Syllabus: Math 561 Practical Big Data Analytics

Course Format: Meetings: Monday, Wednesday, Friday 9:00-9:50 a.m., Math 306. Class time is split (50/50) between lecture and individual and small group work on programming and algorithm development.

Learning Outcomes

1. Understand the theory and foundations of data reduction and information extraction (e.g., associative statistics and data mapping).
2. Develop understanding and practical experience regarding reduction of massive data sets and data streams.
3. Understand the theory and mechanics of distributed computing.
4. Ability to formulate and implement algorithms for processing massively large data sets. Ability to compute histograms, correlation matrices, and linear regression estimators using massively large data sets.
5. Understand the objectives of multiple regression, methods for fitting models, and examining model assumptions. Derive estimators from first principles. Ability to carry out and interpret hypothesis tests. Understand the details of fitting qualitative and quantitative predictors.
6. Proficiency using `Python`, `Hadoop/MapReduce` and `R`.
7. Proficiency in data analytic algorithm design. `Hadoop/MapReduce`

Course Content: Algorithms are the machinery behind the data analytics (the subject matter of the course). To be good at data analytics, one must be competent at programming and have experience with the data and the algorithms of data science. To be expert at data analytics requires an understanding of the foundations and principles behind the algorithms. Why? Applying the algorithms to real problems often requires adapting existing algorithms and creating new ones. To get to the point where innovation is not intimidating but instead an opportunity for creativity, students will work with a set of prototypical algorithms that span much of data analytics. Graduate students will derive and develop key components of the algorithms. Algorithms will be implemented in `Python` or `R` and applied to practical problems involving publicly available data sets.

M 561(G) co-convenes with M 461(U). Course content differs however. In particular, M 561 involves theoretical and foundational learning outcomes absent from M 461. Principally, graduate students will derive estimators from theoretical principles and create innovations to core algorithms.

We shall cover most, but not all of the material in chapters 1 through 8 of the text book. The main topics are

1. Data mappings and the concepts of data reduction. Similarity measures and distance metrics. List, set, and dictionary comprehension.
2. Scalable algorithms and associative statistics. Computing univariate and multivariate statistics from massively large data sets.
3. Distributed computing using MapReduce algorithms and the Hadoop environment. Basics of the command line. Utilizing Elastic MapReduce.
4. Data visualization.
5. Linear regression. Hypothesis testing, regression with factors, interaction, and residual analysis.

6. Healthcare analytics.

7. Cluster analysis. Hierarchical and k-means methods. Optimality.

Textbook: Algorithms for Data Science (ISBN-10: 3319457950)

Prerequisites: Stat 341 or M 421 and at least 2 upper division mathematics courses

Homework: Home work exercises emphasizing theory and proofs will be assigned weekly. Proofs are to be complete and concise. Expectations and grading are commensurate with graduate standing (expectations for undergraduates are less). Completion of 4 to 6 tutorials per month. Tutorials are oriented toward gaining proficiency in programming and algorithm design.

Final exam/Project: The final project will be an analysis of a complex data set requiring innovative algorithm design. Students must work in groups of two or three individuals. Graduate students are responsible for a written paper (75% of the project grade) and oral presentation (25% of the project grade). In lieu of a final, students will present their final project during the meeting time of the final. Oral presentations are expository and aimed at communicating methods and results. The written project must address the theoretical and conceptual aspects of the methods and the paper will be graded in accordance with this requirement.

Grading: Your course grade will be based on homework, tutorials, and the final project. Students are responsible for completing 4 to 6 tutorials per month (due a the first meeting of each month except September). Homework assignments and tutorials are worth 50% and 25% of the course grade, respectively. Home work exercises emphasizing theory and proofs will be assigned weekly. Proofs MUST be complete and concise. Expectations and grading are commensurate with graduate standing (undergraduates generally are not asked to prove theorems).

The final project is worth 25% of the course grade. 30% of the project grade evaluates mathematical writing. You are expected to provide clear and concise explanations and justifications of the methods used in the analysis (undergraduates are not expected to provide this level of mathematical maturity). The final meeting time (for presentation of projects) is Friday, December 14, 10:10 a.m.