DSCI445: Statistical Machine Learning

Fall 2023

http://dsci445-csu.github.io

Dr. Andee Kaplan
Lectures/Labs: TTH 8am - 9:15am Natural Resources Building 109
Office Hours: TH 1:30pm - 3:30pm Statistics Bulding 208
andee.kaplan@colostate.edu

Course Objectives

Statistical Learning refers to a set of tools for modeling and understanding complex datasets. The area combines knowledge from statistics and computer science to tackle these "big data" problems and has become a popular area of work.

By end of course, students will be able to:

1. Formulate prediction problems as statistical machine learning problems (classification, regression, clustering).
2. Choose and apply the appropriate learning methods to their data.
3. Conduct well thought-out statistical machine learning experiments and interpret the results.
4. Write technical reports describing their work.

COVID-19

Important information for students: All students are directed to report any COVID-19 symptoms to the university immediately, as well as exposures or positive test results from a medical provider or home test.

If you suspect you have symptoms, or if you know you have been exposed to a positive person or have tested positive for COVID, (even with a home test), you are directed to fill out the COVID Reporter (https://covid.colostate.edu/reporter/). If you know or believe you have been exposed, including living with someone known to be COVID positive, or are symptomatic, it is important for the health of yourself and others that you complete the online COVID Reporter. Do not ask your instructor to report for you. If you do not have internet access to fill out the online COVID-19 Reporter, please call (970) 491-4600. You

may also report concerns in your academic or living spaces regarding COVID exposures through the COVID Reporter. You will not be penalized in any way for reporting. When you complete the COVID Reporter for any reason, the CSU Public Health office is notified. Students who report symptoms or a positive antigen test through the COVID Reporter may be directed to get a PCR test through the CSU Health Network's medical services for students.

For the latest information about the University's COVID resources and information, please visit the **CSU COVID-19 site**: https://covid.colostate.edu/.

Prerequisites

DSCI 320, DSCI 369 and STAT 341.

Texts

An Introduction to Statistical Learning with Applications in R (2017) by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani – Available free here: http://faculty.marshall.usc.edu/gareth-james/ISL/

Optional Reference: The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2009) by Trevor Hastie , Robert Tibshirani, and Jerome Friedman – Available free here: https://web.stanford.edu/~hastie/ElemStatLearn/

Computing

We will use RStudio (https://rstudio.com), R (https://r-project.org), GitHub (https://github.com). All software is free and open source.

Please install on your own computer or use the class RStudio Server (details to follow).

For your homeworks, you are free to use any other language you like, but I may not be able to help you with your computing if the language is unfamiliar to me.

Classwork and Grading

All graded classwork must be fully **reproducible** and **readable** by the instructor and TA. In other words, we need to be able to run your code and have it produce the product you turned in and your turned in document must be legible by the grader. If this is not the case, it will be reflected in the grading. A copy of your homework will need to be turned in to https://canvas.colostate.edu and the corresponding document used to generate your

homework will need to exist on the server for full credit.

Homework (70%) Homework will be assigned bi-weekly. All homework assignments are due at **11:59pm on the due date**. Each homework assignment will receive equal weight in the final grade and the one lowest homework assignment grades will be dropped. Late work is not accepted except in rare cases (see Documented emergencies below).

Project (30%) There will be a final project that will consist of an analysis of real data using the tools learned in class (this can include participation in an online data science competition, e.g. Kaggle).

You will write a paper and give an in-class presentation. More details will be announced later.

Grades will be assigned according to the following intervals:

A	A-	B+	B	B-	C+	C	D	F
[100, 93]	(93, 90]	(90, 87]	(87, 83]	(83, 80]	(80, 77]	(77, 70]	(70, 60]	(60, 0]

Any grading dispute must be submitted in writing to me within one week after the work is returned.

Extra credit Any extra credit will be announced in lecture only. If you miss lecture, you *may* miss chances for extra credit.

Policy Regarding Academic Honesty

Statisticians and data scientists need to have high ethical standards. Thus, I expect each of you to hold high ethical standards and to act with academic integrity in this class. If you have questions about what integrity means, please feel free to ask me. Behavior that will not be tolerated in this class includes turning in a copy of somebody else's homework or code as your own, copying from somebody's exam, or failure to cite sources.

This course adheres to the CSU Academic Integrity Policy as found on the Students' Responsibilities pages of the CSU General Catalog in the Student Conduct Code. Violations will result in zero points for the assignment as a minimum penalty. In addition, CSU policy requires instructors to report violations to CSU's Office of Conflict Resulution.

Documented Emergencies

If you have a problem that will require you to miss a due date, please discuss this with me in advance if possible. I can grant a rare exception when the reason relates to severe and unavoidable medical or personal emergency. Documentation will be required. Things that

typically are not an emergency: vacation, family reunions, ordinary work commitments, job seeking, or other voluntary events. Please schedule these so that they do not conflict with your classes.

Support Services Available

CSU COVID-19 Recovery Page (https://covidrecovery.colostate.edu) On our road to recovery during these unprecedented times, Colorado State University is committed to the health of our students, faculty and staff, as well as to the health of our university and our ability to continue to empower our community through our land-grant mission of academics, research and outreach.

CSU Health Network Counceling Services A variety of services are offered (151 W. Lake St., Drop-in hours: Monday-Friday 9am-4pm). If you are having difficulty coping, are feeling depressed, or need other psychological assitance, please contact the counceling center.

CSU Disability Center Located in the TILT building. Students with both permanent and temporary limitations and health conditions (physical and mental health) are eligible for support. If you need specific accommodations in this class, please meet with me outside of class to discuss your needs as early in the class as possible.

CSU TILT The Institute for Learning and Teaching has programs to help students improve their study habits, reduce test anxiety, learn about academic integrity, and more.

Disclaimer

I reserve the right to make amendments to the syllabus and the schedule throughout the semester. Any updates will be posted on the class website and announced via e-mail and in class.