STAT 6306 – Introduction to Data Science – Spring 2017

Instructor

	Office	Email	Office Hours
Darren Homrighausen	Heroy 134	dhomrighausen@smu.edu[1]	MW 5:00 - 6:30

Lecture	MW 3:30 PM – 4:45 PM Heroy 0127
Required Text	*An Introduction to Statistical Learning: with Applications in R* James, Witten, Hastie & Tibshirani. ISBN-13: 978-1461471370
Web Site	Canvas
Prerequisites	Multiple regression, Probability, basic linear algebra

Data science is a suite of tools for visualizing, exploring and analyzing data. Applications of it are happening all around you - and if they are done well, they may sometimes even go unnoticed. How does Google web search work? How does Netflix recommend movies to each of its users? How could we predict whether or not a person will develop breast cancer based on genetic information? Can we automatically interpret incoming email to label it as spam? An expert's answer to any one of these questions may very well contain enough material to fill its own course, but basic answers stem from common principles.

This course involves a good deal of both applied work (programming, problem solving, data analysis) and theoretical work (learning, understanding, and evaluating methodologies). We will try to make the class as applied as possible. However, there are necessary detours into more technical aspects.

Upon completing this course, you should be able to tackle data analysis problems by: (1) selecting the appropriate methods and justifying your choices; (2) implementing these methods programmatically (using R, for example) (3) explaining your results to a person outside of statistics.

[1] Please put "STAT 6306" at the beginning of the subject line. I may miss your email if it doesn't include this meta data.

Administrative Remarks

Lectures and Office Hours

Attendance. Attendance is mandatory. You can do it.

Office Hours. I will look into scheduling an office hour once the course begins.

Software

R. In this class you will be provided the opportunity to work with the `R` statistical package. It is a free and widely used statistical computing platform. It is moderately well documented and has a vast user community (download it from `www.r-project.org`). This is the primary software for this class. Though `R` does have memory and speed constraints when confronted with large problems, we will see that it is actually still quite effective.

A convenient interface for exploring `R` is `Rstudio`, which provides both an interpreter and an IDE.

Python. We may additionally use `Python`. It is a flexible platform that is increasingly being used for data analysis tasks.

A good, comprehensive install for Python is via `Anaconda` (from Continuum), which additionally installs some major packages (`Numpy, Scipy, Matplotlib, Scikit-learn, ...`)provides an interface for `Jupyter` and an IDE called `Spyder`. Not that there has been a rare schism in which Python has made a new version 3.x which is no longer backwards compatible with version 2.x. For this class, use Python 2.7.

Some Python resources:

- MIT OCW [click to load website]
- Learn Python the Hard Way, 3rd Edition [click to load website]

Homework and Tests

Homework. There will be seven homework assignments, due every couple of weeks. Submit the homeworks online via Canvas. Homework **must** be submitted as a '.pdf' file. If you use word processing software, export the document to pdf instead of the software's native, proprietary file format.

No homeworks emailed to me will be accepted unless prior arrangements have been made. Homework assignments that are turned in late will be assessed a 30% penalty, regardless of the reason they are late! No homeworks will be accepted after 24 hours following the original submission deadline.

Feel free to discuss homework assignments with others, but realize that the work you hand in must be your own. Homeworks will be equally weighted and comprise 35% of the course grade.

Tests. There will be 3 tests during the semester (**Updated:** Feb. 20, Mar. 20, April 12). Each test is worth 15% of the course grade.

Final Policy. The final will be worth the remaining 20% of the course grade. There will be more details as the course progresses as to the schedule and content.

Miscellaneous

Disability Resources. If you require a special accommodation, such as needing more time to finish exams, contact me as soon as possible.

Class Schedule

Order	Chapter(s)	Topics
1	Skim up to 3.5, 2.2.2	Notation/terminology; General Regression/Classification overview; Risk; Bias/variance trade-off
2	5, 6	Risk estimation; Linear regression; model selection; regularization
3	8, 7.6, 7.7	Classification with sparse logistic regression, tree-based methods
4	8	Boosting/bagging
5	9	Classification with SVMs, Kernelization
6	Notes	Neural networks
7	Notes	Special topics such as parallelization/computing

Note: This schedule is a general guide. We can (and probably will) deviate from this at some point.