UCLA
Computer Science Department

CS 97: Introduction to Data Science   Summer 2023

Course Description: This course is an adapted from CS 188: Data Science of Fundamentals. 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: Yizhou Sun (yzsun@cs.ucla.edu)

Lecture classroom: TBD

Undergrad Tutors:
- TBD
  - Cohort A
- TBD
  - Cohort B
- TBD
  - Cohort C
- TBD
  - Cohort D
- TBD
  - Cohort E
- TBD
  - Cohort F
- TBD
  - Cohort G
- TBD
  - Cohort H

Textbook (Optional):
1. Fundamentals of Machine Learning for Predictive Data Analytics Algorithms, Worked Examples, and Case Studies.
2. Machine Learning: An Algorithmic Perspective, Second Edition Part of: Chapman & Hall/Crc Machine Learning & Pattern Recognition (21 Books) | by Stephen Marsland.
3. Python for data analysis: Data wrangling with Pandas, NumPy, and IPython. O'Reilly Media, Inc., 2012 | by McKinney, Wes.
4. Probabilistic programming and Bayesian methods for hackers., 2015 | by Pilon, Cameron Davidson.

Course material: Lectures, assignments and solutions will be posted on BruinLearn

Class Communication: Important class announcements will be done through online class forum on BruinLearn/Piazza. If you have any questions regarding class materials, they also need to be asked on Piazza.

Piazza: piazza.com/ucla/summer2022/cs97

Grading:

Homework   40%
MIDTERM   25%
Project   30%
Participation   5%

You may discuss problems with friends, but you must write your solutions individually. I expect all students to follow the UCLA Student Conduct Code, which prohibits cheating, fabrication, and multiple submissions.

Tentative Schedule:

Week 1

	Morning (9:00 -- 11:50)	Afternoon (1:00 -- 4:00)	Deadlines
6/26	Introduction; Know Your Data	Discussion and Lab	Homework 0 out
6/27	Linear Regression	Discussion and Lab	
6/28	Regularization; Model Selection	Discussion and Lab	Homework 1 out
6/29	Logistic Regression	Discussion and Lab	
6/30	Classification Evaluation; kNN	Lab tour / seminar	Homework 1 due Homework 2 out Course Project: topic decided

Week 2

	Morning (9:00 -- 11:50)	Afternoon (1:00 -- 4:00)	Deadlines
7/3	SVM and Decision Tree	Discussion and Lab	Homework 2 due Homework 3 out
7/4	No Class	No Class	
7/5	Perceptron and NN	Discussion and Lab	Homework 3 due Homework 4 out
7/6	NN: Design/Training/ Regularization	Discussion and Lab	
7/7	Clustering and K-Means	Lab tour / seminar	Homework 4 due Course Project: first idea implemented

Week 3

	Morning (9:00 -- 11:50)	Afternoon (1:00 -- 4:00)	Deadlines
7/10	Application: Health	Exam (90 mins); Discussion and Lab	
7/11	Application: Text	Discussion and Lab	
7/12	Application: Image	Discussion and Lab	

7/13	Application: Recommender Systems	Discussion and Lab	Course Project: completed; slides prepared
7/14	Course Project Presentation	Course Project Presentation / Closing	