Intro and First Day Stuff

Lecture 1 - CMSE 381

Prof. Mengsen Zhang

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

Mon, Jan 13, 2025

People in this lecture

Dr. Zhang (she/they)
Assistant Professor, CMSE, MSU

Omeiza Olumoye (he/his/him)
Graduate Student, CMSE, MSU

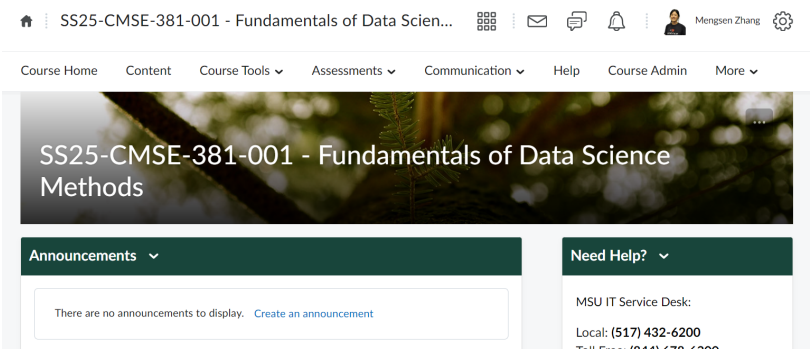
What is this course about?

Topics:

- Fundamental concepts of data science
- Regression
- Classification
- Dimension reduction
- Resampling methods
- Tree-based methods, etc.

