MS1b Statistical Data Mining

Yee Whye Teh
Department of Statistics
Oxford

http://www.stats.ox.ac.uk/~teh/datamining.html

Outline

Outline

Course Structure

Lectures

- Wednesdays 1100-1200, Weeks 1-8.
- Thursdays 1100-1200, Weeks 1,3,5,7.

Problem Sheets

- 7 problem sheets: due Mondays at noon, Weeks 2-8.

Part C students

- Practical classes: Thursdays 1100-1200, Weeks 2,4,6,8.
- Problem classes: Wednesdays time to be decided, Weeks 2-8.

MSc students

- Miniproject: over Easter break.

Outline

Syllabus I

Part I: Dimensionality Reduction

- Principal Components Analysis
- Multidimensional Scaling
- Isomap

Part II: Clustering

- Hierarchical clustering
- K-means
- Vector Quantization
- Mixture Models
- Probabilistic Latent Variable Models and EM algorithm

Part III: Classification and Regression

- Empirical Risk Minimization
- Nearest Neighbours, Prototype Based Methods
- Classification and Regression Trees
- Linear Regression

Syllabus II

- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Naive Bayes
- Bayesian Methods
- Logistic Regression
- Neural Networks

Part IV: Ensemble Methods

- Bootstrap, Bagging
- Random Forests
- Boosting

R

- Learning how to use R for Data Mining

Outline

What is Data Mining?

Traditional Problems in Applied Statistics

Well formulated question that we would like to answer.
Expensive to gathering data and/or expensive to do computation.
Create specially designed experiments to collect high quality data.

Current Situation

Information Revolution

- improvements in data storage devices (both larger and cheaper).
- powerful data capturing devices (bioassays, microphones, cameras, satellites).

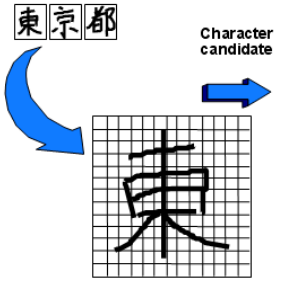
$\rightarrow$ lots of data with potentially valuable information available.
$\rightarrow$ Big Data....

What is Data Mining?

- To gain insight from data.
- Often working with huge datasets.
  - Typically many variables (up to thousands or millions).
  - Often, but not always many observations (dozens to millions).
- Secondary data sources possibly collected for other purposes.
- Uncurated data, missing data, unstructured data, multi-aspect data.
- Gain understanding without specific goals.

Applications of Data Mining

- Pattern Recognition



  - Sorting Cheques
  - Reading License Plates
  - Sorting Envelopes
  - Eye/ Face/ Fingerprint Recognition

Applications of Data Mining

- Business applications
  - Help companies intelligently find information
  - Credit scoring
  - Predict which products people are going to buy
  - Recommender systems
  - Autonomous trading
- Scientific applications
  - Predict cancer occurence/type and health of patients/personalized health
  - Make sense of complex physical, biological, ecological, sociological models

...It is just a nice name for multivariate statistics (‘minus model checking’).

NY Times: Data Mining in Walmart (URL)

November 14, 2004

What Wal-Mart Knows About Customers' Habits

By CONSTANCE L. HAYS

HURRICANE FRANCES was on its way, barreling across the Caribbean, threatening a direct hit on Florida's Atlantic coast. Residents made for higher ground, but far away, in Bentonville, Ark., executives at Wal-Mart Stores decided that the situation offered a great opportunity for one of their newest data-driven weapons, something that the company calls predictive technology.

A week ahead of the storm's landfall, Linda M. Dillman, Wal-Mart's chief information officer, pressed her staff to come up with forecasts based on what had happened when Hurricane Charley struck several weeks earlier. Backed by the trillions of bytes' worth of shopper history that is stored in Wal-Mart's computer network, she felt that the company could "start predicting what's going to happen, instead of waiting for it to happen," as she put it.

