Introduction to Statistical Machine Learning

PSTAT 131/231

Spring 2022

Instructor:	Dr. Katie Coburn (katie_m_coburn@ucsb.edu)
Teaching Assistants:	Hanmo Li (hanmo@pstat.ucsb.edu)
	Lihao Xiao (lihao@pstat.ucsb.edu)
Lecture:	TR 12:30 - 1:45 pm Buchanan 1930
Sections:	T 5:00 - 5:50 pm PHELP 1513 Li
	W 8:00 - 8:50 am PHELP 1513 Xiao
	W 9:00 - 9:50 am PHELP 1513 Xiao
	W 12:00 - 12:50 pm PHELP 1525 Li

Course information

Description

Statistical machine learning is used to discover patterns and relationships in large data sets. Topics will include: data exploration, classification and regression tress, random forests, clustering and association rules. Building predictive models focusing on model selection, model comparison and performance evaluation. Emphasis will be on concepts, methods and data analysis; and students are expected to complete a significant class project using real-world data. Prerequisites: PSTAT 120A-B and PSTAT 126 with a minimum grade of C or better. Credit units: 4.

This course is taught at two levels, one aimed at undergraduates (131 level) and one at graduate students (231 level). Lectures are given at a level designed to be accessible and relevant to all students in the course. Students taking the course at the 231 level will be assigned **additional homework questions** and are expected to do reading in **both** the 131 and 231-level textbooks, listed below. The final project assignment is the same for both levels.

Learning Outcomes

Upon completion of this course, students should be able to:

1. Explain the common statistical learning techniques conceptually and characterize some of them mathematically;
2. Describe and discuss the pros and cons of the common statistical learning methods;
3. Use R, the *tidyverse*, and *tidymodels* effectively for exploratory data analysis, model fitting, and visualization.

Piazza

The course Piazza page can be found here: https://piazza.com/ucsb/spring2022/pstat131231. Sign up with your UCSB school email and with access code **3435**.

Format

As per the Spring Quarter Instruction message from the Office of the Executive Vice Chancellor, dual-mode instruction – defined as livestreaming or classroom recording – **cannot** be provided for courses. With this in mind, lectures will be provided in person **only**. Any lecture slides will be posted on GauchoSpace before the corresponding class. If special circumstances require it, specific lectures may be recorded and shared, but this is an exception. If you cannot attend a lecture in person, let the instructor know, and we may be able to make adjustments as necessary.

The class will progress according to a weekly schedule (provided below) and interact outside of the classroom via Piazza. Students are strongly encouraged to interact on Piazza. The instructional team will monitor Piazza regularly. Piazza should be used for all discussion related to course content, homework, R coding, etc.

Both lecture periods each week will consist of slides that cover related concepts, sometimes including live coding to demonstrate how to apply the concepts using the R language. Section each week will be dedicated to completing an R-based lab.

Materials

The books and software used for this class are as follows. Note that all are freely available online and do not require purchase.

- **Required for PSTAT 131 students**: *An Introduction to Statistical Learning with Applications in R* (or ISLR), by G. James, D. Witten, T. Hastie, R. Tibshirani. Available: https://www.statlearning.com/
- **Required for PSTAT 231 students**: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (or ESL), by T. Hastie, R. Tibshirani, J. Friedman. Available: https://hastie.su.domains/ElemStatLearn/
- **Required for all**:
  - The R statistical environment version 4.1.2, available here: https://www.r-project.org/
  - *R for Data Science*, by H. Wickham, G. Grolemund. Available: https://r4ds.had.co.nz/index.html
  - The RStudio IDE (integrated development environment) Desktop version. Available: https://www.rstudio.com/

Assessments

Your attainment of course learning outcomes will be measured by the following assessments, with the relative weighting indicated in parentheses. All assignments within each category are given equal weight.

Collaboration and open science are fundamental aspects of data science. You can (and are encouraged) to use the Internet, StackOverflow, etc. as a resource when coding. However, you **should not** directly copy code or answers. Your code should be well documented; include comments describing each line. If you use a source for help coding, you **must** cite it.

You will **not** be graded on your completion of lab work or your attendance of labs, but doing the labs is likely to be greatly beneficial for your homework and final project grades. You are strongly encouraged to attend lab sections and follow along with the materials for each lab, which will be posted each week.

