# Intro and First Day Stuff

## Lecture 1 - CMSE 381

Prof. Elizabeth Munch

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

Weds, Aug 31, 2022

# People in this lecture

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

**Emily Bolger** (she/her)
Graduate Student, CMSE, MSU

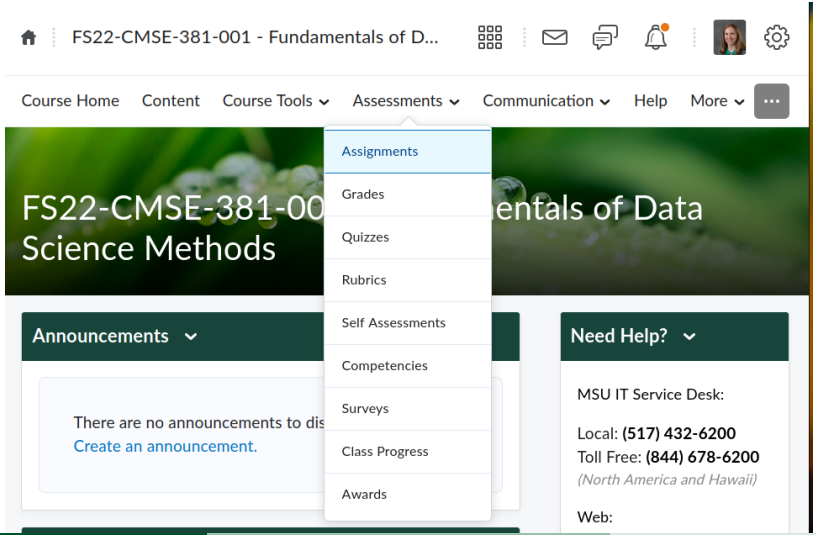
# What is this course about?

Topics:

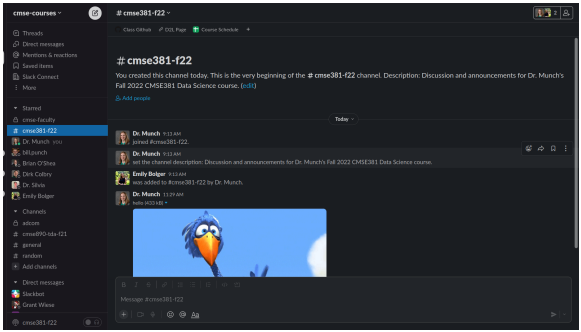
- Fundamental concepts of data science
- Regression
- Classification
- Dimension reduction
- Resampling methods
- Tree-based methods, etc.

# D2L and where to find grades

https://d2l.msu.edu/d2l/home/1579786

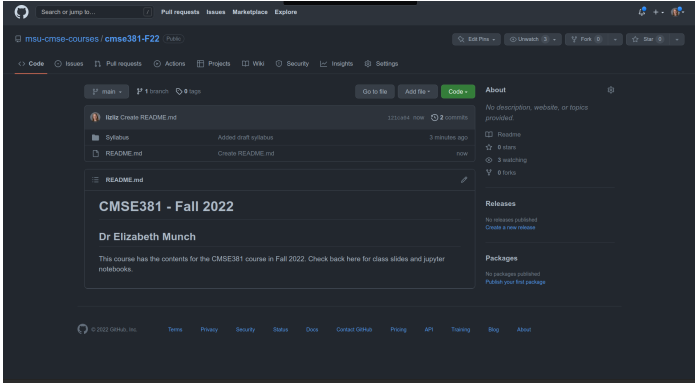


# Slack and where to find announcements/ask questions



# Github and where to find slides and jupyter notebooks

https://github.com/msu-cmse-courses/cmse381-F22/



# Office hours

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

*Dr. Munch*

Time TBD

Zoom & EGR 1511

*Emily Bolger*

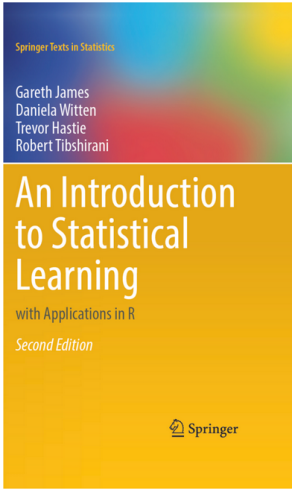
Wed 10:30-noon
Fri 1-2:30

Zoom & EGR 1508A

# Textbook

**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 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.
  - 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

