

Intro and First Day Stuff

Lecture 1 - CMSE 381

Prof. Elizabeth Munch

Michigan State University

::

Dept of Computational Mathematics, Science & Engineering

Mon, Aug 28, 2023

People in this lecture



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



Rachel Roca (she/they)
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/1811231>


 FS23-CMSE-381 - Fundamentals of Data Science Meth... Elizabeth Munch
as Student

Course Home Content Course Tools ▾ Assessments ▾ Communication ▾ Help

FS23-CMSE-381 - Fundamentals of Data Science Methods

Announcements ▾

Welcome! and where to find stuff ×

Elizabeth Munch posted on Aug 16, 2023 2:15 PM •  Edited

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 28.

- I only use D2L for grade communication. So, you should have access to the D2L page as Rachel Roca, our TA, will be posting the grades there.
- Course material will be posted on the github page here:
<https://github.com/msu-cmse-courses/cmse381-F23/>. They're not quite up yet, but for the first day, you will find the slides I will use, as well as a jupyter

Need Help? ▾

MSU IT Service Desk:

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

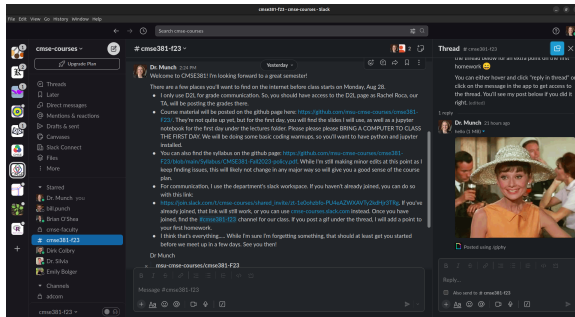
Web:

[D2L Contact Form](#) | [D2L Help Site](#)
[MSU IT Service Status](#) | [Subscribe](#)

Training:

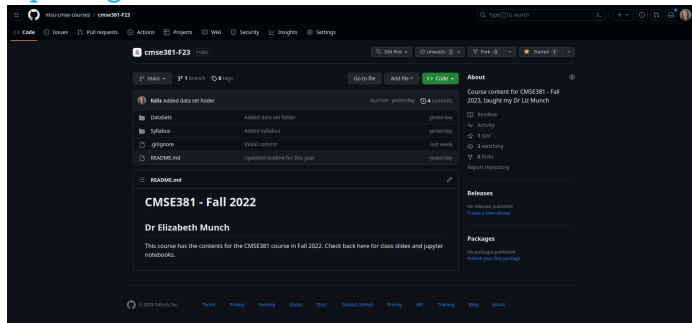
[Educational Technology Training](#)

Slack and where to find announcements/ask questions



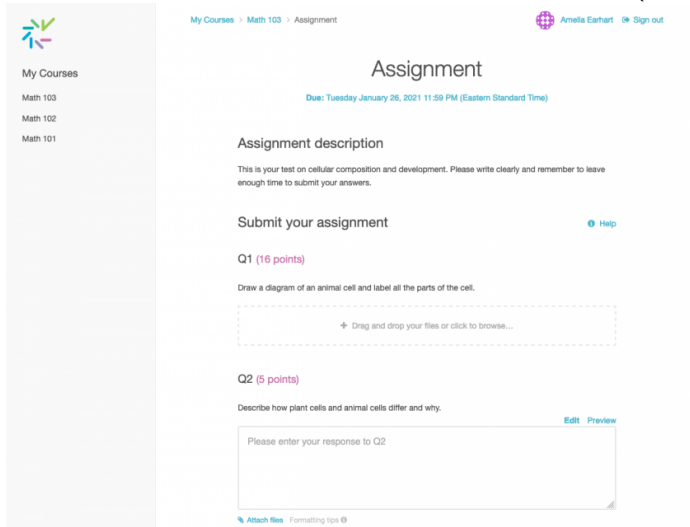
Github and where to find slides and jupyter notebooks

<https://github.com/msu-cmse-courses/cmse381-F23/>



Crowdmark and where to submit homework

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



The screenshot shows the Crowdmark assignment submission page. On the left is a sidebar with the Crowdmark logo and a 'My Courses' section listing '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, followed by the 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.' It features a dashed box with a plus icon and 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.' It includes 'Edit' and 'Preview' links and a text input area with the placeholder 'Please enter your response to Q2'. At the bottom, there are links for 'Attach files' and 'Formatting tips'.