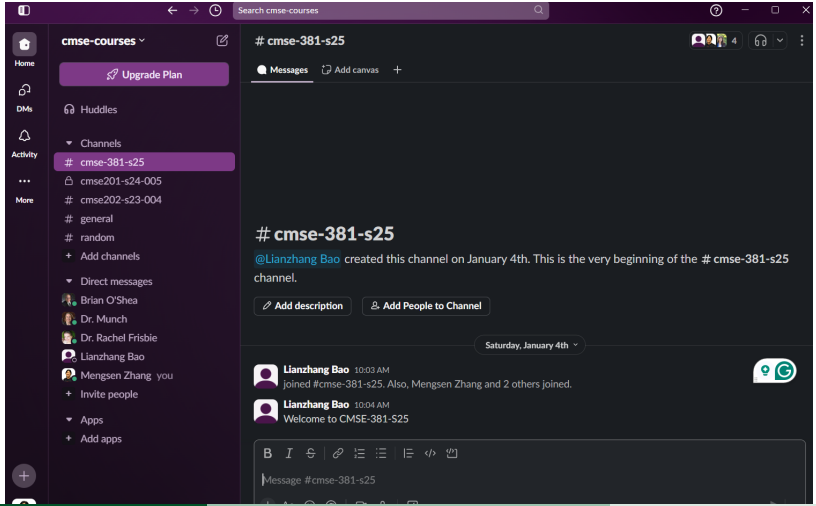
D2L and where to find grades

https://d2l.msu.edu/d2l/home/2143373



Slack and where to find announcements/ask questions

Join `cmse-courses` slack: https://tinyurl.com/cmse-courses-slack-invite

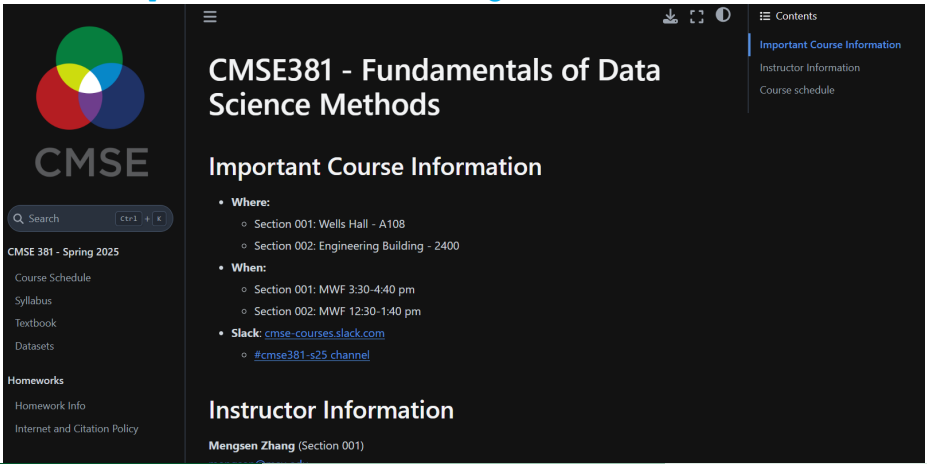


Course Website and where to find slides and jupyter notebooks

https://cmse.msu.edu/CMSE381

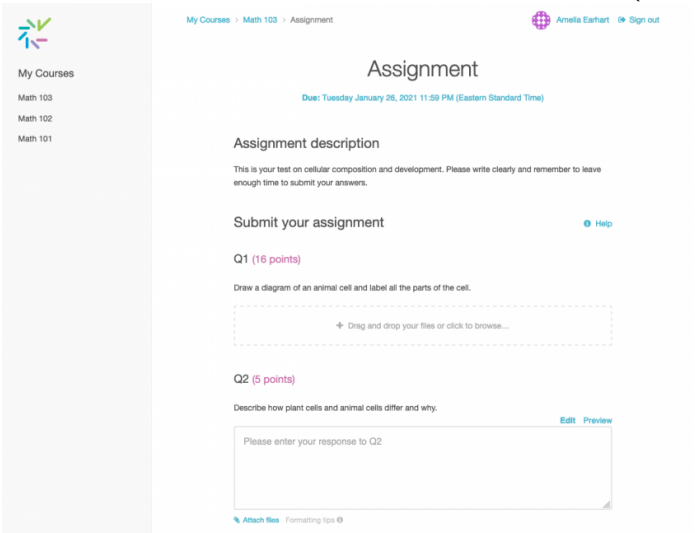
–or–

https://msu-cmse-courses.github.io/CMSE381-F25/



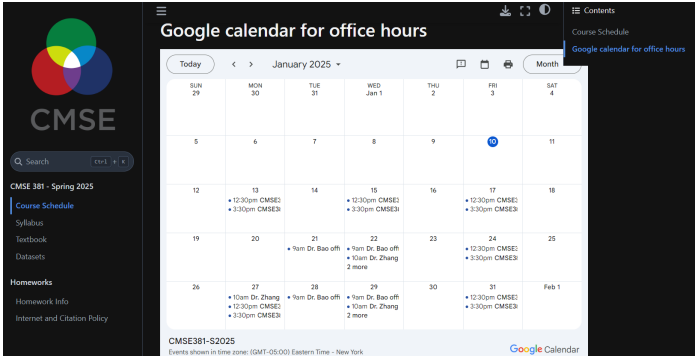
Crowdmark and where to submit homework

No URL: You will get an automated email from the system (I think.....?)



Office hours

Zoom link: Look up on calendar on the website



Dr. Zhang

Time MW 10-11 am (Starting 1/22)

Zoom & EGR 1514

Omeiza Olumoye

Time T-Th 3-4 PM

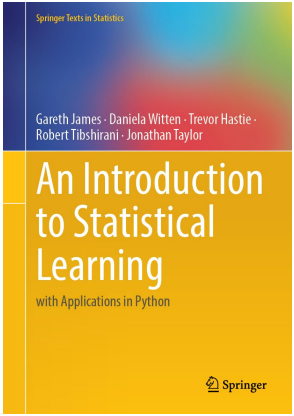
Zoom 959 4655 2093 (PW: cmse381)

& EGR (Room TBD)

Textbook

Free download

https://www.statlearning.com/



Class Structure

- Class is a combination of lecture time, and group work/coding time.
  - Bring computer every day
  - Jupyter notebooks
  - Python
- Once a week, there will be a short check-in quiz. This will be basic content realted to lectures since the last class. Possible questions include checking on definitions, or basic understanding of major ideas.
  - 10 points per quiz
  - Drop two lowest grades

Class Structure Pt 2

- Homeworks due once a week, midnight of the day marked in the schedule (mostly Sundays).
  - 20 points per homework
  - Drop two lowest grades
  - Sliding scale:
    - ★ 24 hours late: 5% penalty.
    - ★ 48 hours late: 15% penalty.
    - ★ >48 hours: No late work accepted.
- Three Midterms
  - See schedule for dates
  - 100 points each
  - Not cumulative
- One Project
  - Analyze dataset using tools in class, submit written report
  - 100 points
  - Due at the end of the semester

Approximate schedule

Up to date version: https://msu-cmse-courses.github.io/CMSE381-S25/Course_Info/Schedule.html

