Data Science Research Seminar

Jessica Clark and Kunpeng Zhang

Fall 2018

Learning Goals

By the end of the semester, students should be able to:

1. Understand and evaluate the use of basic machine learning and data science in research papers.
2. Implement data science techniques in Python.
3. Formulate, conduct, and position data science research for publication in business school journals.

Course Description

The design of this class is to work towards a deeper understanding of data science and especially research on data science, based on a combination of readings on fundamental material, such as textbook chapters or classic papers, plus more recent research that uses those fundamentals. We will dig deep into the material in our class discussions, so we may end up spending entire class periods discussing one paper -- meaning discussing the fundamentals underlying what is presented in the paper, plus the particular contributions of the paper, and the general notions of contributions to research in data science, technical information systems, design science, etc., as well as how one actually goes about conducting research (really, not in some ideal world).

We will especially focus on data science papers that have been published in top business school journals, and discuss distinguishing characteristics of such papers. We plan to invite Smith faculty who have used machine learning in their research (across DO&IT, marketing, finance, and possibly others) to guest lecture throughout the semester.

The second component of the class will be practical: learning essential tools that data science researchers use to conduct their research. The two components of the class will be taught concurrently, as data science skills are essential to conducting data science research. We will require that students complete a project for the class using these tools, ideally one which could lead to a research paper.

Course Outline

- Week 1: Introduction to data science
    - Predictive vs. explanatory modeling; formulating a predictive problem
    - Basic machine learning: overview of classifiers (and implementation in Python via scikit-learn)
    - “Bring your own paper” – discussion of how to position data science research for publication in top-tier business journals
    - Evaluation tools: cross-validation, confusion matrices, ROC curves, learning curves
    - Complexity control, bias-variance tradeoff
- Week 2: data collecting, text mining, and natural language processing
    - Writing a data scraper in Python or R
    - Processing web data
    - Web data research questions
    - Ethics in web scraping: privacy, others?
    - Featurizing text data
    - NLP research topics: sentiment analysis, classification, summarization, others
    - Use of NLP in IS research
- Week 3: Bayesian network models (probabilistic graphical models)
    - Directed graphical models – Bayes nets
    - Undirected graphical models – Markov networks
    - Inference – (collapsed) MCMC, variational inference
- Week 4: Unsupervised learning
    - Measuring (dis)similarity and evaluating the output of clustering methods
    - Traditional clustering methods: K-means, DBSCAN
    - Spectral clustering
    - Hierarchical clustering
    - Latent Dirichlet Allocation
    - SVD/PCA
    - Non-negative Matrix factorization
    - Questions: improvements in predictive performance, efficiency gains, interpretability?
- Week 5: Recommender systems
    - Collaborative filtering
    - Content-based RS
    - Hybrid recommender systems
    - The Netflix Prize and matrix-factorization-based recommender systems
    - Ethics in data mining: privacy
    - Ethics in data mining: algorithmic bias
- Week 6: Social Network Analysis
    - Basic network characteristics – centralities
    - Random graphs
    - Community detection – modularity maximization based
    - Information propagation – SIS, SIR models

- Week 7: Deep learning
    - Strategies for overfitting
    - Deep neural nets (DNN)
    - Convolutional neural nets (CNN)
    - Sequence learning (recurrent neural nets, RNN)
- Week 8: Semi-supervised learning (time permitting)
    - Mixture models (mixture Gaussian)
    - EM algorithm
    - Transductive learning
    - Active learning and acquiring labels via crowdsourcing