The experts mined the data and found that the stores would indeed need certain products - and not just the usual flashlights. "We didn't know in the past that strawberry Pop-Tarts increase in sales, like seven times their normal sales rate, ahead of a hurricane," Ms. Dillman said in a recent interview. "And the pre-hurricane top-selling item was beer."

Thanks to those insights, trucks filled with toaster pastries and six-packs were soon speeding down Interstate 95 toward Wal-Marts in the path of Frances. Most of the products that were stocked for the storm sold quickly, the company said.

NY Times: Career in Statistics (URL)

August 6, 2009

For Today's Graduate, Just One Word: Statistics

By STEVE LOHR

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.

"People think of field archaeology as Indiana Jones, but much of what you really do is data analysis," she said.

Now Ms. Grimes does a different kind of digging. She works at Google, where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

Ms. Grimes is an Internet-age statistician, one of many who are changing the image of the profession as a place for dronish number nerds. They are finding themselves increasingly in demand — and even cool.

"I keep saying that the sexy job in the next 10 years will be statisticians," said Hal Varian, chief economist at Google. "And I'm not kidding."

The rising stature of statisticians, who can earn $125,000 at top companies in their first year after getting a doctorate, is a byproduct of the recent explosion of digital data. In field after field, computing and the Web are creating new realms of data to explore — sensor signals, surveillance tapes, social network chatter, public records and more. And the digital data surge only promises to accelerate, rising fivefold by

NY Times: R (URL)

January 7, 2009

Data Analysts Captivated by R's Power

By ASHLEE VANCE

To some people R is just the 18th letter of the alphabet. To others, it's the rating on racy movies, a measure of an attic's insulation or what pirates in movies say.

R is also the name of a popular programming language used by a growing number of data analysts inside corporations and academia. It is becoming their lingua franca partly because data mining has entered a golden age, whether being used to set ad prices, find new drugs more quickly or fine-tune financial models. Companies as diverse as Google, Pfizer, Merck, Bank of America, the InterContinental Hotels Group and Shell use it.

But R has also quickly found a following because statisticians, engineers and scientists without computer programming skills find it easy to use.

"R is really important to the point that it's hard to overvalue it," said Daryl Pregibon, a research scientist at Google, which uses the software widely. "It allows statisticians to do very intricate and complicated analyses without knowing the blood and guts of computing systems."

It is also free. R is an open-source program, and its popularity reflects a shift in the type of software used inside corporations. Open-source software is free for anyone to use and modify. I.B.M., Hewlett-Packard and Dell make billions of dollars a year selling servers that run the open-source Linux operating system.

Types of Data Mining

Unsupervised Learning

'Unclassified' data from which we would like to uncover hidden 'structure' or groupings

- Given detailed phone usage from many people, find interesting groups of people with similar behaviour.
- Shopping habits for people using loyalty cards: find groups of 'similar' shoppers.
- Given expression measurements of 1000s of genes for 100s of patients, find groups of functionally similar genes.

Goal: Hypothesis generation, visualization.

Types of Data Mining

Supervised Learning

A database of ‘classified’ examples with predefined groupings

- Given detailed phone usage of many users *along with their historic churn*, predict when/if people are going to change contracts again.
- Given expression measurements of 1000s of genes for 100s of patients *along with a binary variable indicating absence or presence of a specific cancer*, predict if the cancer is present for a new patient.
- Given expression measurements of 1000s of genes for 100s of patients *along with survival length*, predict survival time.

Goal: Prediction.

Further Readings

- Leo Breiman: Statistical Modeling: The Two Cultures (URL)
- NY Times: Big Data's Impact In the World (URL)
- Economist: Data, Data Everywhere (URL)
- McKinsey: Big data: The Next Frontier for Competition (URL)

Other recent news on Big Data, Data Mining, Machine Learning:

- New York Times: Sure, Big Data Is Great. But So Is Intuition (URL)
- New York Times: How Many Computers to Identify a Cat? 16,000 (URL)
- New York Times: Scientists See Promise in Deep-Learning Programs (URL)
- New Yorker: Is "Deep Learning" a Revolution in Artificial Intelligence? (URL)