new ideas are first available in R.
- It offers maximum flexibility in an analysis because it is more an object oriented scripting language than it is a statistical package.
- Its disadvantage is that it has a steep learning curve.
- R will be demonstrated in class.

SAS

- Its main advantage is that it runs on all platforms, can communicate across platforms, and can warehouse data across platforms.
- Unbeatable data management, manipulation, and reporting capabilities.
- Interfaces with industry standard data base software – Oracle, Microsoft, IBM, SAP.
- Its statistical capabilities are actually a minor factor in its popularity.
- Enterprise Miner will be demonstrated in class.

XLMiner

- Excel add in – builds on existing skills.
- Easily learned – menu driven.
- Minimal feature bloat – what is essential is present and most of what is inessential is absent.
- Useful – with subsampling can use it on the job for idea development, checking results reported to you, etc.

Vocabulary

- Supervised learning = directed data mining
- Unsupervised learning = undirected data mining
- Input = independent variable = attribute = field = feature = predictor
- Output = dependent variable = target
- Observation = record = case = example = instance
- Learning data = training data
- Validation data = evaluation data = hold-out sample = test data

Several disciplines, each with its own vocabulary, contribute to the subject of data mining, which causes this heterogeneity in terminology. Fortunately, the technical meaning of each of these terms is close to its dictionary meaning.

Blank page