Intro and First Day Stuff

Lecture 1 - CMSE 381

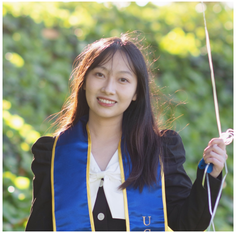
Prof. Elizabeth Munch

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

Mon, Aug 26, 2024

People in this lecture

Dr. Munch (she/her)
Depts of CMSE and Math



Christy Lu (she/her)
Graduate Student, CMSE, MSU

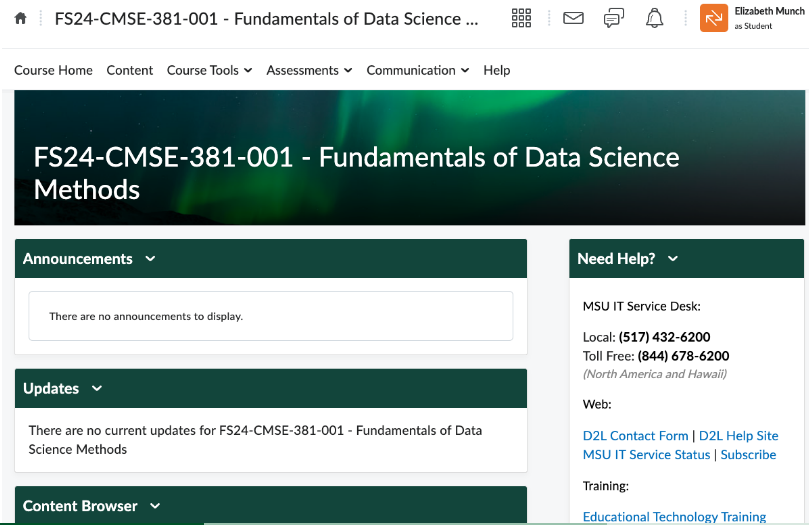
What is this course about?

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



FS24-CMSE-381-001 - Fundamentals of Data Science ...

Elizabeth Munch
as Student

FS24-CMSE-381-001 - Fundamentals of Data Science Methods

Announcements

There are no announcements to display.

Updates

There are no current updates for FS24-CMSE-381-001 - Fundamentals of Data Science Methods

Content Browser

Need Help?

MSU IT Service Desk:

Local: **(517) 432-6200**
Toll Free: **(844) 678-6200**
(North America and Hawaii)

Web:

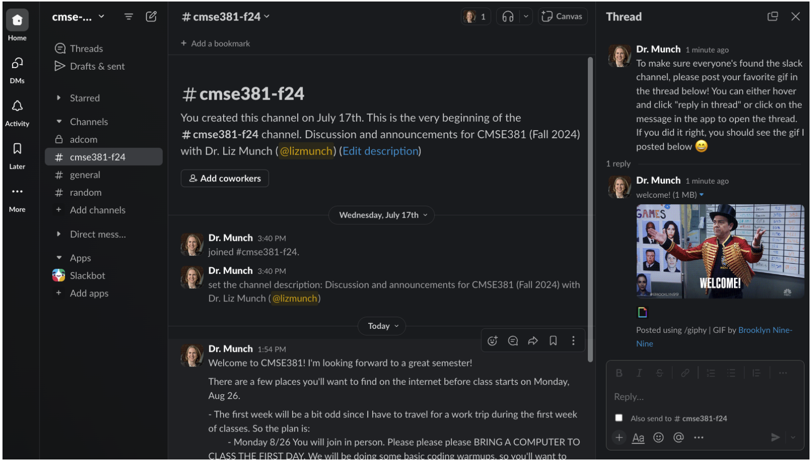
D2L Contact Form | D2L Help Site
MSU IT Service Status | Subscribe

Training:

Educational Technology Training

Slack and where to find announcements/ask questions

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



cmse381-f24

You created this channel on July 17th. This is the very beginning of the # cmse381-f24 channel. Discussion and announcements for CMSE381 (Fall 2024) with Dr. Liz Munch (@lizmunch) (Edit description)

Dr. Munch 3:40 PM
joined #cmse381-f24.

Dr. Munch 3:40 PM
set the channel description: Discussion and announcements for CMSE381 (Fall 2024) with Dr. Liz Munch (@lizmunch)

Dr. Munch 1:54 PM
Welcome to CMSE381! I'm looking forward to a great semester!

There are a few places you'll want to find on the internet before class starts on Monday, Aug 26.

