STAT 897: Applied Data Mining and Statistical Learning

Syllabus

- **Instructor**: Lingzhou Xue (lxx6@psu.edu)
- **Teaching Assistant**: Joshua Snoke (jvs140@psu.edu)

Course Information and Requirement

Overview

Data mining and statistical learning use a variety of computational tools for understanding large and complex datasets. In some cases, the focus is on building models to predict a quantitative or qualitative output based on a collection of inputs. In others, the goal is simply to find relationships and structure from data with no specific output variable.

This course takes an applied approach to understand methodology, motivation, assumptions, strengths, and weaknesses of the most widely applicable methods in this field.

Prerequisites

Students enrolling for this course should have taken **STAT 501** which deals with simple and multiple regression, correlation, polynomial models, step-wise and piece-wise regression and rudimentary logistic regression. Students are also expected to know basics of probability and conditional distributions and expectations. A knowledge of linear algebra and multivariate calculus is beneficial in understanding some of the concepts underlying the methods.

Required Textbook

An Introduction to Statistical Learning with Applications in R (2013) by G. James, D. Witten, T. Hastie, and R. Tibshirani

The textbook is available at http://bookstore.mbsdirect.net/psude.htm. This is the URL for the book supplier for Penn State World Campus. Penn State Library has an electronic version freely available. You may search for An Introduction to Statistical Learning and download the files at http://link.springer.com.ezaccess.libraries.psu.edu.

Optional Textbooks

- *All of Statistics: A Concise Course in Statistical Inference* by L. Wasserman
- *The Elements of Statistical Learning* by T. Hastie, R. Tibshirani, and J. Friedman
- *Pattern Recognition and Machine Learning* by C. M. Bishop
- *Pattern Recognition and Neural Networks* by B. Ripley

Statistical Softwares

R: http://cran.r-project.org/ to download R for free. We strongly recommend downloading and working in R-Studio. You can use other programs such as Minitab or SAS, but the class notes and assignments will focus on R. The weekly Data Analysis Assignments will be submitted in R Markdown (http://rmarkdown.rstudio.com/).

If you are using R for the first time, please take some time to follow through the following introduction: https://onlinecourses.science.psu.edu/statprogram/node/50. There are a number of tutorials available on the YouTube, e.g., https://www.youtube.com/playlist?list=PLqzoL9-eJTNBDdKgJgJzaQcY6OXmsXAHU. Another helpful resource is R Programming for Data Science by Roger Peng. This book is freely available on the web.

Python is also widely used in academia and industry. We may also use a bit of Python in places where no suitable R implementation is available. If students would like to use Python in assignments, that will probably be fine and even encouraged, but please get in touch with me about it. Despite the name, R Markdown can also work with Python code blocks.

Grading Policy and Requirements

Your reading assignment, topics, assignments, and due dates will be posted on Canvas. Note that there is no exam in this course.

Weekly R Lab Assignments (15%) There will be weekly R labs to be graded on Canvas. Any questions marked incorrectly will be reviewed and solution will be released after the deadline. Late labs will not be accepted without a prior notice and a legitimate excuse.

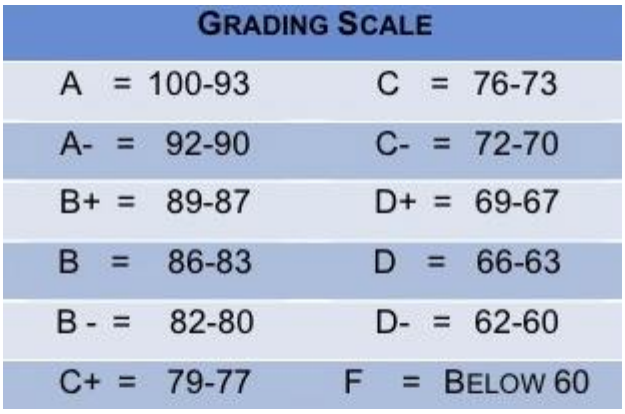
Each lab consists of 2 parts. The first part is to follow through the text's lab on your own time. Do **NOT** submit your work on the text's lab. After you feel comfortable with code and material discussed in the text's lab, please complete the R Lab. The second part consists of about 8 questions that test your knowledge of the commands and output covered in the text's lab. You have sufficient time to complete this part on Canvas and your answers will be submitted when time is up. Please allow sufficient time to take the labs without interruption.

