Lecture 1 Introduction

ECE 625: Data Analysis and Knowledge Discovery

Di Niu

Department of Electrical and Computer Engineering
University of Alberta

January 13, 2021

Outline

Data Mining

- It is the computational process of discovering patterns in large datasets (“big data") at the intersection of artificial intelligence, machine learning, statistics, and database systems.
- IEEE ICDM CFP: “draws researchers and application developers from a wide range of data mining related areas such as statistics, machine learning, pattern recognition, databases and data warehousing, data visualization, knowledge-based systems, and high performance computing.”
- 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

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.

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.
- 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 the textbook and try out examples.