- The first week will be a bit odd since I have to travel for a work trip during the first week of classes. So the plan is:

    - Monday 8/26 You will join in person. Please please please BRING A COMPUTER TO CLASS THE FIRST DAY. We will be doing some basic coding warmups, so you'll want to

Thread

Dr. Munch 1 minute ago
To make sure everyone's found the slack channel, please post your favorite gif in the thread below! You can either hover and click "reply in thread" or click on the message in the app to open the thread. If you did it right, you should see the gif I posted below 😁

1 reply

Dr. Munch 1 minute ago
welcome! (1 MB)

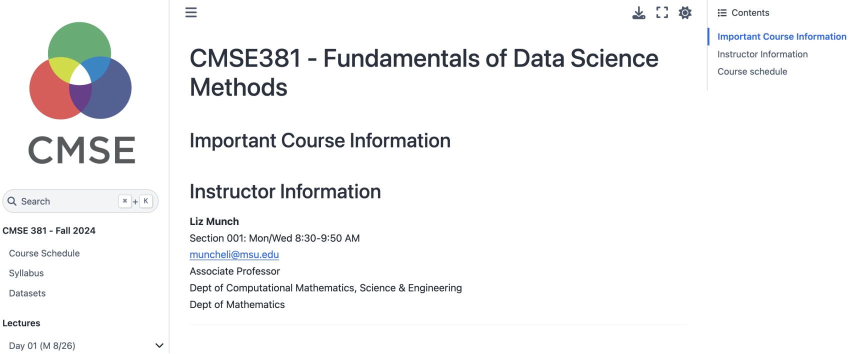
Posted using /giphy | GIF by Brooklyn Nine-Nine

Course Website and where to find slides and jupyter notebooks

https://cmse.msu.edu/CMSE381

–or–

https://msu-cmse-courses.github.io/CMSE381-F24/



CMSE381 - Fundamentals of Data Science Methods

Important Course Information

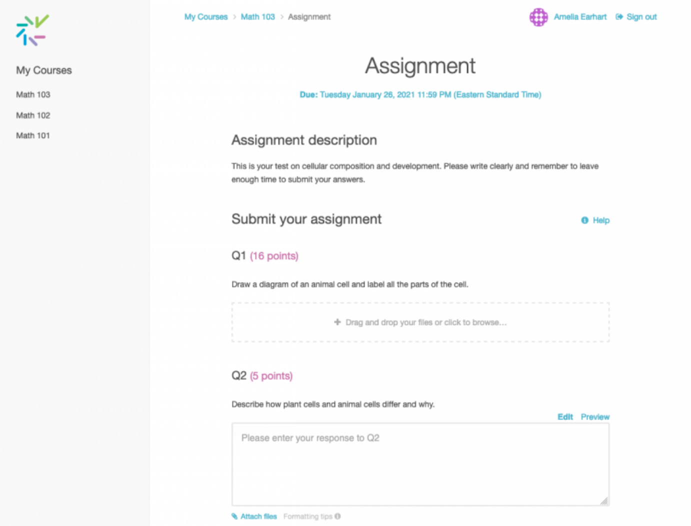
Instructor Information

Liz Munch
Section 001: Mon/Wed 8:30-9:50 AM
muncheli@msu.edu
Associate Professor
Dept of Computational Mathematics, Science & Engineering
Dept of Mathematics

Note the syllabus link above!

Crowdmark and where to submit homework

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



My Courses > Math 103 > Assignment
Amelia Earhart
Sign out
My Courses
Math 103
Math 102
Math 101

Assignment

Due: Tuesday January 26, 2021 11:59 PM (Eastern Standard Time)

Assignment description

This is your test on cellular composition and development. Please write clearly and remember to leave enough time to submit your answers.

Submit your assignment

Q1 (16 points)

Draw a diagram of an animal cell and label all the parts of the cell.

+ Drag and drop your files or click to browse...

Q2 (5 points)

Describe how plant cells and animal cells differ and why.

Office hours

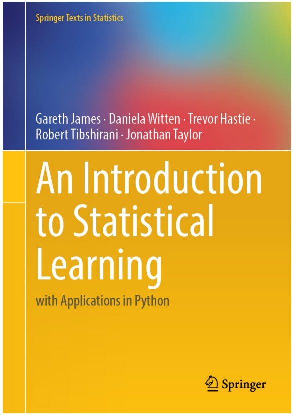
Zoom link: `https://bit.ly/3FTuRqG`

Dr. Munch	*Christy Lu*
Time TBD (Starting next week)	Time TBD
Zoom & EGR 1511	Zoom & EGR (Room TBD)

Textbook

Free download

https://www.statlearning.com/


Springer Texts in Statistics
Gareth James · Daniela Witten · Trevor Hastie · Robert Tibshirani · Jonathan Taylor
An Introduction to Statistical Learning
with Applications in Python
Springer

