Program Cover Sheet --- STA307/CSC307: Data Mining & Predictive Modeling

I. Basic Course Information

Data Mining is a 300-level course that satisfies optional requirements in statistics and computer science. The course meets twice weekly. Students must have completed STA215 and a programming course (CSC220, CSC230, CSC250, or CRI215) prior to enrolling. The course is suitable to students from many disciplines as it provides an introduction to the mining of large data sets, which have become ubiquitous in recent years.

Predictive modeling in statistics has a very long history, dating back to the time of Gauss and Legendre, when they used least squares/linear regression to make predictions on specific outcomes. Modern data mining started to take shape a few decades ago when new advances were made in CART (Classification and Regression Trees), neural networks, and text mining. In the past two decades, dramatic changes occurred when powerful new algorithms relying on modern computing addressed problems in signal analysis, image analysis, drug discovery, manufacturing process control, searches in large databases, and network analysis.

II. Learning Goals

Data mining is a field of study on mathematical models and algorithms originated from different disciplines, primarily statistics and machine learning. Techniques of data mining have been used successfully in science, engineering, biomedical research, business, political campaigns, database marketing, and genomic mining.

The primary goal of this course is to introduce students to a variety of statistical and computer science techniques that are widely used in modern data mining. The techniques in classification may include linear discriminant analysis, logistic regression, neural networks, decision trees, random trees and random forests, and support vector machines. The techniques in clustering will briefly touch upon hierarchical clustering (covered in detail in STA306), before introducing K-means clustering, self-organizing maps, spectral clustering, and ensemble methods. Additional techniques may include two-stage modeling, text mining, various missing value imputation techniques, association rules (market basket analysis), and independent component analysis. Advanced techniques such as genetic algorithms, kernel techniques, local likelihood, MARS (Multivariate Adaptive Regression Splines), relative risk trees, LARS (Least Angle Regression), and basis expansion may be discussed as well, so that students can use such methods in their final projects. The field is still evolving rapidly, so topics are likely to change with each offering of the course to keep abreast of the most promising recent developments.

In this course, students will study many large-scale applications. The course will also guide students to read through the literature for general background, historical perspective, potential misuse and abuse of data mining techniques, and the ethical conduct of data mining.

The course will be a mix of theory and applications, and will include extensive use of modern computing power. The College is equipped with state-of-art technology for students to build sophisticated models in the investigation of massive data sets. For main techniques covered in class, students will have the opportunity to actually analyze massive data sets with millions of cases and tens or hundreds of variables that were beyond the reach of Gauss and Legendre. The course may use the R statistical language or the SAS package, both of which are available to students on campus. R is available free on Linux, Windows, and Macintosh platforms.

III. Student Assessment

Students will be assessed through a combination of weekly problem sets, exams during the semester, and a final project or final exam either alone or in combination.

IV. Learning Activities

Learning activities will consist of a combination of lectures, discussions, student presentations, and programming and analysis assignments. The specific choice will depend upon the individual instructor. Outside of class, students are expected to do a significant amount of individual and group work to achieve the learning goals. Students are encouraged to work together to more fully develop a deep understanding of data mining and predictive modeling. The ability to do many tasks on laptop computers will permit students to work as a group in the library in addition to in the School of Science computer laboratories.

Updated 11/4/15