COGS 239: Introduction to Statistical Learning

Professor Ramesh Srinivasan, Cognitive Sciences

Version 0 09/26/24

E-mail: `r.srinivasan@uci.edu`
Office Hours: **By appointment**
Office: SBSG 2336

Course Description

Introduction to statistical learning for scientific research. Topics include linear regression, logistic regression, model selection, cross-validation, ridge and Lasso regularization, classification trees, ensemble methods, random forests, unsupervised learning, PCA, PLS regression, k-means clustering, Gaussian mixture models, covariance models, Topological Data Analysis (TDA) .

Prerequisites

Knowledge of Calculus, computer programming, and undergraduate Statistics

Course Organization

For each topic there are lectures covering coding data analysis and visualization methods, and one advanced lecture covering the mathematical basis of the methods and more advanced methods. In addition, tutorial materials will be provided each week for you to study at your own pace. A rough weekly schedule is provided in **Table 1**.
I will do my best to record Lectures and make them available in the class YouTube channel.

Textbook

James G, Witten D, Hastie T, Tibshirani R, Taylor J. (2023) An Introduction to Statistical Learning (with Applications in Python). Springer.
This book is free for UCI students, and a download link is available on the class webpage. There is also a version of this book with Applications in R.

Introduction		Reading Ch 2.1
Week 1	(M)	What is statistical learning?
Exploratory Data Analysis		Reading Ch 2.2
Week 1	(W)	Exploratory Data Analysis
Week 1	(F)	Visualization and Communication
Week 2	(M)	Trade-off between Prediction and Interpretability
Linear Regression		Reading Ch 3.1-3.3, 4.6
Week 2	(W)	Prediction with Regression Models
Week 2	(F)	Multiple Linear Regression
Week 3	(M)	Generalized Linear Models
Linear Classifiers		Reading Reading 4.1-4.5, 5.1
Week 3	(W)	Prediction with linear classifiers
Week 3	(F)	Comparison of classifiers
Week 4	(M)	Cross-validation on Classification problems
Model Selection		Reading 6.1-6.2
Week 4	(W)	Ridge and Lasso Regularization
Week 4	(F)	Model Selection and Interpretation
Week 5	(M)	Model Optimization and Elastic Net Regularization
Decision Trees		Reading 8.1-8.2.2
Week 5	(W)	Predictive analysis with Decision Trees
Week 5	(F)	Ensemble Methods: Random Forests
Week 6	(M)	Information Theoretic Basis of Decision Trees
Ensemble Methods		Reading 8.2.1-8.2.4
Week 7	(W)	Ensemble Methods: Bagging and Boosting
Week 7	(F)	Interpretation of Decision Tree Models
Week 8	(M)	Unsupervised learning: Clustering
Clustering		Reading 12.1, 12.4
Week 8	(W)	Clustering: K-means and Gaussian mixture models
Week 8	(F)	Fitting multivariate probability distributions
Week 9	(M)	Graphical Models
PCA and PLS		Reading 6.4.1-6.4.2, 12.2, 6.3
Week 9	(W)	Challenges of working with high-dimensional data
Week 9	(F)	Principal Components Analysis
Week 10	(M)	Partial Least Squares Regression
Covariance Models and Topological data Analysis		
Week 10	(W)	Topological Data Analysis
Week 10	(F)	Topological Data Analysis

Table 1: Weekly Schedule

Laptop Requirement

The software used in this class will work on Windows/Mac/Linux. Please contact me if there is an issue with this. I would like you to have a laptop in class. If you need to borrow one, I can find you one.

Assignments

Homework

Homework will be assigned regularly in the class, and usually due the following week.All submissions will be through Github classroom

Midterms

There are two in class/take home midterms for this course. The midterms will cover both theoretical material and data analysis problems.

Final

The final exam is to do a Project. The project will take the methods you have learned and apply them to analyze and visualize a publicly available data set. You project must include (1) Exploratory Data Analysis, (2) Visualization, (3) Predictive analytics, and (4) Interpretation of the predictive models. It would likely be more interesting if you did your own project with some data that interests you. If you need help selecting something let me know.

The Project is submitted in the form of a Github repository that provides the code to complete the data analysis and visualization presented along with a very brief write up.

Grading Policy

Midterms: 40% of grade
Homework: 40% of grade.
Project: 20% of grade.

Late Work Policy

It is essential that you complete all the work of the class. You cannot learn this material skipping over a lesson. The material for each week in the class builds on previous weeks, so if you fall behind by a week it is very difficult to recover.
If you are not able to meet the deadline posted for homework, you should still submit the assignment late for partial credit and do the assignments in order. If you submit late work please

notify me by email as I may not be aware of your late submission.

Concurrent enrollment

This course may be offered concurrently as COGS 105 (undergraduate) and COGS 205C (graduate). When offered concurrently, the two courses will overlap in two lectures each week. For the third lectures, COGS 105 will have a practicum where the coding of methods is discussed. Students in COGS 205B will be expected to be able to do the coding without detailed instruction, and will instead get an additional lecture each week with advanced methods and a stronger theoretical character. The homework for 205B students will incorporate additional theoretical problems, and 205B students have two midterms.

GitHub

This course is organized from GitHub Classroom. To use GitHub Classroom, you need to make an account on https://github.com. Please make this a professional account, with a username that clearly identifies you, and use your UCI email. I highly recommend you maintain a GitHub account with your work in this class and other classes you take. When I review postdoc applications, I look for the candidates' GitHub page.

What would make this process the most effective is if you were to learn to use git software to interact with your GitHub code repositories. This can either be done with a commmand line interface (CLI) or using GitHub Desktop or by installing a plugin inside VS Code. A tutorial video is provided on the class GitHub page.

https://desktop.github.com/

Inside VS Code there are some instructions of how to set up the Git plugin. You may have to also install Git on your computer. Again, please watch the tutorial video.

Command Line Interface (CLI) is actually the best way to work with GitHub.

Software

The course will be taught in the Python programming language. Thus I recommend you follow the instructions below to set up an appropriate Python environment for the class, to be able to run my code.
However, you are free to carry out the assignments in the class in any programming language you would like. All of the methods discussed here are available in R (along with the textbook) and should be available in Matlab.

Anaconda Python

All of the examples in this class make use of the Python programming language. Install Anaconda Python Libraries on your (Windows, Mac, or Linux) computer.

Anaconda Python - https://www.anaconda.com/products/individual

Why Anaconda Python?

Your computer may already have Python installed on it, which in principle you could configure to use for this class. It is much easier to simply install the Anaconda distribution of Python which is free and comes with an nearly complete library of software for scientific computing as well as a number of other useful tools like the *Jupyter Lab* IDE.

VS Code

My preferred IDE for Jupyter Notebooks is Visual Studio Code.

https://code.visualstudio.com/

There is no requirement to use VS Code. You can use Jupyter Notebooks or Jupyter Lab or Spyder, if you are already used to them and like them. In industry, many people like PyCharm, but I think its overkill for a class. All of the materials in this class should work in any IDE.

Disability services, academic dishonesty, and copyright policy

Disability Services link: https://dsc.uci.edu/
Academic Dishonesty link: https://aisc.uci.edu/students/academic-integrity/index.php
Copyright policy link: http://copyright.universityofcalifornia.edu/use/teaching.html