# Approximate schedule

Up to date version: https://bit.ly/3SE0Phl

| Lec # | Date | Topic | Reading | Homeworks |
|---|---|---|---|---|
| 1 | W Aug 31 | Intro / First day stuff / Python Review Pt 1 | 1 | |
| 2 | F Sep 2 | What is statistical learning? / Python Review Pt 2 | 2.1 | |
| | M Sep 5 | No class - Labor day | | |
| 3 | W Sep 7 | Assessing Model Accuracy | 2.2 | HW #1 Due |
| 4 | F Sep 9 | Linear Regression | 3.1 | |
| 5 | M Sep 12 | More Linear Regression | 3.2 | |
| 6 | W Sep 14 | Even more linear regression | 3.3 | HW #2 Due |
| 7 | F Sep 16 | Probably more linear regression | | |
| 8 | M Sep 19 | Intro to classification, Logisitic Regression | 4.1, 4.2, 4.3 | |
| 9 | W Sep 21 | More logistic regression | | HW #3 Due |
| 10 | F Sep 23 | Review | | |
| 11 | M Sep 26 | ***Midterm #1*** | | |
| 12 | W Sep 28 | [No class, Dr Munch out of town] | | |
| 13 | F Sep 30 | [No class, Dr Munch out of town] | | |
| 14 | M Oct 3 | Leave one out CV | 5.1.1, 5.1.2 | |
| 15 | W Oct 5 | k-fold CV | 5.1.3 | |
| 16 | F Oct 7 | More k-fold CV | 5.1.4 | |
| 17 | M Oct 10 | CV for classification | 5.1.5 | HW #4 Due |
| 18 | W Oct 12 | Resampling methods: Bootstrap | 5.2 | |
| 19 | F Oct 14 | Subset selection | 6.1 | |
| 20 | M Oct 17 | Shrinkage: Ridge | 6.2.1 | HW #5 Due |
| 21 | W Oct 19 | Shrinkage: Lasso | 6.2.2 | |
| 22 | F Oct 21 | Dimension Reduction | 6.3 | |

| Lec # | Date | Topic | Reading | Homeworks |
|---|---|---|---|---|
| | M Oct 24 | No class - Fall break | | |
| 21 | W Oct 26 | More dimension reduction; High dimensions | 6.4 | |
| 22 | F Oct 28 | Polynomial & Step Functions. | 7.1,7.2 | HW #6 Due |
| 23 | M Oct 31 | Review | | |
| 24 | W Nov 2 | ***Midterm #2*** | | |
| 25 | F Nov 4 | Basis functions, Regression Splines | 7.3,7.4 | |
| 26 | M Nov 7 | Smoothing Splines; Local regression; GAMs | 7.5-7.7 | |
| 27 | W Nov 9 | Decision Trees | 8.1 | |
| 28 | F Nov 11 | Ensemble methods | 8.2 | HW #7 Due |
| 29 | M Nov 14 | Maximal Margin Classifier | 9.1 | |
| 30 | W Nov 16 | SVC | 9.2 | |
| 31 | F Nov 18 | SVM | 9.3, 9.4, 9.5 | HW #8 Due |
| 32 | M Nov 21 | More SVM | | |
| 33 | W Nov 23 | Single layer NN | 10.1 | |
| | F Nov 25 | No class - Thanksgiving | | |
| 35 | M Nov 28 | Multi Layer NN | 10.2 | HW #9 Due |
| 36 | W Nov 30 | CNN | 10.3 | |
| 37 | F Dec 2 | Unsupervised Learning & Clustering | 12.1, 12.4 | |
| 38 | M Dec 5 | More Clustering | 12.4 | HW #10 Due |
| 39 | W Dec 7 | Review | | |
| 40 | F Dec 9 | Midterm #3 | | |

# Grade distribution

| | *Estimated Points* |
|---|---|
| Homeworks | (10 homeworks - 2 lowest grades) $\times$ 20 points = 160 |
| Quizzes | (12 Quizzes - 2 lowest grades) $\times$ 10 points = 100 |
| Midterm | (3 Midterms) $\times$ 100 = 300 |
| TOTAL: | 560 |

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 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.
- Download the jupyter notebook and the csv file from github.
- Get started!

# Next time

- Friday: What is statistical learning?
- No homework or quiz until next week

| Lec # | Date | Topic | Reading | Homeworks |
|---|---|---|---|---|
| 1 | W Aug 31 | Intro / First day stuff / Python Review Pt 1 | 1 | |
| 2 | F Sep 2 | What is statistical learning? / Python Review Pt 2 | 2.1 | |
| | M Sep 5 | No class - Labor day | | |
| 3 | W Sep 7 | Assessing Model Accuracy | 2.2 | HW #1 Due |
| 4 | F Sep 9 | Linear Regression | 3.1 | |
| 5 | M Sep 12 | More Linear Regression | 3.2 | |
| 6 | W Sep 14 | Even more linear regression | 3.3 | HW #2 Due |
| 7 | F Sep 16 | Probably more linear regression | | |
| 8 | M Sep 19 | Intro to classification, Logisitic Regression | 4.1, 4.2, 4.3 | |
| 9 | W Sep 21 | More logistic regression | | HW #3 Due |
| 10 | F Sep 23 | Review | | |
| 11 | M Sep 26 | ***Midterm #1*** | | |
| 12 | W Sep 28 | [No class, Dr Munch out of town] | | |
| 13 | F Sep 30 | [No class, Dr Munch out of town] | | |
| 14 | M Oct 3 | Leave one out CV | 5.1.1, 5.1.2 | |
| 15 | W Oct 5 | k-fold CV | 5.1.3 | |
| 16 | F Oct 7 | More k-fold CV | 5.1.4 | |
| 17 | M Oct 10 | CV for classification | 5.1.5 | HW #4 Due |
| 18 | W Oct 12 | Resampling methods: Bootstrap | 5.2 | |
| 19 | F Oct 14 | Subset selection | 6.1 | |
| 20 | M Oct 17 | Shrinkage: Ridge | 6.2.1 | HW #5 Due |
| 21 | W Oct 19 | Shrinkage: Lasso | 6.2.2 | |
| 22 | F Oct 21 | Dimension Reduction | 6.3 | |