UCLA
Computer Science Department

CS 97: Introduction to Data Science Summer 2024

Course Description: The fundamental question this course aims to address is: given data arising in real-world, how does one analyze that data so as to understand the corresponding phenomenon. This course will cover topics in machine learning, data analytics, and statistical modeling classically employed for prediction. The course will be a blend of theoretical and practical instruction, providing a comprehensive, hands-on overview of the Data Science domain. The course will seek to teach students the data science lifecycle: data selection and cleaning, feature engineering, model selection, and prediction methodologies.

Instructor: Shichang Zhang (shichang@ucla.edu)

Time: Mon - Fri (9:00 am - 11:50 am & 1:00 pm - 4:00pm) from June 24th to July 12th

Classroom: Boelter Hall 3400

Teaching Assistants:

- Amin Doosti (doostiamin@ucla.edu)
- Howard Zhu (howardzhu8@gmail.com)
- Weikai Li (weikaili@cs.ucla.edu)
- Zongyue Qin (qinzongyue@cs.ucla.edu)

Course Material Lectures, discussions, labs, and assignments. Self-contained slides will be used.

Textbooks (Optional):

- Fundamentals of Machine Learning for Predictive Data Analytics Algorithms, Worked Examples, and Case Studies. — by John D. Kelleher, Brian Mac Namee, Aoife D'Arcy
- Machine Learning: An Algorithmic Perspective, Second Edition Part of: Chapman & Hall / CRC Machine Learning & Pattern Recognition (21 Books) — by Stephen Marshland.
- Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. O'Reilly Media, Inc., 2012 — by McKinney, Wes.
- Probabilistic programming and Bayesian methods for hackers., 2015 — by Pilon, Cameron Davidson.

Class Communication

- Important class announcements will be made through the online class forum BruinLearn. Course-related files like lecture slides and homework will be posted on BruinLearn. `https://bruinlearn.ucla.edu/courses/188172`
- Questions and discussions of the course material on Piazza. If you have any questions regarding class materials, please ask them on Piazza. `https://piazza.com/ucla/summer2024/cs97`

Grading

- Homework 40% (Homework 1 to 4, 10% each)
- Midterm Exam 25%
- Final Course Project 30%
- Participation 5%

Academic Integrity: You are encouraged to discuss homework problems with friends, but you must write your solutions individually. All students are expected to follow the UCLA Student Conduct Code, which prohibits cheating, fabrication, and multiple submissions.

Tentative Schedule: The schedule might change depending on the progress of the class.

Week	Date	Lecture (9:00 am – 11:50 am)	1:00 pm – 4:00 pm	Deadlines
Week 1	6/24	Introduction. Know Your Data	Discussion and Lab	Homework 0 out
	6/25	Linear Regression	Discussion and Lab	
	6/26	Regularization. Model Selection	Discussion and Lab	Homework 1 out
	6/27	Logistic Regression	Discussion and Lab	
	6/28	Classification Evaluation. KNN	Seminar Discussion and Lab	Homework 1 due Homework 2 out Project: topic decided
Week 2	7/1	SVM and Decision Tree	Discussion and Lab	Homework 2 due Homework 3 out
	7/2	Perceptron and NN	Discussion and Lab	Homework 3 due Homework 4 out
	7/3	NN: Design/Training/Regularization	Discussion and Lab	
	7/4	No class	No class	
	7/5	Clustering and K-Means	Seminar Discussion and Lab	Homework 4 due Project: first idea implemented
Week 3	7/8	Application: Health	Exam (90 minutes) Q & A	
	7/9	Application: Text	Discussion and Lab	
	7/10	Application: Image	Discussion and Lab	
	7/11	Application: Recommender Systems	Discussion and Lab	Project: completed & slides prepared
	7/12	Project Presentation	Project Presentation Closing	