Weekly Quiz (10%) There are weekly quizzes based on the course material. The quiz is timed but time allotted is more than sufficient. The quiz will be submitted when the time limit is up. Please allow sufficient time to take this quiz without interruption. Quizzes will be reviewed and explanations will be released after the deadline. Late quizzes will not be accepted without a legitimate excuse.

Data Analysis Assignments (25%) There are weekly data analysis assignments. For each of them, you will apply the material and R codes to real data. The objective here is for you to apply the concepts to data and write your own codes. There is no time limit but you must submit it before due dates. Late submissions will not be accepted without a legitimate excuse.

Projects (50%) There will be 2 Individual Projects worth **30%** and 1 Team Project worth **20%**. The instructions for each project will be posted on Canvas. The individual projects include sample R Code for a similar analysis that you may use as a template for the project. For example, follow through Project1ex.R posted in the Project 1 folder before you attempt the project and then adapt that code as necessary to analyze a different dataset. However, you are welcome to try different techniques. In that case please also submit your codes. The Team Project at the end of the course is more open ended and only guidelines will be provided.

Participation (Extra Credit) Students are encouraged to communicate and participate in the discussion board on Canvas. Please post class questions and answer others’ questions if you can, rather than emailing the instructor or TA. In this way, everyone can benefit from it. Top three active participants will receive the extra credit **2%, 1.5% and 1%** respectively.



GRADING SCALE	
A = 100-93	C = 76-73
A- = 92-90	C- = 72-70
B+ = 89-87	D+ = 69-67
B = 86-83	D = 66-63
B - = 82-80	D- = 62-60
C+ = 79-77	F = BELOW 60

A course percentage above 90% will earn A or A-, and above 80% will earn B+, B- or B.

Note: All labs/quizzes/dropboxes will be closed at 11:55 PM EST on the due date. Formal instruction will end on the last day of class. However, you will continue to be able to access the course materials for one year from the day the course began.

Academic Integrity:

It is encouraged for you to work with your colleagues regarding homework assignments, but the solutions you submit MUST be your own. Furthermore, it is to be understood that no collaboration is to occur regarding the examinations. For any material or ideas obtained from other sources, such as the text or things you see on the web, in the library, etc., a source reference must be given. Direct quotes from any source must be identified as such. This course will abide by the Penn State Academic Integrity Policy.

Accommodations for Students with Disabilities:

Penn State welcomes students with disabilities into the University's educational programs. If you have a disability-related need for reasonable academic adjustments in this course, contact the Office for Disability Services (ODS) at 814-863-1807 (V/TTY). For further information regarding ODS, please visit the Office for Disability Services website. In order to receive consideration for course accommodations, you must contact ODS and provide documentation (see the documentation guidelines). If the documentation supports the need for academic adjustments, ODS will provide a letter identifying appropriate academic adjustments. Please share this letter and discuss the adjustments with your instructor as early in the course as possible. You must contact ODS and request academic adjustment letters at the beginning of each semester.

Code of Mutual Respect and Cooperation:

The Eberly College of Science Code of Mutual Respect and Cooperation embodies the values that we hope our faculty, staff, and students possess and will endorse to make The Eberly College of Science a place where every individual feels respected and valued, as well as challenged and rewarded.

Campus Emergency:

In case of weather-related delays at the University, this online course will proceed as planned. Your instructor will inform you if there are any extenuating circumstances regarding content or activity due dates in the course due to weather delays. If you are affected by a weather-related emergency, then please contact your instructor at the earliest possible time to make special arrangements.