Program Cover Sheet --- STA 307: Data Mining & Predictive Modeling

I. Basic Course Information

Data Mining is a 300-level course that satisfies optional requirements in the Data Science & Statistics and Applied Mathematics major. The course meets twice weekly. Students must have completed an introduction to statistics course and a programming course (CSC 120 or MAT 203) 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.
Prerequisites: (STA 215 or STA 216 or (ECO 105 and (MAT 125 or MAT 127)) or (STA145 and MAT 127)) and a programming course (MAT 203 or CSC 120).

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 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

This course introduces some of the statistical methods that underlie data mining, with a focus on classification and factorization. These methods provide the basis of supervised and unsupervised learning approaches that form the backbone of modern analysis of large data sets. Techniques of data mining have been used successfully in science, engineering, biomedical research, business, political campaigns, and marketing.

The primary goal of this course is to introduce students to a variety of statistical and data science techniques that are widely used in modern data mining, including clustering and classification algorithms. 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 may include hierarchical clustering, K-means clustering, self-organizing maps, spectral clustering, and ensemble methods. As the field of data mining is rapidly evolving, the instructor has discretion to choose topics from the above list, or to introduce new topics to keep abreast of the most promising recent developments.

In this course, students will study many large-scale applications. The course will be a mix of theory and applications, and will include extensive use of modern computing power. For the main techniques covered in class, students will have the opportunity to actually analyze massive data sets that were beyond the reach of Gauss and Legendre. The course will use a state-of-the art statistical language or package.

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.

Approved: 12-8-2025