

PSTAT 131/231: DATA MINING

Spring 2018

Instructor:	Alexander Franks	Time:	Tu/Th 5:00 – 6:15
Email:	afranks@pstat.ucsb.edu	Place:	PSYCH 1924

Course Pages:

1. <https://gauchospace.ucsb.edu/courses/course/view.php?id=24758>
2. <https://piazza.com/ucsb/spring2018/pstat131/>
3. <https://rstudio.lsit.ucsb.edu>

Office Hours:

Professor Franks: Tuesday 10:00am - 12:00pm in SH 5522

Yuanbo: Wednesday 2:00-3:00pm, 5:00-6:00pm in SH 5431-V (wangy@pstat.ucsb.edu)

Franky: 2:00 - 4:00pm in SH 5432-S (fmeng@umail.ucsb.edu)

Main References (free online):

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani *An Introduction to Statistical Learning*, Springer, 2014; <http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf>).
- Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer 2009 *Elements of Statistical Learning*; <http://www.stanford.edu/~hastie/ElemStatLearn/> (231 students only)
- Garrett Golemund and Hadley Wickham *R for Data Science*; <http://r4ds.had.co.nz/>

Objectives: This course is taught at two levels, one aimed at undergraduates (131 level) and one aimed at graduate students (231 level). The lectures will be given at a level accessible and relevant to all of the students in the course. Students taking the course at the 231 level will have additional homework and quiz problems and will be graded on a different curve.

At the end of the course, a successful student will:

- comprehend many of the most common statistical learning algorithms
- be able to effectively use R for exploratory analysis, model fitting, and visualization

Prerequisites: PSTAT 120 A-B and 126 (or equivalent). Some familiarity with R is expected.

Tentative Course Outline:

- Supervised learning and classification (weeks 1-4)
- Unsupervised learning and clustering (weeks 5-8)
- Advanced topics (as time allows) (weeks 9-10)

Grading Policy:

- Homework (30%).
 - There will be 4 homeworks, due roughly every two weeks on Fridays at midnight.
 - You are allowed (and encouraged) to work with a partner. You need only turn in one homework per pair.
 - If you need help finding a partner, start with the “Search for Teammates” tool on Piazza. Homework solutions must be done in RMarkdown and turned in on GauchoSpace. We have provided a homework template that you can use to get started.
 - All code should be well documented. Homework not submitted online before the deadline will be considered late (10 point deduction). 24 hours after the deadline homework will not be accepted and no credit will be awarded. Do not wait until the night before it is due to start working!
- Midterm Exam (20%). **May 10, 2018 5PM** during class.
- Participation (20%). Lab attendance and quizzes.
 - Lab attendance is required and essential. Your attendance will be noted at the end of each lab.
 - Occassional pop quizzes (multiple-choice) will be given out in during lectures.
 - There is no make-up for missed quizzes but your lowest quiz score will be dropped when calculating your final grade.
- Final Project (30%). You are allowed (and encouraged) to work with a partner. Please refer to the final project information sheet for more information.

Homeworks, Midterm and Final Project submission:

- All files will be submitted electronically via GauchoSpace
- Submit a zip file containing:
 1. R markdown code (.Rmd file, template provided)
 2. Any additional files as needed
 3. Generated PDF file

Important Dates:

Midterm May 10, 2018
 Final Project Due Date June 13, 2018

Course Policies:

- Please sign up for Piazza: piazza.com/ucsb/spring208/pstat131. All questions related to course content, homework, R programming and quizzes should be done on Piazza. *Only questions pertaining to grades or personal issues should be handled over email.*
- Regular attendance is essential and expected!