CMSE381_S2025_Schedule : Sheet1

Lec #	Date		Topic	Reading	HW
1	M	1/13	Intro / Python Review	1	
2	W	1/15	What is statistical learning	2.1	
3	F	1/17	Assessing Model Accuracy	2.2.1, 2.2.2	
	M	1/20	MLK - No Class		
4	W	1/22	Linear Regression	3.1	
5	F	1/24	More Linear Regression	3.1	HW #1 Due Sun 1/26
6	M	1/27	Multi-linear Regression	3.2	
7	W	1/29	Probably More Linear Regression	3.3	
8	F	1/31	Last of the Linear Regression		HW #2 Due Sun 2/1
9	M	2/3	Intro to classification, Bayes classifier, KNN classifier	2.2.3	
10	W	2/5	Logistic Regression	4.1, 4.2, 4.3.1-3	
11	F	2/7	Multiple Logistic Regression / Multinomial Logistic Regression	4.3.4-5	HW #3 Due Sun 2/9
	M	2/10	***Project Day & Review***		
	W	2/12	**Midterm #1**		
12	F	2/14	Leave one out CV	5.1.1, 5.1.2	
13	M	2/17	k-fold CV	5.1.3	
14	W	2/19	More k-fold CV	5.1.4-5	
15	F	2/21	k-fold CV for classification	5.1.5	HW #4 Due Sun 2/23
16	M	2/24	Subset selection	6.1	
	M	2/10	***Project Day & Review***		
	W	2/12	**Midterm #1**		
12	F	2/14	Leave one out CV	5.1.1, 5.1.2	
13	M	2/17	k-fold CV	5.1.3	
14	W	2/19	More k-fold CV	5.1.4-5	
15	F	2/21	k-fold CV for classification	5.1.5	HW #4 Due Sun 2/23
16	M	2/24	Subset selection	6.1	
17	W	2/26	Shrinkage: Ridge	6.2.1	
18	F	2/28	Shrinkage: Lasso	6.2.2	
	M	3/3	Spring Break		
	W	3/5	Spring Break		
	F	3/7	Spring Break		
19	M	3/10	PCA	6.3	
20	W	3/12	PCR	6.3	
	F	3/14	***Review***		HW #5 Due Sun 3/16
	M	3/17	**Midterm #2**		
21	W	3/19	Polynomial & Step Functions	7.1-7.2	
22	F	3/21	Step Functions; Basis functions; Start Splines	7.2-7.4	HW #6 Due Sun 3/23
23	M	3/24	Regression Splines	7.4	
24	W	3/26	Decision Trees	8.1	
25	F	3/28	Random Forests	8.2.1, 8.2.2	HW #7 Due Sun 3/30
26	M	3/31	Maximal Margin Classifier	9.1	
27	W	4/2	SVC	9.2	
28	F	4/4	SVM	9.3, 9.4	HW #8 Due Sun 4/6
29	M	4/7	Single Layer NN	10.1	
30	W	4/9	Multi Layer NN	10.2	
19	M	3/10	PCA	6.3	
20	W	3/12	PCR	6.3	
	F	3/14	***Review***		HW #5 Due Sun 3/16
	M	3/17	**Midterm #2**		
21	W	3/19	Polynomial & Step Functions	7.1-7.2	
22	F	3/21	Step Functions; Basis functions; Start Splines	7.2-7.4	HW #6 Due Sun 3/23
23	M	3/24	Regression Splines	7.4	
24	W	3/26	Decision Trees	8.1	
25	F	3/28	Random Forests	8.2.1, 8.2.2	HW #7 Due Sun 3/30
26	M	3/31	Maximal Margin Classifier	9.1	
27	W	4/2	SVC	9.2	
28	F	4/4	SVM	9.3, 9.4	HW #8 Due Sun 4/6
29	M	4/7	Single Layer NN	10.1	
30	W	4/9	Multi Layer NN	10.2	
31	F	4/11	CNN	10.3	HW #9 Due Sun 4/13
32	M	4/14	Unsupervised learning / clustering	12.1, 12.4	
33	W	4/16	Virtual: Project Office Hours		
	F	4/18	***Review***		
	M	4/21	**Midterm #3**		
	W	4/23			
	F	4/25			**Project Due**
			No final exam		

Grade distribution

	Estimated Points
Homeworks	(9 homeworks - 2 lowest grades) $\times$ 20 points = 140
Quizzes	(10 Quizzes - 2 lowest grades) $\times$ 10 points = 80
Midterm	(3 Midterms) $\times$ 100 = 300
Final Project	100
TOTAL:	620 (Subject to change!)

Section 1

Intro to class

What is Statistical Learning?

