Lecture 1 Introduction

STAT 441/505: Applied Statistical Methods in Data Mining

Linglong Kong

Department of Mathematical and Statistical Sciences
University of Alberta

Winter, 2016

Outline

Data Mining

Statistical Machine Learning

Software and Remark

Data Mining

- Wikipedia: Data mining is an interdisciplinary subfield of computer science.
- It is the computational process of discovering patterns in large data sets (“big data") involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.
- The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amount of data, not the extraction (mining) of data itself.
- It also is a buzzword and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence, machine learning, and business intelligence.

Data Mining

- The book *Data mining: Practical machine learning tools and techniques with Java* (which covers mostly machine learning material) was originally to be named just *Practical machine learning*, and the term data mining was only added for marketing reasons.
- Often the more general terms (large scale) data analysis and analytics - or, when referring to actual methods, artificial intelligence and machine learning - are more appropriate.
- Data mining = data analysis and analytics/artificial intelligence and machine learning + marketing
- Applied Statistical Methods in Data Mining = Applied Statistical Methods + marketing
- Indeed, this course was formerly Applied Statistical Methods.

Machine Learning

- Wikipedia: Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence.
- Machine learning is closely related to computational statistics; a discipline that aims at the design of algorithms for implementing statistical methods on computers.
- Machine learning and pattern recognition *can be viewed as two facets of the same field*.
- Machine learning tasks are typically classified into three broad categories, supervised learning, unsupervised learning, and reinforcement learning.

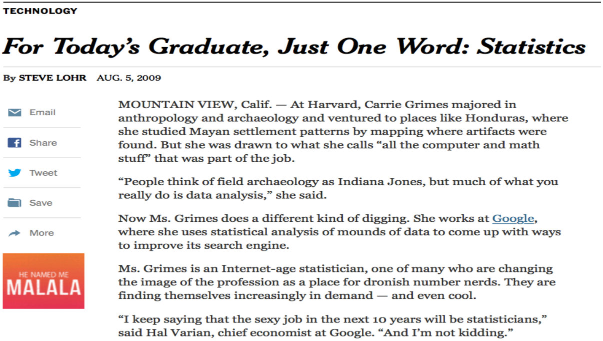
Data Mining and Machine Learning

- Machine learning is sometimes conflated with data mining, although that focuses more on exploratory data analysis.
- Machine learning and data mining often employ the same methods and overlap significantly.
  - Machine learning focuses on prediction, based on known properties learned from the training data.
  - Data mining focuses on the discovery of (previously) unknown properties in the data.
- The two areas overlap in many ways: data mining uses many machine learning methods, but often with a slightly different goal in mind.
- On the other hand, machine learning also employs data mining methods as unsupervised learning or as a preprocessing step to improve learner accuracy.

Statistical Machine Learning

- This courses is not exactly data mining (too vast), nor exactly machine learning, nor exactly multivariate analysis.
    - Data mining is too vast.
    - Machine learning emphasizes on prediction accuracy and on large scale applications.
    - There are quite a few methods beyond the classical multivariate analysis.
- So what do we do in this course? Statistical machine learning!
- Statistical machine learning merges statistics with the computational sciences - computer science, systems science and optimization.
  `http://www.stat.berkeley.edu/~statlearning/`.
- Statistical machine learning emphasizes models and their interpretability, and precision and uncertainty.

Statistical Machine Learning



TECHNOLOGY

For Today's Graduate, Just One Word: Statistics

By STEVE LOHR AUG. 5, 2009

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."

- Quote of the Day, New York Times, August 5, 2009
  "I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding." HAL VARIAN, chief economist at Google.

Supervised Learning

- Data: response Y and covariate X.
- In the regression problem, Y is quantitative (e.g. price and blood pressure).
- In the classification problem, Y takes categorical data (e.g. survived/died, digits $0-9$).
- In regression, techniques include linear regression, model selection, nonlinear regression, ...
- In classification, techniques include logistic regression, linear and quadratic discriminant analysis, support vector machine, ...
- There are many other supervised learning methods, like tree-based methods, Ensembles (Bagging, Boosting, Random forests), and so on.

Unsupervised Learning

- No response, just a set of covariates.
- objective is more fuzzy - find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- Difficult to know how well your are doing.
- Different from supervised learning, but can be useful as a pre-processing step for supervised learning.
- Methods include cluster analysis, principal component analysis, independent component analysis, factor analysis, canonical correlation analysis, ...

Summary and Remark

- Install software **R**, if necessary, play demos, browse documentation.
- In my opinion, the best way to learn in this course is to try everything in **R**.
- Once it works, then think why, and how to write it in your own way.
- READ TEXTBOOK and TRY DATA EXAMPLES.