Zoom link: <https://bit.ly/3FTuRqG>

Dr. Munch

Wednesdays and Thursdays

Both 11am - Noon

Zoom & EGR 1511

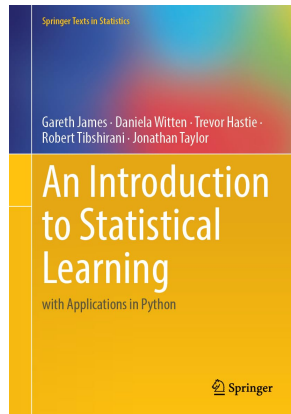
Rachel Roca

Tuesdays 3:30 - 5pm,
and Fridays 11:30 - 1pm

Zoom & 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.
 - ▶ 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

Up to date version: <https://tinyurl.com/CMSE381-FS23-Schedule>

Up to date version: <https://tinyurl.com/CMSE381-FS23-Schedule>

Lec #	Date		Reading
	Mon Oct 23	No class - Fall break	
	Wed Oct 25	Midterm #2	
20	Fri Oct 27	Dimension Reduction	6.3
21	Mon Oct 30	More dimension reduction; High dimensions	6.4
22	Wed Nov 1	Polynomial & Step Functions	7.1, 7.2
23	Fri Nov 3	Step Functions	7.2
24	Mon Nov 6	Basis functions, Regression Splines	7.3, 7.4
25	Wed Nov 8	Decision Trees	8.1
26	Fri Nov 10	Random Forests	8.2.1, 8.2.2
27	Mon Nov 13	Maximal Margin Classifier	9.1
28	Wed Nov 15	SVC	9.2
29	Fri Nov 17	SVM	9.3, 9.4
30	Mon Nov 20	Single layer NN	10.1
31	Wed Nov 22	Overflow/project day?	
	Fri Nov 24	No class - Thanksgiving	
	Mon Nov 27	Review	
	Wed Nov 29	Midterm #3	
32	Fri Dec 1	Multi Layer NN	10.2
33	Mon Dec 4	CNN	10.3
34	Wed Dec 6	Unsupervised Learning & Clustering	12.1, 12.4
35	Fri Dec 8	Overflow/Project day?	
		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
TOTAL:	660

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 ($0.2 \cdot \%you - 0.3 \cdot \%george$) > 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 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 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.
- Clone the class repository (or download the jupyter notebook and the csv file from github)
- Get started!

Using git

- `git clone`
`git@github.com:msu-cmse-courses/cmse381-F23.git`
- from inside the folder you just made, run `git pull` any time you want to download new content

Next time

- Weds: What is statistical learning?
- Homework due Friday
- Quiz sometime this week
- Office hours:
 - ▶ Dr. Munch: Weds and Thurs 11-12
(At least this week, subject to change)
 - ▶ Rachel Roca: Tues 3:30 - 5pm and Fri 11:30 - 1pm

Lec #	Date			Reading
1	Mon	Aug 28	Intro / First day stuff / Python Review Pt 1	1
2	Wed	Aug 30	What is statistical learning?	2.1
	Fri	Sep 1	Assessing Model Accuracy	2.2.1, 2.2.2
3	Mon	Sep 4	No class - Labor day	
4	Wed	Sep 6	Linear Regression	3.1
5	Fri	Sep 8	More Linear Regression	3.1/3.2
6	Mon	Sep 11	Even more linear regression	3.2.2
7	Wed	Sep 13	Probably more linear regression	3.3
8	Fri	Sep 15	Intro to classification, Logistic Regression	2.2.3, 4.1, 4.2, 4.3
9	Mon	Sep 18	More logistic regression	
10	Wed	Sep 20	Multiple Logistic Regression / Multinomial Logistic Regression	
11	Fri	Sep 22	Overflow/Project day?	
	Mon	Sep 25	Review	
	Wed	Sep 27	Midterm #1	
	Fri	Sep 29	No class - Dr Munch out of town	
12	Mon	Oct 2	Leave one out CV	5.1.1, 5.1.2
13	Wed	Oct 4	k-fold CV	5.1.3
14	Fri	Oct 6	More k-fold CV,	5.1.4-5
15	Mon	Oct 9	k-fold CV for classification	5.1.5
16	Wed	Oct 11	Resampling methods: Bootstrap	5.2
17	Fri	Oct 13	Subset selection	6.1
18	Mon	Oct 16	Shrinkage: Ridge	6.2.1