Statistical Learning

- Subfield of statistics
- Emphasizes models and their interpretability, precision, and uncertainty

Machine Learning

- Machine learning has a greater emphasis on large scale applications and prediction accuracy.

Nowadays....to sound pedantic or techie?

Why should you care?

Data is cheap (or even free), learning how to analyze data is critical.

- Web data, e-commerce (Amazon, JD, Alibaba)
- Car sales (Tesla, Ford, and GM)
- Sports team (MSU, Lions, etc)
- Politics and government

Learning Tools as Black Boxes? Or Math Apocalypse?

- Need to understand the machinery enough to
  - know what tool to use
  - know how to interpret output of the tool
- Don't need to rebuild the entire box from scratch

Example: Email spam

	george	you	your	hp	free	hpl	!	our	re	edu	remove
spam	0.00	2.26	1.38	0.02	0.52	0.01	0.51	0.51	0.13	0.01	0.28
email	1.27	1.27	0.44	0.90	0.07	0.43	0.11	0.18	0.42	0.29	0.01

if (%george < 0.6) & (%you > 1.5) then spam else email.

if $(0.2 \cdot$ %you $- \ 0.3 \cdot$ %george$) > 0$ then spam else email.

Supervised learning

- Outcome measurement Y (also called dependent variable, response, target, label).
- Vector of p predictor measurements X (also called inputs, regressors, covariates, features, independent variables).
- In the regression problem, Y is quantitative (e.g price, blood pressure).
- In the classification problem, Y takes values in a set of distinct categories (survived/died, cancer class of tissue sample, types of language).

Unsupervised learning

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- Objective is fuzzier: often explore the intrinsic relation between samples (e.g.,clustering) or features (e.g. dimensionality reduction)
- Difficult to know how well you are are doing
- Different from supervised learning but can be useful as a pre-processing step for supervised learning.

Generative AI discussion

Definition via Wikipedia:

> *Generative artificial intelligence (AI) is artificial intelligence capable of generating text, images, or other media, using generative models. Generative AI models learn the patterns and structure of their input training data and then generate new data that has similar characteristics.*

Examples:

- ChatGPT
- Bard
- DALL-E

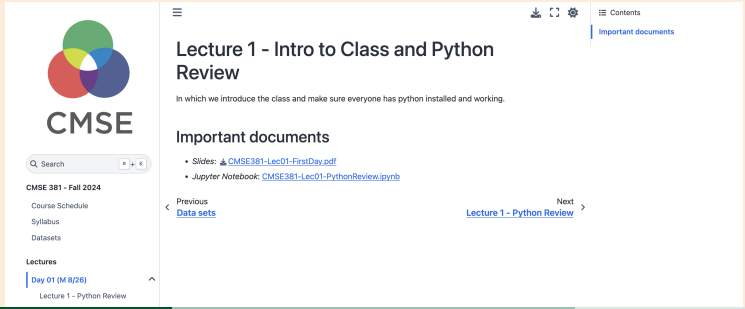
- Get in a group of about 4.
- Open this google doc: tinyurl.com/CMSE381-S25-genAI
- In your group, brainstorm cases where someone might use generative AI in the context of our class.
- Once you have added a few, start adding arguments for or against whether we should allow the use of that context in class.

Section 2

Python Review Lab: Pt 1

Plan for the lab

- Find a group of 4 or so.
- Find the class website (cmse.msu.edu/CMSE381) or (msu-cmse-courses.github.io/CMSE381-S25/) and download the jupyter notebook for the Python Review Lab.
- Get started!



Next time

- Weds: What is statistical learning?
- No class coming Monday (1/20)
- First HW Due Sunday, 9/8
- Quiz sometime **this** week
- Office hours:
  - Maintained on the website
  - Dr. Zhang: Monday and Wednesday 10-11 am
    (Starting next week)
  - Omeiza Olumoye: Tuesdays and Thursdays 3-4 pm

CMSE381_S2025_Schedule : Sheet1

Lec #	Date		Topic	Reading	HW
1	M	1/13	Intro / Python Review	1	
2	W	1/15	What is statistical learning	2.1	
3	F	1/17	Assessing Model Accuracy	2.2.1, 2.2.2	
	M	1/20	MLK - No Class		
4	W	1/22	Linear Regression	3.1	
5	F	1/24	More Linear Regression	3.1	HW #1 Due Sun 1/26
6	M	1/27	Multi-linear Regression	3.2	