

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>

🏠 | SS25-CMSE-381-001 - Fundamentals of Data Scienc...  |  |  |  |  Mengers Zhang 

Course Home Content Course Tools ▾ Assessments ▾ Communication ▾ Help Course Admin More ▾

SS25-CMSE-381-001 - Fundamentals of Data Science Methods

Announcements ▾

There are no announcements to display. [Create an announcement](#)

Need Help? ▾

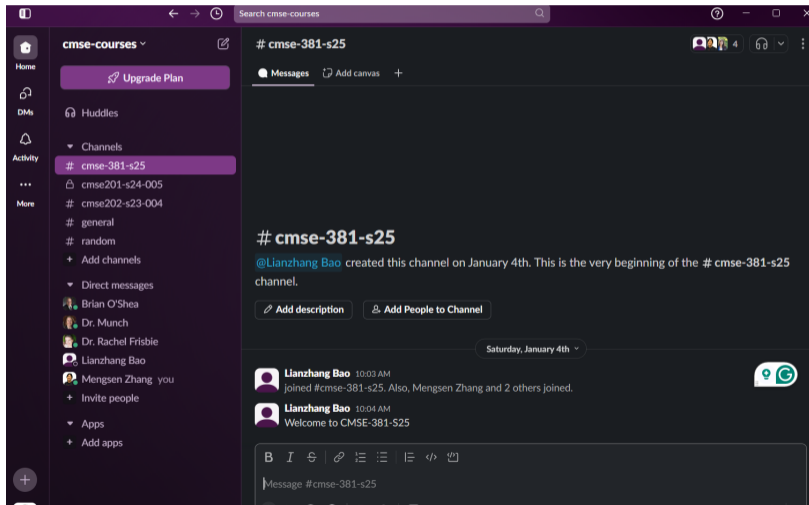
MSU IT Service Desk:

Local: (517) 432-6200

Toll Free: (844) 276-2800

Slack and where to find announcements/ask questions

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

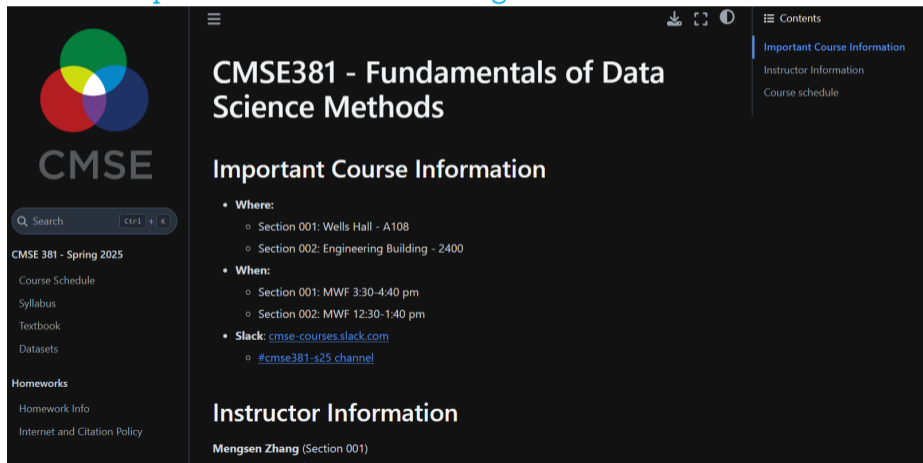


Course Website and where to find slides and jupyter notebooks

<https://cmse.msu.edu/CMSE381>

—or—

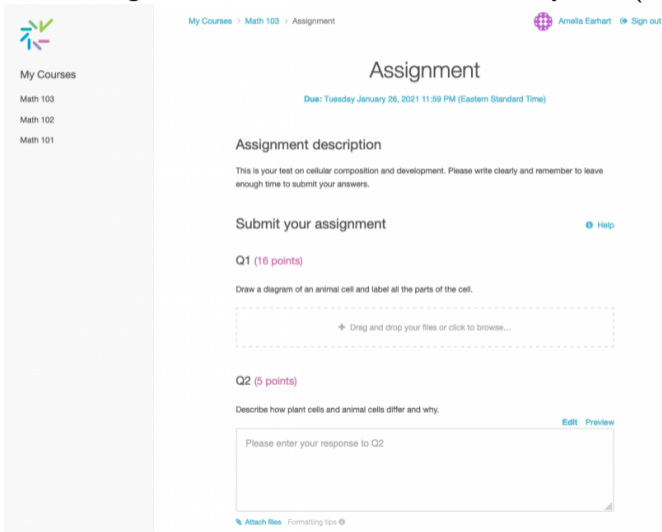
<https://msu-cmse-courses.github.io/CMSE381-F25/>



