SYLLABUS: 897D APPLIED DATA MINING AND STATISTICAL LEARNING

- Instructor: Jason Morton (morton@math.psu.edu), Associate Professor of Mathematics and Statistics.

Overview. Data mining and statistical learning methods use a variety of computational tools for understanding large, 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 the 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 their expectations. A knowledge of matrix algebra and multivariate calculus is beneficial in understanding some of the concepts underlying the methods.

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

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

The course notes are publicly available online and are linked within each Canvas lesson.
Further Reading. Some highly recommended texts are:

- *Applied Predictive Modeling (2013)* by Max Kuhn and Kjell Johnson. This book has considerable overlap with the course textbook ISLR, and contains a number of case studies, datasets and R codes. You will sometimes be given a set of optional reading from this book (APM). A pdf version is freely downloadable.


- *The Elements of Statistical Learning, 2nd ed.*, by T. Hastie, R. Tibshirani, & J. Friedman. Our required course text is a simplified version of this book. You may want to refer to it if you want to go more into depth on any topic we cover from the course text, particularly from a theoretical perspective.

Some other recommended texts are:

- Statistical Learning from a Regression Perspective by R. A. Berk
- Modern Multivariate Statistical Techniques by A. J. Izenman
- Pattern Recognition and Machine Learning by C. M. Bishop
- Classification & Regression Trees by L. Breiman, J.H. Friedman, R.A. Olshen, & C. J. Stone
- Pattern Recognition and Neural Networks by B. Ripley
- Principles of Data Mining by H. Mannila, P. Smyth & D. J. Hand

Statistical Software. In this class we will primarily use the open source statistical software R. Go to [http://cran.r-project.org/](http://cran.r-project.org/) to download R for free. Weekly Data Analysis Assignments will be submitted in R Markdown ([http://rmarkdown.rstudio.com/](http://rmarkdown.rstudio.com/)).

We strongly recommend downloading R-Studio from [https://www.rstudio.com/](https://www.rstudio.com/) and working in that environment (which makes R Markdown easy to use). If you are unfamiliar with R (or need to brush up), please take some time to follow through the following introduction: [https://onlinecourses.science.psu.edu/statprogram/node/50](https://onlinecourses.science.psu.edu/statprogram/node/50)
If you are using R the first time, there are a number of very good tutorials available on the YouTube. You may check out the following for a comprehensive training. [https://www.youtube.com/playlist?list=PLqzoL9-eJTNBDdKgJgJzaQcY60XmsXAHU](https://www.youtube.com/playlist?list=PLqzoL9-eJTNBDdKgJgJzaQcY60XmsXAHU)

Another good R resource for Data Scientists is R Programming for Data Science by Roger Peng. This book is also freely downloadable and refers to videos as and when necessary.

**Python** is also widely used in academia and industry for the kind of problems this course covers. We may 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.** There is no exam in this course.

Your reading assignment, the lesson topics, Canvas assignments, and due dates are posted on the course schedule on Canvas.

*Weekly R Lab Assignments (15%).* There will be weekly R labs automatically graded in Canvas. Any questions marked incorrectly by Canvas will be reviewed and solution released within a stipulated time after the deadline. Late labs will not be accepted without a proper University approved 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 the code and material discussed in the text’s lab, please complete the Canvas R Lab. This second part consists of approximately 8 quiz questions that test your knowledge of the commands and output covered in the text’s lab. You have sufficient time to complete this part of the lab in Canvas and your answers will automatically submit when the time limit is up. Please allow sufficient time to take the labs without interruption.

*Weekly Quiz (10%).* There are weekly quizzes on the course material. The quiz is timed but time allotted is more than sufficient. The quiz will automatically submit when the time limit is up. Please allow sufficient time to take this quiz without interruption. Quizzes will
be reviewed and solutions released within a stipulated time after the deadline. Late quizzes will not be accepted without a proper University approved excuse.

*Data Analysis Assignments (25%).* There are weekly data analysis assignments for this course. For each of them you will need to apply the weekly material and R code to data (some of which is available within R). The objective of these assignments is for you to apply the concepts to data and write your own code, following the R Labs for the same week. There is no time limit but you must submit within the due dates posted in the Schedule. Late submissions will not be accepted without a proper University approved excuse.

*Projects (50%).* There will be 2 Individual Projects worth 25% together, and 1 Team Project worth 25%. The instructions for each project will posted to 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 the different dataset in the project. However, you are welcome to try different techniques. In that case please submit your codes for the benefit of the class. The Team Project at the end of the course is more open ended and only guidelines will be provided.

   All labs/quizzes/dropboxes close at 11:55 PM EST/EDT on the due date. Late submissions will not be accepted without a proper University approved excuse.

   A course percentage above 90% will earn A- or A and a course percentage above 80% will earn at least a B-, B, or B+. Ideally there should not be any grade below B.

   Note: 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.

Please note that in the past, I have caught students cheating in online courses and pursued full disciplinary action.

**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.