

STAT 6306 – Introduction to Data Science – Spring 2018

Instructor

	Office	Email	Office Hours
Darren Homrighausen	Heroy 134	dhomrighausen@smu.edu ¹	Tues. 2 - 3pm, Thurs. 11 - 12pm

Lecture TR 3:30 PM – 4:45 PM Harold Simmons Hall 0207

Required Text *An Introduction to Statistical Learning: with Applications in R*
James, Witten, Hastie & Tibshirani. ISBN-13: 978-1461471370

Web Sites <https://github.com/STAT6306/> and Canvas

Prerequisites A strong class in multiple regression (and hence its prerequisite of an intro statistics class) and a very basic introduction to probability are the core ideas. Having the most basic exposure to linear algebra (multiplying matrices, mainly) is very helpful. Like usual, we will cover the relevant ideas in class as a review. But, the material tends to come too fast if it is your first exposure to the topics.

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.

¹Please 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. We will schedule office hours 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. Some R resources:

- [An Introduction to R](#)
- [Intro to R](#) (video series by Google Developers)
- [R studio](#) (by RStudio)
- [Simplest introduction to R](#)

Also, we will be lightly using **Github** for some aspects of the class. Some **Github** resources:

- [Comprehensive tutorial](#)
- [Github commands cheat sheet](#)
- [Try Github tutorial](#) (A quick online application going over basics)

Homework and Tests

Homework. There will be seven or so homework assignments, due every couple of weeks. Submit the homeworks online via Canvas. Homework **must** be generated from an R Markdown file. For our purposes, the best format to knit to is PDF, though HTML is permissible as well (do not submit DOC files, however).

IMPORTANT: Treat the homework submission as if you are creating a report for a boss that is statistically trained but unfamiliar with the topic being discussed in the homework. In particular, do not include output that is not specifically asked for by me or referenced by you. I reserve the right to penalize any submission with an egregious amount of excess output 10% and request a resubmission.

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. Each test is worth 15% of the course grade. There will be more details as the course progresses as to the schedule and content.

Final Policy. The final will be worth the remaining 20% of the course grade and be on May 10th at 3:00 pm. 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. Out-of-core methods
3	8, 7.6, 7.7	Classification with sparse logistic regression, discriminant analysis, tree-based methods
4	8	Boosting/bagging
5	9	Classification with SVMs, Kernelization
6	Notes	Neural networks
∞	Notes	Special topics

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

From The Provost

- **Disability Accommodations:** Students needing academic accommodations for a disability must first register with Disability Accommodations & Success Strategies (DASS). Students can call 214-768-1470 or visit <http://www.smu.edu/Provost/ALEC/DASS> to begin the process. Once registered, students should then schedule an appointment with the professor as early in the semester as possible, present a DASS Accommodation Letter, and make appropriate arrangements. Please note that accommodations are not retroactive and require advance notice to implement.
- **Religious Observance:** Religiously observant students wishing to be absent on holidays that require missing class should notify their professors in writing at the beginning of the semester, and should discuss with them, in advance, acceptable ways of making up any work missed because of the absence. (See University Policy No. 1.9.)
- **Excused Absences for University Extracurricular Activities:** Students participating in an officially sanctioned, scheduled University extracurricular activity should be given the opportunity to make up class assignments or other graded assignments missed as a result of their participation. It is the responsibility of the student to make arrangements with the instructor prior to any missed scheduled examination or other missed assignment for making up the work. (University Undergraduate Catalogue)
- **Student Learning Outcomes:** Please include in your syllabi all student learning outcomes, both those specific to your course, as well as those that satisfy major and general education requirements.
- **Final Exams:** Final course examinations shall be given in all courses where they are appropriate, and some form of final assessment is essential. Final exams or final assessments must be administered as specified in the official examination schedule, and shall not be administered during the last week of classes or during the Reading Period. Please state clearly in the syllabus the date/time and form of the final exam or assessment.