Program Cover Document --- STA-494 Seminar in Statistics: Data Mining & Predictive Modeling

I. Basic Course Information

Data Mining is primarily a junior/senior level course. It is an optional course for MATC, MATA, MATT, Mathematics and Statistics minors. The course is scheduled for two 80-minute meetings each week. Its prerequisite is (1) STA-305 or ECON-231 or STA-220 or PSYC-303, or (2) MAT-316 and permission from the instructor. Students from Computer Science, Business, Biology, Engineering, Psychology, Political Science, and other branches of Science and Social Studies are welcome.

II. Learning Goals

Data mining is a field of study on mathematical models and algorithms originated from different disciplines including statistics, machine learning, neural networks, fuzzy logic, and evolutionary computation. Techniques of data mining have been used successfully in science, engineering, biomedical research, business, political campaign, data base marketing, and genomic mining. The US federal government also has very extensive efforts on data mining, including government services and homeland security (see, for example, documents prepared by the US General Accounting Office, May 2004, http://www.gao.gov/new.items/d04548.pdf and by US Department of Defense, March 2004, http://www.epic.org/privacy/profiling/tia/tapac_report.pdf).

The primary goal of this course is to introduce students to a variety of statistical techniques that are widely used in modern data mining. The techniques include decision trees, link functions, logistic regression, neural networks, tree/net, prior vector and profit/loss matrix, two-stage modeling, text mining, various missing value imputation techniques, association rules (market basket analysis), self-organizing maps, and independent component analysis. The course may also discuss DNA Microarray data analysis, a sub-area of data mining, including the use of statistical techniques to improve the quality of DNA Microarray data. Advanced techniques such as random forests, kernel techniques, local likelihood, MARS (Multivariate Adaptive Regression Splines), relative risk tree, LARS (Least Angle Regression), basis expansion, and support vector machines will be discussed to motivate top students toward projects that are beyond regular course work.

Data Mining was described by an article in Amstat News (an official publication of American Statistical Association, 9/2003) as a <<defining event>> that will impact the future of statistics. MIT Technology Review (Jan/Feb/2001) and Bayesian Machine Learning (Feb, 2004) ranked data mining as one of the ten emerging technologies that will change the world. It was observed that in the 1999 Joint Statistical Meeting there was only one paper on DNA Microarray, but there were over a hundred in 2002 and some 1,200 papers in 2003 (Professor J. Cabrera, Rutgers University, 2004). This is exponential growth with astonishing rate. In an Internet search, out of 3598 statistics jobs, 36% were in the area of data mining (Professor R. Hannum, Denver University, 2004). Other success stories can be found at sas.com, spss.com, IBM Intelligent Miner website and google.com.

Predictive modeling in statistics has a very long history, dating back to the time of Gauss (1823) and Legendre (1821) when they used least square/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 10-15 years, dramatic changes occurred when researchers used neural networks, CART, and other techniques to investigate a variety of problems including signal detection, satellite image analysis, drug discovery, models of gene regulation, medical diagnosis, disease stratification, automotive warranty early warning system, manufacturing process control, and as a delight to many people, data mining tools are also used on drowsy driver detection and auto/truck safety (after all, modern automobiles have more computational capability than the Apollo spacecraft that went to the Moon).

In the past 8 or 9 years, the techniques are also used successfully in business applications such as web mining, targeted marketing, financial crimes detection, loan failure, credit card holder balances, insurance claim losses, customer catalog orders, cell phone usage, customer relationship management, financial services, plus other examples in planning and services from government and big organizations.

In this course, students will study many large-scale applications. The course will also guide students to read through general background, historical development, potential misuse and abuse, and ethical conducts of data mining. In addition, the course will ask students to identify what data mining can do and cannot do in the human quest for prediction, special events detection and hidden causes identification.

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.

III. Student Assessment

This course is intended to be highly homework intensive. Weekly problem sets will constantly provide students opportunities to show their understanding of the material. At the same time students will receive weekly feedback on their work and their progress. A combination of computing assignments, quizzes and tests throughout the course will provide further valuable information both for the instructor and the individual students. Students will be encouraged to present their course projects, individually or with their group members, in the class or at the annual TCNJ Celebration of Student Achievements.

IV. Learning Activities

The specific choices of learning activities will depend upon the instructor, but it is expected that they will consist of some combination of lectures, pre-class reading assignments, group work, student presentations, individual homework, computer assignments, quizzes, tests and final exam. This course will cover a lot of topics in data mining, and hence students are expected to complete pre-class reading assignments and to pass related quizzes in order to understand and appreciate the course materials. We also plan to invite guest speakers to talk on new development and applications of data mining (no charge to the College if the speaker is from SAS Inc.). Students will be asked to summarize the talk as part of their homework assignments.