Class Structure

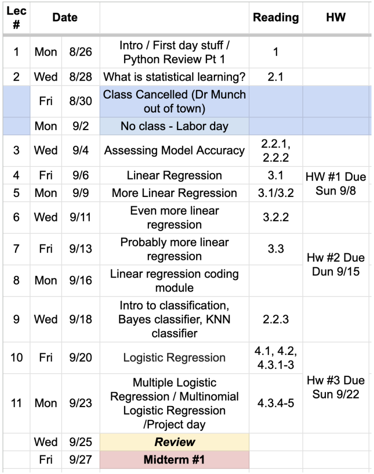
- Class is a combination of lecture time, and group work/coding time.
  - Bring computer every day
  - Jupyter notebooks
  - Python
- Once a week, there will be a short check-in quiz. This will be basic content realted to lectures since the last class. Possible questions include checking on definitions,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

Class Structure Pt 2

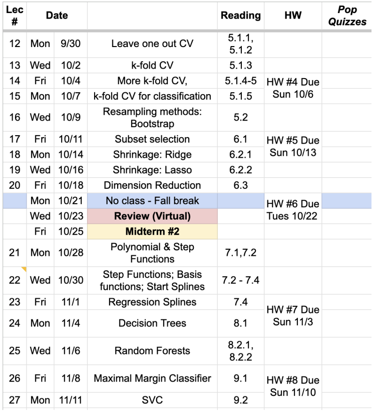
- Homeworks due once a week, midnight of the day marked in the schedule (mostly Sundays).
    - 20 points per homework
    - Drop two lowest grades
    - Sliding scale:
        - 24 hours late: 5% penalty.
        - 48 hours late: 15% penalty.
        - $>$48 hours: No late work accepted.
- Three Midterms
    - 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

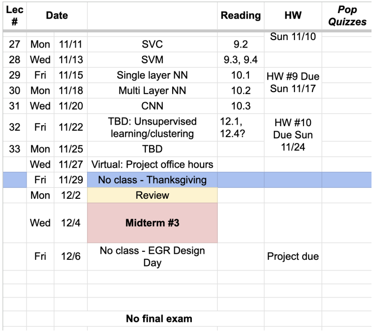
Up to date version: https://msu-cmse-courses.github.io/CMSE381-F24/Course_Info/Schedule.html



Lec #	Date			Reading	HW
1	Mon	8/26	Intro / First day stuff / Python Review Pt 1	1	
2	Wed	8/28	What is statistical learning?	2.1	
	Fri	8/30	Class Cancelled (Dr Munch out of town)		
	Mon	9/2	No class - Labor day		
3	Wed	9/4	Assessing Model Accuracy	2.2.1, 2.2.2	
4	Fri	9/6	Linear Regression	3.1	HW #1 Due Sun 9/8
5	Mon	9/9	More Linear Regression	3.1/3.2	
6	Wed	9/11	Even more linear regression	3.2.2	
7	Fri	9/13	Probably more linear regression	3.3	Hw #2 Due Dun 9/15
8	Mon	9/16	Linear regression coding module		
9	Wed	9/18	Intro to classification, Bayes classifier, KNN classifier	2.2.3	
10	Fri	9/20	Logistic Regression	4.1, 4.2, 4.3.1-3	Hw #3 Due Sun 9/22
11	Mon	9/23	Multiple Logistic Regression / Multinomial Logistic Regression /Project day	4.3.4-5	
	Wed	9/25	***Review***		
	Fri	9/27	**Midterm #1**		



Lec #	Date			Reading	HW	*Pop Quizzes*
12	Mon	9/30	Leave one out CV	5.1.1, 5.1.2		
13	Wed	10/2	k-fold CV	5.1.3		
14	Fri	10/4	More k-fold CV,	5.1.4-5	HW #4 Due Sun 10/6	
15	Mon	10/7	k-fold CV for classification	5.1.5		
16	Wed	10/9	Resampling methods: Bootstrap	5.2		
17	Fri	10/11	Subset selection	6.1	HW #5 Due Sun 10/13	
18	Mon	10/14	Shrinkage: Ridge	6.2.1		
19	Wed	10/16	Shrinkage: Lasso	6.2.2		
20	Fri	10/18	Dimension Reduction	6.3		
	Mon	10/21	No class - Fall break		HW #6 Due Tues 10/22	
	Wed	10/23	**Review (Virtual)**			
	Fri	10/25	**Midterm #2**			
21	Mon	10/28	Polynomial & Step Functions	7.1,7.2		
22	Wed	10/30	Step Functions; Basis functions; Start Splines	7.2 - 7.4		
23	Fri	11/1	Regression Splines	7.4	HW #7 Due Sun 11/3	
24	Mon	11/4	Decision Trees	8.1		
25	Wed	11/6	Random Forests	8.2.1, 8.2.2		
26	Fri	11/8	Maximal Margin Classifier	9.1	HW #8 Due Sun 11/10	
27	Mon	11/11	SVC	9.2		



