Large Scale Data Analysis (33:136:487)

Instructor: Spiros Papadimitriou

This course introduces students to fundamental statistical techniques for analyzing large-scale business data. The main goal is to provide systematic training in statistical models for massive datasets as well as programming and exploratory data analysis in real-world settings. The course equips students to develop context-sensitive models and perform model checking and diagnosis. Topics include parametric inference, logistic regression, support vector machines, model selection, similarity and clustering. Students are provided the opportunity to learn a comprehensive set of data analysis techniques through lessons, demonstrations, and hands-on programming.

Class goals: (i) approach problems data-analytically: think carefully and systematically about whether and how data can improve performance; (ii) be able to interact competently on the topic of data mining analytics: know the basics of data mining processes, techniques, and concepts; and (iii) receive hands-on experience mining data: you should be able to follow up on ideas or opportunities that present themselves. Throughout the curriculum, special emphasis is placed on techniques that are relevant to large volumes of data. For each topic, the course covers enough theoretical background so students can understand how to apply and interpret each class of models, and then illustrates how it's used via hands-on exploratory analysis using Python. Students will also be expected to complete homework assignments, participate and follow along in-lecture exploratory analysis, building up to a guided class project on a substantial dataset (e.g., analysis and classification of Amazon product reviews).

List of topics

- Introduction
  - What is “data science”
  - Data and models
  - Supervised vs. unsupervised tasks
- Predictive models
  - Entropy and information gain
  - Decision trees
- Introduction to Python for data science
  - Python basics
  - NumPy and Matplotlib
  - Introduction to scikit-learn
- Parametric models and linear classifiers
  - Decision boundaries and line/hyperplane equations
  - Linear discriminant analysis
  - Logistic regression
  - Support vector machines
  - Kernel trick
- Model evaluation
  - Overfitting and model testing

    - Cross-validation
    - ROC curves
- Probability
    - Introduction to probability
    - Naive Bayes
- Text classification
    - Vector-space (bag-of-words) model
    - Feature extraction, stemming, stopwords, and TF-IDF
- MapReduce and Apache Spark (overview)
- Similarity
    - Distance measures and metrics
    - Euclidean and Lp distances
    - Jaccard similarity
    - k-NN classifiers
- Clustering
    - k-Means
    - Hierarchical clustering

Textbook

- “Data Science for Business (What you need to know about data mining and data-analytic thinking)”, Foster Provost and Tom Fawcett, O'Reilly, 2013.

Grading policies

40%	Programming assignments/projects
30%	Midterm
30%	Final