Departmental Course Syllabus --- STA-494 Seminar in Statistics: Data Mining & Predictive Modeling

I. Basic information on course and instructor

A. Purpose statement: The primary goal of this course is to introduce students to a collection of data mining techniques that have been used successfully in science, engineering, biomedical research, business, political campaign, data base marketing, and genomic mining. The course will also guide students to discuss potential misuse and abuse, and ethical conducts of data mining. In addition, the course will ask students to identify what data mining can do and cannot do in the human quest for prediction.

B. Course description: An introduction to Data Mining and Predictive Modeling. Topics include decision trees, logistic regression, neural networks, tree/net, two-stage modeling, text mining, association rules (market basket analysis), self-organizing maps, and independent component analysis.

B. Course prerequisite: (1) STA-305 or ECON-231 or STA-220 or PSYC-303, or (2) MAT-316 and permission from the instructor.

II. Learning goals

A. Content goals: The course will discuss standard data mining techniques such as decision trees, link functions, logistic regression, multilayer perceptron, tree/net, prior vector and profit/loss matrix, two-stage modeling, various missing value imputation techniques, singular value decomposition, clustering analysis and Kohonen topology-preserving maps, and principal component and its use to aid neural network prediction.

B. Performance goals: At the completion of the course, students should demonstrate competence with the theory in the forms of derivations, calculations, and mathematical proofs. Students are also expected to be skillful with computer usage in the mining of massive data sets. Furthermore, students are expected to build competing models for specific data set and be able to use a variety of statistical criteria to select a model that best suit the purpose of the study. Students are expected to know the non-linear characteristic of both the models and the modeling process.

III. Student assessment

A. Assessment Plan: This course is intended to be highly homework intensive. Weekly problem sets will constantly provide students opportunities to show their understanding of the material. At the same time students will receive weekly feedback on their work and their progress. A combination of computing assignments, quizzes and tests throughout the course will provide further valuable information both for the instructor and the individual students. Students will be encouraged to present their course projects, individually or with their group members, in the class or at the annual TCNJ Celebration of Student Achievements.

B. Rationale: Through the use of regular feedback from homework, quizzes, student presentations and examinations, students will be able to see and correct their misunderstandings and improve their performance.

C. Methods and criteria: We will use the assessment of homework, quizzes, student presentations, and examinations to evaluate student accomplishment of the course learning goals. These assessment tools are similar to the manner in which students will need to use their knowledge in the future and are an appropriate way to assess the accomplishment of course learning goals.

IV. Learning activities

A. Summary of learning activities: The specific choices of learning activities will depend upon the instructor, but it is expected that they will consist of some combination of lectures, pre-class reading assignments, group work, student presentations, individual homework, computer assignments, quizzes, tests and final exam. This course will cover a lot of topics in data mining, and hence students are expected to complete pre-class reading assignments and to pass related quizzes in order to understand and appreciate the course materials.

B. Calendar or outline: A guide to the organization of the course, a schedule of assessment tools, and a plan for the coverage of topics should be provided to the students. Homework, quizzes, and examinations should be spaced at appropriate intervals throughout the semester. A sample course outline is at the end of this document.

C. Rationale: By giving students a multitude of ways to learn and do mathematics, the learning activities promote a deeper understanding of the course materials and contribute to the learning goals of these programs. A regular spacing of assessment tools insures students' continual regular feedback on their work.

Sample Course Outline

Textbook:

(a) *Principles of Data Mining*, by David J. Hand, Heikki Mannila and Padhraic Smyth, the MIT press, 2001. Or

(b) *Data Mining Techniques,* Michael Berry & Gordon Linoff, 2004, Wiley.

1. Multiple Regression (a brief review)
2. Various sampling plans prior to the analysis of a massive data set
3. Data partition: training data, validation data, and test data
4. CART (Classification and Regression Tree): χ^2-test, Gini coefficient, entropy, the growth/pruning and the interpretation of the tree
5. The prior vector and profit/loss matrix
6. Link function and stepwise logistic regression
7. Gains chart, banana chart, lift chart, and assessment and comparison of models
8. Deploying a predictive model, scoring a new data set, and reporting results
9. Neural networks-I: connections of input, hidden, and output units, the biological inspiration, error functions, combination functions, activation functions, and MLP (multilayer perceptron)

10. Neural networks-II: altitude, width, and the Gaussian radial basis function networks (optional)
11. Neural networks-III (Tree/Net): using CART decision tree to improve the Neural network performance
12. Neural network-IV: principal component analysis and its use to aid neural network prediction
13. Neural networks-V: neighborhood functions, Kohonen’s topology-preserving maps, and various clustering techniques (optional)
14. Missing-value imputation
15. Two-stage modeling for categorical and interval targets
16. Association rules (market basket analysis)
17. Text mining, an introduction: data cleansing, sparse matrix, singular value decomposition and its use to improve predictive modeling
18. Advanced topics in data mining (optional)
19. Potential misuse and abuse of data mining
21. What data mining can do and cannot do? And a comparison of data mining and statistical design of experiments