Lec #	Date			Reading	HW	*Pop Quizzes*
27	Mon	11/11	SVC	9.2	Sun 11/10	
28	Wed	11/13	SVM	9.3, 9.4		
29	Fri	11/15	Single layer NN	10.1	HW #9 Due Sun 11/17	
30	Mon	11/18	Multi Layer NN	10.2		
31	Wed	11/20	CNN	10.3		
32	Fri	11/22	TBD: Unsupervised learning/clustering	12.1, 12.4?	HW #10 Due Sun 11/24	
33	Mon	11/25	TBD			
	Wed	11/27	Virtual: Project office hours			
	Fri	11/29	No class - Thanksgiving			
	Mon	12/2	Review			
	Wed	12/4	**Midterm #3**			
	Fri	12/6	No class - EGR Design Day		Project due	
			No final exam			

Grade distribution

	Estimated Points
Homeworks	(10 homeworks - 2 lowest grades) × 20 points = 160
Quizzes	(12 Quizzes - 2 lowest grades) × 10 points = 100
Midterm	(3 Midterms) × 100 = 300
Final Project	100
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?

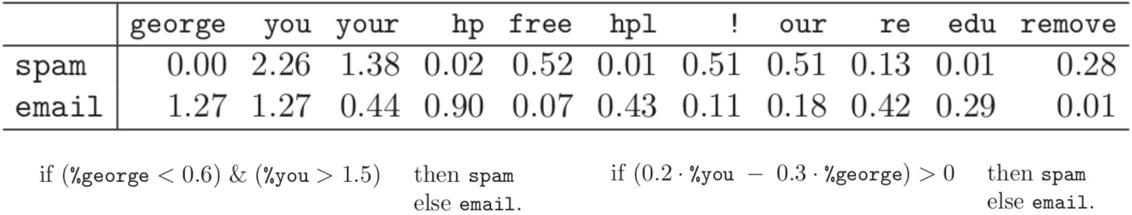
Data is cheap (or even free), learning how to analyze data is critical.

- 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 $(\texttt{\%george} < 0.6)$ & $(\texttt{\%you} > 1.5)$ then $\texttt{spam}$
else $\texttt{email}$.

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

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

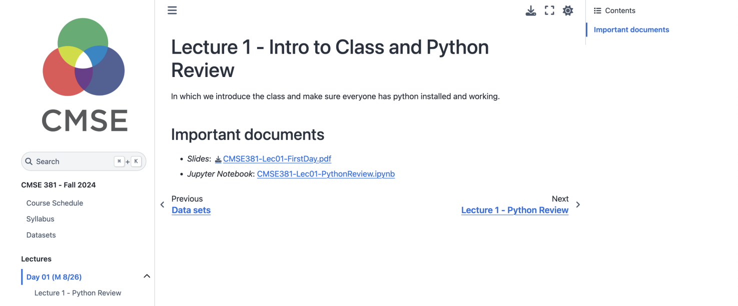
- Get in a group of about 4.
- Open this google doc (MSU Login required): tinyurl.com/CMSE381-genAI
- In your group, brainstorm cases where someone might use generative AI in the context of our class.
- Once you have added a few, start adding arguments for or against whether we should allow the use of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



Lecture 1 - Intro to Class and Python Review

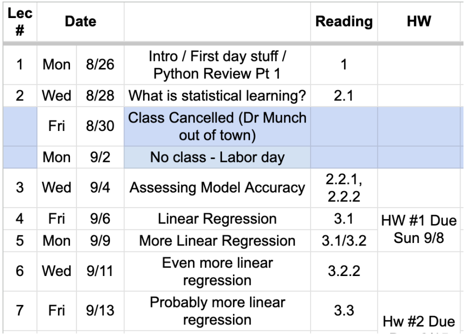
In which we introduce the class and make sure everyone has python installed and working.

Important documents

- *Slides*: CMSE381-Lec01-FirstDay.pdf
- *Jupyter Notebook*: CMSE381-Lec01-PythonReview.ipynb

Next time

- Weds: What is statistical learning?
- First HW Due Sunday, 9/8
- Quiz sometime **this** week
- Office hours:
  - Maintained on the website
  - Dr. Munch: Monday and Friday 11-12 (Starting next week)
  - Christy Lu: Times TBD



Lec #	Date			Reading	HW
1	Mon	8/26	Intro / First day stuff / Python Review Pt 1	1	
2	Wed	8/28	What is statistical learning?	2.1	
	Fri	8/30	Class Cancelled (Dr Munch out of town)		
	Mon	9/2	No class - Labor day		
3	Wed	9/4	Assessing Model Accuracy	2.2.1, 2.2.2	
4	Fri	9/6	Linear Regression	3.1	HW #1 Due Sun 9/8
5	Mon	9/9	More Linear Regression	3.1/3.2	
6	Wed	9/11	Even more linear regression	3.2.2	
7	Fri	9/13	Probably more linear regression	3.3	Hw #2 Due