

Intro and First Day Stuff

Lecture 1 - CMSE 381

Prof. Lianzhang Bao

Michigan State University

::

Dept of Computational Mathematics, Science & Engineering

Mon, Jan 13, 2025

People in this lecture



Dr. Bao (he/him)
Dept of CMSE



Christy Lu (she/her)
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/2066703>

🏠 SS25-CMSE-381-002 - Fundamentals of Data Scien...      LB Lianzhang Bao 

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

SS25-CMSE-381-002 - Fundamentals of Data Science Methods

Announcements ▾

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

Updates ▾

There are no current updates for SS25-CMSE-381-002 - Fundamentals of Data Science Methods

Content Browser ▾

 Bookmarks  Recently Visited

There is no content to display. [Create some content](#)

Need Help? ▾

MSU IT Service Desk:

Local: **(517) 432-6200**

Toll Free: **(844) 678-6200**

(North America and Hawaii)

Web:

[D2L Contact Form](#) | [D2L Help Site](#)

[MSU IT Service Status](#) | [Subscribe](#)

Training:

[Educational Technology Training](#)

Calendar ▾

Slack and where to find announcements/ask questions

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

cmse-381-s25



Messages Add canvas +

cmse-381-s25

You created this channel on January 4th. This is the very beginning of the # cmse-381-s25 channel.

Add description

Add People to Channel

Saturday, January 4th



Lianzhang Bao 10:03 AM

joined #cmse-381-s25. Also, Mengsen Zhang and 2 others joined.



Lianzhang Bao 10:04 AM

Welcome to CMSE-381-S25

Course Website and where to find slides and jupyter notebooks

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

—or—

<https://msu-cmse-courses.github.io/CMSE381-S25/>



CMSE

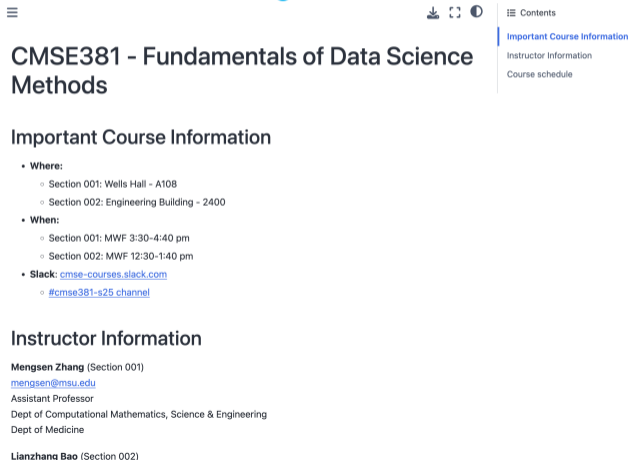
Q Search ✕ + K

CMSE 381 - Spring 2025

- Course Schedule
- Syllabus
- Textbook
- Datasets

Homeworks

- Homework Info
- Internet and Citation Policy



☰ 📄 🔍 🌐 🔊 ☰ Contents

CMSE381 - Fundamentals of Data Science Methods

- Important Course Information
- Instructor Information
- Course schedule

Important Course Information

- **Where:**
 - Section 001: Wells Hall - A108
 - Section 002: Engineering Building - 2400
- **When:**
 - Section 001: MWF 3:30-4:40 pm
 - Section 002: MWF 12:30-1:40 pm
- **Slack:** cmse-courses.slack.com
 - [#cmse381-s25 channel](#)

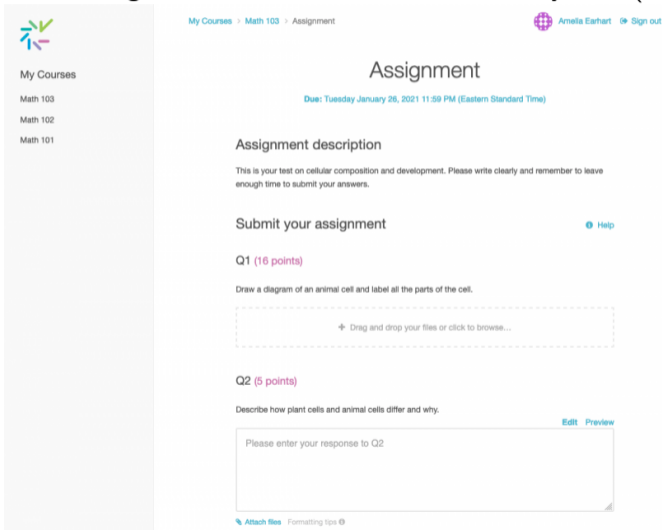
Instructor Information

Mengsen Zhang (Section 001)
mengsen@msu.edu
Assistant Professor
Dept of Computational Mathematics, Science & Engineering
Dept of Medicine

Lianzhang Bao (Section 002)

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 this 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 the question 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 text 'Please enter your response to Q2'. At the bottom left of the input area are links for 'Attach files' and 'Formatting tips'.

