Course #: STAT 5370  
(CRN 15577)  
Course Title: Data Mining and Statistical Learning  
Credit Hrs: 3  
Term: Fall 2014  
(Instruction 08/25/2014-12/04/2014)  
Course Meetings & Location: Classroom Building C203  
10:30-11:50am TR  
Prerequisite Courses:  
Linear models and generalized linear models; Some programming experiences would be plus, though not required.  
Instructor: Xiaogang Su  
Office Location: Bell Hall 320  
Contact Info:  
Phone: (915) 747-6860 [O]  
xsu@utep.edu  
Fax: (915) 744-6502  
Office Hours: 1:30-2:30pm TR  
Class Web page: https://sites.google.com/site/xgsu00/stat-5370  
Textbook(s), Materials:  
Course Description and Learning Outcomes:  
With the advent of computers and database management tools, vast amounts of data are being generated in various fields. As Rutherford D. Roger states, “We are drowning in information and starving for knowledge.” Data Mining is the process of exploring and analyzing, by automatic or semiautomatic means, large quantities of observational data in order to discover meaningful patterns and models. By applying data mining techniques, data miners can fully exploit data patterns and behavior, gain insider understanding of the data, and produce new knowledge that decision-makers can act upon.  
Data Mining emerges as an interdisciplinary field with joint inputs from statistics, computer science, machine learning, and artificial intelligence. This course is intended to cover some commonly-used data mining techniques, with more focus on the most technical part - statistical learning algorithms. The materials are arranged in two main categories: unsupervised learning and supervised learning. A tentative outline of the specific topics is provided below.
1. Introduction to Data Mining;
2. R Preliminaries: Downloading and Installing; Introduction to R; Dealing with Large Data Sets in R
3. Data Preparation: Data Cleaning; Sampling and Partitioning Data; Missing Value Imputation; Exploration.

Part I: Unsupervised Learning
5. Principal Components: PCA and its extensions such as Principal Curves and Surfaces; Factor Analysis (Exploratory and Confirmatory);
7. Web Mining: Google PageRank

Part II: Supervised Learning
8. Linear Regression: Least Squares Estimation; Assessment and Validation; Model Diagnostics; Generalized LS; Partial/Total LS
9. Variable Screening and Selection: Variable Relevance/Importance - the RELIEF algorithm; Univariate/Bivariate Screening; Stepwise Variable Selection and Regularization - ridge regression, LASSO and its variants, etc.;
10. Logistic Regression: Model Assessment, Validation, and Regularization; Classification Errors, Odds Ratio, ROC, etc.; Generalized Linear Models (GLM);
11. Regression/Classification Trees: Pros and Cons; Impurity Measures; Pruning; Tree Size Selection; and Interpretation; Generalization to Regression, Censored Survival Data, and Longitudinal Data;
12. Ensemble Models: Boosting; Bagging and Random Forests; Stacking.

Course Activities/Assignments: Presentations and projects will be assigned throughout the semester. NO LATE COURSEWORK WILL BE ACCEPTED, EXCEPT EXTREME SCENARIOS.

Assessment of Course Objectives: Each student will be evaluated by the quality of his/her own assigned presentations as well as their contribution to the discussions during other students’ presentations.

Course Schedule:

<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>08/25</td>
<td>Class starts</td>
</tr>
<tr>
<td>10/31</td>
<td>Class drop deadline</td>
</tr>
<tr>
<td>12/08 - 12/12</td>
<td>Final Exam Period</td>
</tr>
<tr>
<td>Holidays</td>
<td></td>
</tr>
<tr>
<td>09/01</td>
<td>Labor Day</td>
</tr>
<tr>
<td>11/27-11/28</td>
<td>Thanksgiving</td>
</tr>
</tbody>
</table>
Grading Policy: There will be a number of computer assignments and a final project. The assignments make up 60% and the final project makes up 30% to your final score. For the final project, students are given the freedom to select data from whatever field they are interested in. Students should make their own plans to collect data, raise interesting research questions, and consult the instructor for the adequacy of the project. Also, each student will have the opportunity to present their work in class. There will also be a few in-class quizzes or exams, which make up 15%. No make-up exam will be given and no late project submission is accepted without justifiable reasons.

Letter grades are determined according to the following scale:

Grades Score
A  90-100
B  80-89
C  70-79
D  60-69
F  <60

Make-up Policy: All other assignments must be turned in on time.

Attendance Policy: Class attendance is required and helpful to decide borderline grades. If a student has to be absent from a particular seminar, he/she will be responsible for catching up with course material.

Academic Integrity Policy: Please see http://academics.utep.edu/Default.aspx?tabid=23785

Civility Statement: This is a class where participation is required. You will be participating in classroom discussions. All students will be treated with respect.

Disability Statement: If you have a disability and need classroom accommodations, please contact The Center for Accommodations and Support Services (CASS) at 747-5148, or by email to cass@utep.edu, or visit their office located in UTEP Union East, Room 106. For additional information, please visit the CASS website at www.sa.utep.edu/cass.

Military Statement: If you are a military student with the potential of being called to military service and/or training during the course of the semester, you are encouraged to contact me as soon as possible.