- **Homework** (50%). Homework is assigned **every week** and is due by Sundays at 11:59 PM PDT. Each homework is worth 50 points. One homework is assigned each week, except for the last week, for a total of **nine** homework assignments. Your **lowest** homework score will be dropped.
- **Final Project** (50%). The final project is a significant portion of the class and allows students to demonstrate their understanding of the material by working with a data set of their own choosing. Each student should begin thinking about their project and looking for a data set immediately; a page

on GauchoSpace provides a list of possible resources. You'll receive a lot of guidance about the project over the course of the quarter.

The project will require a large amount of work and cannot be satisfactorily completed in only a day (or even a week). You are **strongly advised** to work on it in stages throughout the quarter. See GauchoSpace for more detailed information and a rubric.

Tentative Schedule

The tentative weekly schedule is indicated below. The topics and reading are subject to change based on the progress of the class.

Week	Date	Topic	Reading	Assignments	Final Project Stage
Week 1	Mar. 29	Introduction & basics of ML	Ch 1 ISLR & ESL	Homework 1	Decide on topic
	Mar. 31	Exploratory data analysis			and locate data set(s)
Week 2	Apr. 5	Bias-variance tradeoff	Ch. 2 ISLR & ESL	Homework 2	Load and tidy data
	Apr. 7	Linear regression	Ch. 3 ISLR & ESL		
Week 3	Apr. 12	Logistic regression	Ch. 4 ISLR & ESL	Homework 3	
	Apr. 14	More on linear & logistic			
Week 4	Apr. 19	Resampling methods	Ch. 5 ISLR, 7 ESL	Homework 4	Run and write up descriptive analyses
	Apr. 21	Discriminant analysis	Ch. 4 ISLR & ESL		
Week 5	Apr. 26	Discriminant analysis cont.		Homework 5	
	Apr. 28	Linear model selection	Ch. 6 ISLR & ESL		
Week 6	May 3	Regularization	Ch. 5 ESL	Homework 6	Run models
	May 5	Beyond linearity	Ch. 7 ISLR, 9 ESL		
Week 7	May 10	Decision trees	Ch. 8.1 ISLR	Homework 7	
	May 12	Tree-based methods	Ch. 8.2 - 8.3 ISLR, 10 ESL		Write up results
Week 8	May 17	K-means	Ch. 12.4 ISLR, 14 ESL	Homework 8	
	May 19	Hierarchical clustering			Work on draft of paper
Week 9	May 24	PCA		Homework 9	
	May 26	SVM (I)	Ch. 9 ISLR, 12 ESL		
Week 10	May 31	SVM (II)			Make edits
	June 2	Neural networks	Ch. 10 ISLR, 11 ESL		
	June 6			Final Project Due	Begin final draft

Time Commitment

The course is 4 credit units; each credit unit corresponds to an approximate time commitment of 3 hours. You should expect to allocate 12 hours per week to the course. If you find yourself spending considerably more time on the course on a regular basis, please let the instructor or TAs know so that we can help you balance the workload.

A suggested allocation of this time is as follows:

- Reading and class time: 3 hours (25%)
- Homework: 4.5 hours (37.5%)
- Sections: 1 hour (8.3%)
- Final project: 4.5 hours (37.5%)

Course Policies

Communication

There are five means of communication with other students or the instructional team: during/after class, Piazza, office hours, email, and individual appointments. **Please use them in that order**.

1. **During/after class**. The easiest way to get in touch with me is to walk up after class (if meeting in person) or message me in Zoom chat during/after class (if meeting online).
2. **Piazza**. Consider Piazza your primary communication resource for the course; it is a way to stay connected with the instructor, the TAs, and your classmates throughout the term. You can start and participate in conversations, ask and answer questions, create discussions for specific purposes as you see fit (*e.g.*, forming a study group), and exchange direct messages with anyone in the class.
3. **Office hours**. Office hours are offered at a minimum of twice weekly. One session will be conducted by the instructor, the other(s) by various combinations of TAs. These are opportunities to interact informally, ask questions, and discuss course material or assignments.
4. **Email**. Please use email with discernment for simple communication. A response is guaranteed within 48 weekday hours (so if you email on Friday afternoon, you may not receive a reply until Tuesday afternoon). In light of this response policy, bear in mind that you are likely to receive replies to messages or posts in Piazza much faster than replies to email. If your message is time-sensitive, please indicate so in the subject, and we will do our best to respond promptly.
    - Note: I am notoriously slow at replying to email – if the matter is urgent you might need to email me multiple times, or catch me in person/at office hours. Don't feel bad about doing this and please be assured there is never anything personal about a delayed reply.
5. **Appointment**. You can schedule 20-minute appointments with me as needed. These appointments may be either on Zoom or in person. This mode of communication is best suited to more complex or nuanced communication regarding personal matters. If you schedule an appointment, you will be prompted to indicate what you wish to discuss.