The screenshot shows the course website for CMSE381 - Fundamentals of Data Science Methods. The page has a dark theme. On the left is a sidebar with a logo consisting of four overlapping circles (green, yellow, red, blue) and the text 'CMSE'. Below the logo is a search bar with 'Search' and 'Ctrl + K' buttons. The sidebar lists navigation links: 'CMSE 381 - Spring 2025', 'Course Schedule', 'Syllabus', 'Textbook', 'Datasets', 'Homeworks', 'Homework Info', and 'Internet and Citation Policy'. The main content area features a hamburger menu icon at the top left, a title 'CMSE381 - Fundamentals of Data Science Methods', and a section titled 'Important Course Information'. This section contains three bullet points: 'Where:' with sub-points for Section 001 (Wells Hall - A108) and Section 002 (Engineering Building - 2400); 'When:' with sub-points for Section 001 (MWF 3:30-4:40 pm) and Section 002 (MWF 12:30-1:40 pm); and 'Slack:' with sub-points for the channel 'cmse-courses.slack.com' and '#cmse381-s25_channel'. Below this is the 'Instructor Information' section, listing 'Mengsen Zhang (Section 001)'. On the right side, there is a 'Contents' menu with links for 'Important Course Information', 'Instructor Information', and 'Course schedule'.

Crowdmark and where to submit homework

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



The screenshot shows a web interface for an assignment. On the left is a sidebar with a logo and a list of courses: 'My Courses', 'Math 103', 'Math 102', and 'Math 101'. The main content area has a breadcrumb trail 'My Courses > Math 103 > Assignment' and a user profile for 'Amelia Earhart' with a 'Sign out' link. The title 'Assignment' is centered, with a due date 'Due: Tuesday January 26, 2021 11:59 PM (Eastern Standard Time)'. Below is the 'Assignment description' section, which states: 'This is your test on cellular composition and development. Please write clearly and remember to leave enough time to submit your answers.' The 'Submit your assignment' section includes a 'Help' link. The first question, 'Q1 (16 points)', asks to 'Draw a diagram of an animal cell and label all the parts of the cell.' Below this is a dashed box containing the text '➔ Drag and drop your files or click to browse...'. The second question, 'Q2 (5 points)', asks to 'Describe how plant cells and animal cells differ and why.' To the right of the question are 'Edit' and 'Preview' links. Below the question is a text input area with the placeholder 'Please enter your response to Q2'. At the bottom left of the input area are links for 'Attach files' and 'Formatting tips'.

Zoom link: Look up on [calendar on the website](#)

The image shows a screenshot of the CMSE website on the left and a Google Calendar on the right. The website sidebar includes the CMSE logo, a search bar, and navigation links for 'CMSE 381 - Spring 2025', 'Course Schedule', 'Syllabus', 'Textbook', 'Datasets', 'Homeworks', 'Homework Info', and 'Internet and Citation Policy'. The Google Calendar is titled 'Google calendar for office hours' and shows a monthly view for January 2025. The calendar grid displays office hours for CMSE381-S2025, with events listed for various dates including 13, 15, 17, 19, 20, 21, 22, 23, 24, 26, 27, 28, 29, 30, and 31. The events are categorized by time slots: 10am Dr. Zhang, 12:30pm CMSE2, 3:30pm CMSE3, 9am Dr. Bao off, and 9am Dr. Bao off. The calendar also shows '9am Dr. Bao off' and '10am Dr. Zhang' on Jan 21 and 22, and '10am Dr. Zhang' and '10am Dr. Zhang' on Jan 29.

Dr. Zhang

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

Zoom & EGR 1514

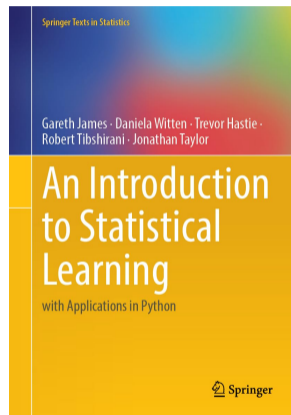
Omeiza Olumoye

Time T-Th 3-4 PM

Zoom 959 4655 2093 (PW: cmse381)
& EGR (Room TBD)

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 related 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	
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	
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		

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
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	
	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
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
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
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	$(9 \text{ homeworks} - 2 \text{ lowest grades}) \times 20 \text{ points} = 140$
Quizzes	$(10 \text{ Quizzes} - 2 \text{ lowest grades}) \times 10 \text{ points} = 80$
Midterm	$(3 \text{ Midterms}) \times 100 = 300$
Final Project	100
<hr/>	
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 · %you - 0.3 · %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!

The screenshot displays the course website for CMSE381. On the left is a sidebar with the CMSE logo (three overlapping circles in green, red, and blue) and the text 'CMSE'. Below the logo is a search bar and a navigation menu for 'CMSE 381 - Fall 2024' with links for 'Course Schedule', 'Syllabus', 'Datasets', and 'Lectures'. The 'Lectures' section is expanded to show 'Day 01 (M 8/26)' and 'Lecture 1 - Python Review'. The main content area is titled 'Lecture 1 - Intro to Class and Python Review' and includes a sub-header 'Important documents' with links to 'CMSE381-Lec01-FirstDay.pdf' and 'CMSE381-Lec01-PythonReview.ipynb'. Navigation arrows for 'Previous Data sets' and 'Next Lecture 1 - Python Review' are visible. A 'Contents' sidebar on the right lists 'Important documents'.

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	