

Intro and First Day Stuff

Lecture 1 - CMSE 381

Prof. Rongrong Wang

Michigan State University

::

Dept of Computational Mathematics, Science & Engineering

January 7, 2024

People in this lecture

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

Maryclare Martin (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/1871821>

🏠 ... SS24-CMSE-381-001 - Fundamentals of Data Scienc...  ...    ...

Course Home Content Course Tools ▾ Assessments ▾ Communication ▾ Help Course A

SS24-CMSE-381-001 - Fundamentals of Data Science Methods

Announcements ▾

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

Updates ▾

There are no current updates for SS24-CMSE-381-001 - Fundamentals of Data Science Methods

Need Help?

MSU IT Service

Local: **(517) 43**

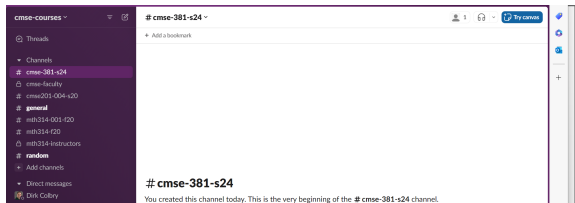
Toll Free: **(844)**
(North America)

Web:

[D2L Contact F](#)

[MSU IT Service](#)

Slack and where to find announcements/ask questions



Expected answer time from the instructors: within 48 hours

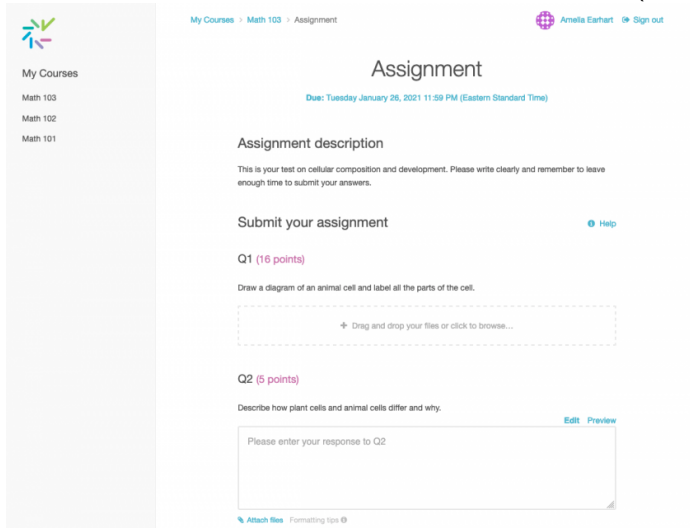
Github and where to find slides and jupyter notebooks

<https://github.com/rrwng/CMSE381SS24/>

The screenshot shows the GitHub interface for the repository `rrwng/CMSE381SS24`. At the top, there's a navigation bar with links for Code, Issues, Pull requests, Actions, Projects, Wiki, Security, Insights, and Settings. Below this, the repository name `CMSE381SS24` is displayed as public, with buttons for Pin, Unwatch (1), Fork (0), and Star (0). The main content area shows the `main` branch with 1 branch and 0 tags. A search bar for files is present, along with buttons for 'Add file' and 'Code'. The commit history shows a recent commit by `rrwng` titled 'Update README.md' from 3 days ago, with 2 commits. The README file is selected, showing the title 'CMSE 381 Spring 2024' and the text 'This the course website for CMSE318 SS24'. On the right sidebar, there are sections for 'About' (no description), 'Releases' (no releases published), and 'Packages' (no packages published).

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, 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.' 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 Meeting ID: 91303872529, Passcode: CMSE381

Dr. Wang

Wednesdays
Both 5pm - 7pm

Zoom

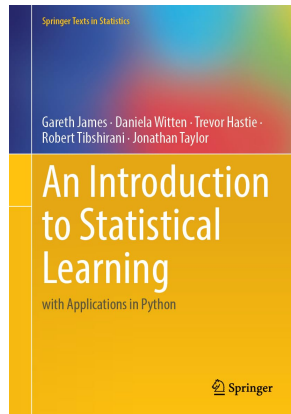
Maryclare Martin

Tuesdays and Thursday 12:30 -
2:30pm,

Zoom (Meeting ID: 91303872529,
Passcode: CMSE381)

Free download

<https://www.statlearning.com/>



Class Structure

- Monday and Wednesday: lecture by Dr. Wang; Friday: recitation by TA and LA.
- Class is a combination of lecture time, and group work/coding time.
 - ▶ Bring computer every day
 - ▶ Jupyter notebooks
 - ▶ Python
- Every Friday, 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.
 - ▶ Drop three lowest grades

Class Structure Pt 2

- Homeworks due once a week, midnight of the day marked in the schedule.
 - ▶ Drop two lowest grades
 - ▶ Sliding scale:
 - ★ 24 hours late: 25% penalty.
 - ★ 48 hours late: 50% penalty.
 - ★ >48 hours: No late work accepted.
- In-class assignments due every Friday midnight.
 - ▶ drop two lowest grades
 - ▶ to be completed during the recitation
 - ▶ full credit for answering 2/3 of the questions correctly
- Two Midterms
 - ▶ See schedule for dates
 - ▶ Not cumulative
- Final exam

Approximate schedule

Schedule will be updated throughout the semester

Grade distribution

Estimated Points

Homeworks	$(11 \text{ homeworks} - 2 \text{ lowest grades}) \times 10 \text{ points} = 90$
Quizzes	$(11 \text{ Quizzes} - 3 \text{ lowest grades}) \times 3 \text{ points} = 24$
In-class assignments	$(12 - 2 \text{ lowest grades}) \times 2 \text{ points} = 20$
Midterm	$(2 \text{ Midterms}) \times 30 = 60$
Final exam	36
TOTAL:	230

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 abundant and powerful,
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 independent measurements X (also called inputs, predictor, regressors, covariates, features, independent variables).
- In the regression problem, Y is quantitative (e.g price, blood pressure).
- In the classification problem, Y categorical, i.e., 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.

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 https://github.com/rrwng/CMSE381SS24.git`
- from inside the folder you just made, run `git pull` any time you want to download new content

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-genAI
- In your group, brainstorm cases where someone might use generative AI in the context of our class.
- Download a copy of this google doc file and add your ideas to it, also add arguments for or against whether we should allow the use of that context in class.
- turn the file in along with the completed python-review notebook.

- Weds: What is statistical learning and model accuracy?
- Homework due next Wednesday
- no Quiz this week
- Office hours:
 - ▶ Dr. Wang: Wednesday 5-7pm
 - ▶ Maryclare: Tues and Thur 12:30pm - 2:30pm