Dr. Bao

Tu-W 9am - 10am

Zoom & EGR 2507L

Christy Lu

Time TBD

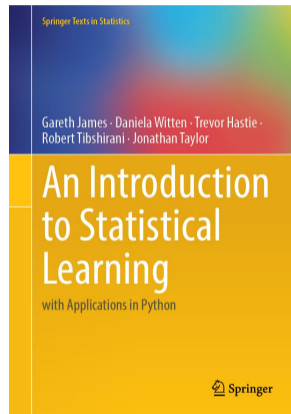
Zoom & EGR (Room TBD)

Details on the calendar posted on the course webpage

<https://msu-cmse-courses.github.io/CMSE381-S25/>

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

Lec #	Date	Topic	Reading	HW	Pop Quizzes	Notes
1	M 1/13	Intro / Python Review	1			
2	W 1/15	What is statistical learning	2.1		Q1	
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		Q2	
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		Q3	
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		Q4	
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		Q5	
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		Q6	
	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		Q7	
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		Q8	
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		Q9	
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		Q10	
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	$(10 \text{ homeworks} - 2 \text{ lowest grades}) \times 20 \text{ points} = 160$
Quizzes	$(12 \text{ Quizzes} - 2 \text{ lowest grades}) \times 10 \text{ points} = 100$
Midterm	$(3 \text{ Midterms}) \times 100 = 300$
Final Project	100
<hr/>	
TOTAL:	660 (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.

Very blurred distinction at this point....

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

- Need to know what tool to use
- Need to 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 finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).

Unsupervised learning

- No outcome variable, just a set of predictors (features) measured on a set of samples.
- Objective is fuzzier: find groups of samples that behave similarly, find features that behave similarly, find linear combinations of features with the most variation.
- 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

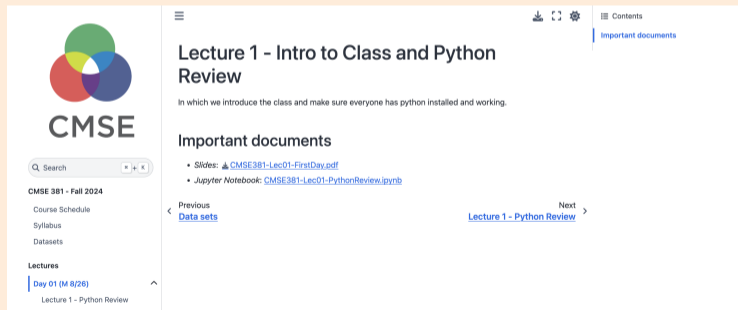
- Get in a group of about 4.
- Open this google doc (MSU Login required): tinyurl.com/CMSE381-genAI
- In your group, brainstorm cases where someone might use generative AI in the context of our class.
- Once you have added a few, start adding arguments for or against whether we should allow the use of 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) and download the jupyter notebook for the Python Review Lab.
- Get started!



The screenshot displays the CMSE 381 website interface. On the left is a navigation sidebar with the CMSE logo (four overlapping circles in green, red, yellow, and blue) and the text "CMSE". Below the logo is a search bar and a menu for "CMSE 381 - Fall 2024" containing links for "Course Schedule", "Syllabus", "Datasets", and "Lectures". Under "Lectures", "Day 01 (M 8/26)" is selected, and "Lecture 1 - Python Review" is listed below it. The main content area on the right features a hamburger menu, download, refresh, and settings icons, and a "Contents" dropdown menu. The page title is "Lecture 1 - Intro to Class and Python Review". Below the title is a paragraph: "In which we introduce the class and make sure everyone has python installed and working." Underneath is a section titled "Important documents" with two bullet points: "Slides: CMSE381-Lec01-FirstDay.pdf" and "Jupyter Notebook: CMSE381-Lec01-PythonReview.ipynb". At the bottom of the main content area are navigation arrows and links for "Previous Data sets" and "Next Lecture 1 - Python Review".

Next time

- Weds: What is statistical learning?
- First HW Due Sunday, 1/26
- Quiz sometime **this** week
- Office hours:
 - ▶ Maintained on the website
 - ▶ Dr. Bao: Tuesday and Wednesday 9-10 (Starting next week)
 - ▶ Christy Lu: Times TBD

Lec #	Date	Topic	Reading	HW	Pop Quizzes	Notes
1	M 1/13	Intro / Python Review	1			
2	W 1/15	What is statistical learning	2.1		Q1	
3	F 1/17	Assessing Model Accuracy	2.2.1, 2.2.2			
	M 1/20	MLK - No Class				
	W 1/22	Linear Regression	2.3		Q2	