Syllabus: Math 461 Practical Big Data Analytics

Course Format: Two meetings per week: Tuesday, Thursday 9:30-10:50, Math 305. Class time is split (50/50) between lecture and individual and small group work on programming and algorithm development.

Learning Outcomes

1. Understand the purpose 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 mechanics of distributed computing.
4. Ability to 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 and examining model assumptions. Ability to carry out and interpret hypothesis tests.
6. Competency using `Python` and `R`.

Course Content: Algorithms are the machinery behind the data analytics (the subject matter and focal point of the course). To be good at data analytics, one must be a competent programmer and have experience with the data and the algorithms of data science. To gain an understanding of algorithm design and good programming practices, students will work with a set of prototypical algorithms that are representative of data analytics. To learn how to function as a data scientist in a relatively short time, the student will be actively engaged in turning the algorithms into code and using them with real data.

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, and one of M 221 or M 273, or consent of instructor.

Home work: Home work exercises emphasizing applications of the algorithms will be assigned weekly. Completion of 4 tutorials per month. Tutorials are oriented toward gaining proficiency in programming and

algorithm design.

Grading: Your course grade will be based on homework, tutorials, and a final project. Students are responsible for completing 4 tutorials per month (due a the first meeting of each month except September). Homework assignments and tutorials are worth 40% of the course grade, respectively. The final project is worth 20%. The final project will be an analysis of a complex data set. Undergraduate students are responsible for a written paper or (with instructor approval) an oral presentation.

Final exam/Project: In lieu of a final, students will be graded on their final project. The written project must describe the data, objectives, and results. Students must work on projects in groups of two or three individuals.