Extra Credit

If the class reaches a 90% submission rate of ESCI surveys at the end of the quarter, the **entire class** will receive 2 free points on the final project.

Grades

Your overall grade in the course will be calculated as the weighted average of the proportions of total possible points in each assessment category according to the weightings indicated in the Assessments section and reported as a percentage rounded to two decimal places. Tentatively, letter grades will be assigned according to the rubric below – a curve is possible, but not guaranteed.

A	93% – 100.00%
A-	90% – 92.99%
B+	87% – 89.99%
B	83% – 86.99%
B-	80% – 82.99%
C+	77% – 79.99%
C	73% – 76.99%
C-	70% – 72.99%
D+	67% – 69.99%
D	60% – 66.99%
F	0% – 59.99%

You can keep track of your marks on individual assessments, your marks in each assessment category, and your overall grade in the GauchoSpace gradebook. Please notify the instructor or TAs of any errors in grade *entry*; please do not attempt to negotiate the grades themselves. If at the end of the course you believe your grade was unfairly assigned, you are entitled to contest it according to the procedure outlined here in the UCSB General Catalog.

Deadlines

You receive two free late homework submissions without penalty. This policy applies only to homeworks. There is no need to notify the instructor or TAs to use these late allowances; simply submit within one week of the original deadline.

Non-exempted homeworks turned in within one week of the deadline will be awarded 50% credit. No credit will be awarded for homework turned in more than one week late; please plan ahead and submit your work on time.

The final project deadline is firm and no late submissions will be accepted.

Extensions

Extensions may be granted based on individual circumstances at the instructor's discretion.

Conduct

Please be mindful of maintaining respectful and kind communication. You are expected to uphold the UCSB student code of conduct in your behavior when in class, in section, or interacting with other students or the instructional team. You can find the student code of conduct on the Office of Student Conduct website from this page. If you are uncomfortable with the conduct of another individual for any reason, please notify the instructor or TAs.

Academic Integrity

Please maintain integrity in learning. Your work in the course must be your own. Any form of plagiarism, cheating, misrepresentation of individual effort on assignments and assessments, falsification of information or documents, or misuse of course materials compromises your own learning experience, that of your peers, and undermines the integrity of the UCSB community. Any evidence of dishonest conduct will be discussed with the student(s) involved and reported to the Office of Student Conduct. If you are reading this, thank you for paying close attention to the syllabus. Make an instructor-only post on Piazza with the code word "integrity.'' Depending on the nature of the evidence and the violation, penalty in the course may range from loss of credit to automatic failure. For a definition and examples of dishonesty, a discussion of what constitutes an appropriate response from faculty, and an explanation of the reporting and investigation process, see the OSC page on academic integrity.

Accommodations

Reasonable accommodations will be made for any student with a qualifying disability. Such requests should be made through the Disabled Students Program (DSP). More information, instructions on how to access accommodations, and information on related resources can be found on DSP website. Remote learning may present unique accommodation needs requiring additional flexibility; students receiving accommodation via DSP are invited to discuss this with the instructor if desired.

Student Evaluations

Toward the end of the term, you will be given an opportunity to provide feedback about the course. Your suggestions and assessments are essential to improving the course, so please take the time to fill out the evaluations thoughtfully.

Student Resources

Any students in need are encouraged to make use of the following resources.

- Financial Crisis Response Team
  https://food.ucsb.edu/about/committees/financial-crisis-response-team
- Food Security and Basic Needs (Food, Housing, Technology) Advising Center
  https://food.ucsb.edu/resources/basic-needs-advocates
- Undocumented Student Services
  http://www.sa.ucsb.edu/dreamscholars/home
- Campus Advocacy, Resources, and Education (CARE)
  https://care.ucsb.edu/
  24/7 Confidential Phone: (805) 893-4613
- The Trevor Project
  https://www.thetrevorproject.org/
- Counseling and Psychological Services (CAPS)
  https://caps.sa.ucsb.edu/
  24/7 Counselors: (805) 893-4411, press 2