

Workspace 'Workspace' in 'Intro'

Page 1 (row 1, column 1)



(X)

‘

CMSE 381: Fundamentals of Data Science Methods

Jan 9th , 2023



SPARTANS WILL.

About me



→ YuYing Xie

→ Xie
YuYing

→ XYY

← Clark ← Chr X Chr Y Y

xyy@msu.edu ·

xx@msu.edu



SPARTANS WILL.

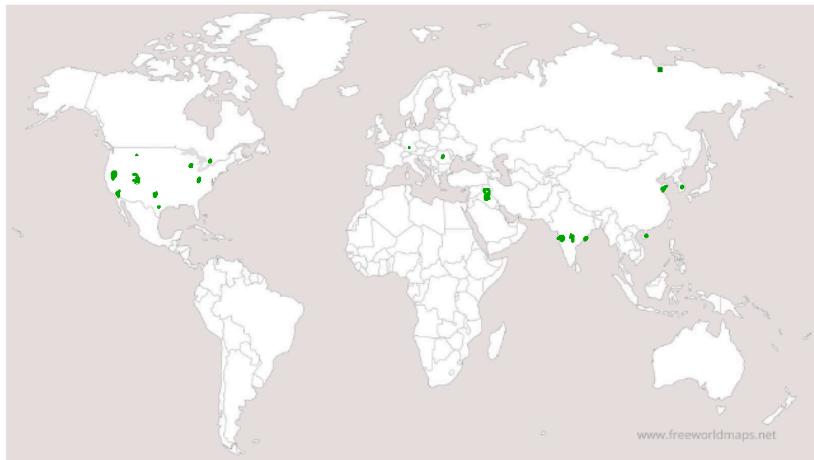
About TA: Robert Termuhlen (termuhle@msu.edu)

Office Hours: Tue & Thur 2 – 3 pm

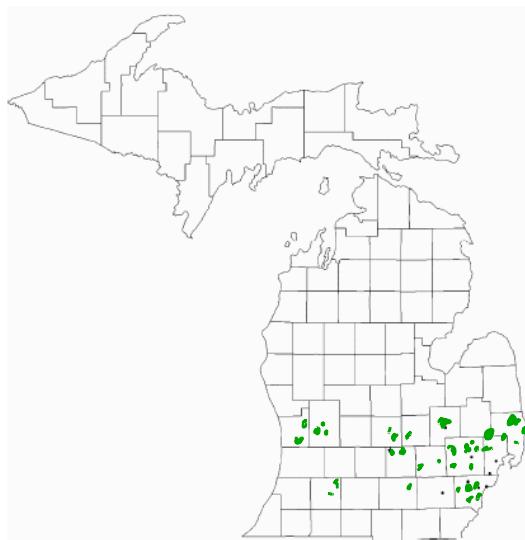


SPARTANS WILL.

Introduce yourself. Annotate you



www.freeworldmaps.net



Get Printable Maps From:
WaterproofPaper.com



What is the course about?

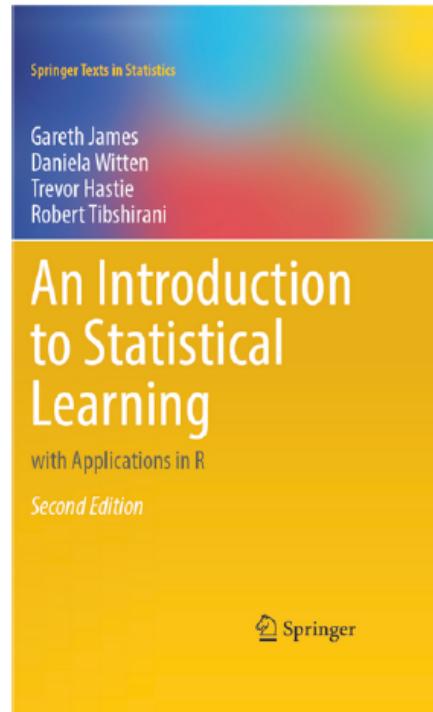
- Topics:
 - Fundamental concepts of data science
 - Regression
 - Classification
 - Dimension reduction
 - Resampling methods
 - Tree-based methods, etc.

Deep learning

- Textbook:
 - An Introduction to Statistical Learning, with applications in R. / paper ↗
 - The Elements of Statistical Learning: Data Mining, Inference, and Prediction ↙
 - Also have a youtube list



- Textbook:
 - An Introduction to Statistical Learning, with applications in R
 - Online version available!
 - <https://www.youtube.com/user/joshstarmers>



Github and where to find slides and jupyter notebooks

MICHIGAN STATE
UNIVERSITY

https://github.com/yuyingxie/CMSE381_SP2023

The screenshot shows the GitHub repository page for 'yuyingxie / CMSE381_SP2023'. The repository is public and contains one branch ('main') and one tag ('0'). The repository has 3 commits, the latest being 'yuyingxie syllabus' at 'ec9ea5e' 15 hours ago. It also contains a 'DataSets' folder with a 'dataset' file, a 'Syllabus' folder with a 'syllabus' file, and a 'test' file. A message encourages adding a README, with a 'Add a README' button. The repository has 0 stars, 1 watching, and 1 fork. It has no releases or packages published. The Languages section shows 100.0% Jupyter Notebook usage. The footer includes links to GitHub's Terms, Privacy, Security, Status, Docs, Contact GitHub, Pricing, API, Training, Blog, and About pages.

Search or jump to... Pull requests Issues Codespaces Marketplace Explore

yuyingxie / CMSE381_SP2023 Public

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main 1 branch 0 tags Go to file Add file Code

yuyingxie syllabus dataset ec9ea5e 15 hours ago 3 commits

DataSets dataset 2 days ago

Syllabus syllabus 15 hours ago

test test 2 days ago

Help people interested in this repository understand your project by adding a README. Add a README

About No description, website, or topics provided. 0 stars 1 watching 1 fork

Releases No releases published Create a new release

Packages No packages published Publish your first package

Languages Jupyter Notebook 100.0%

© 2023 GitHub, Inc. Terms Privacy Security Status Docs Contact GitHub Pricing API Training Blog About

Yuying Xie

Zoom Link: shorturl.at/ghpNR

Dr. Xie

Time: Tues 12:00-1:30 pm

Zoom or Egr 1513

Bob Termuhlen

Time: ~~TBA~~ Tue & Thur 2-3 pm

Zoom or **Egr 1513**

1508 A



Yuying Xie

Instruction Format

- **In Class:**

- Stop me if you have any question!
- We will have coding section in most of the classes (bring your laptop)
- What happens in CMSE 381 **stays in** CMSE 381



- **Grade Breakdown:**

- 20% for Homework
- 15% for Quizzes
- 20% Midterm I
- 20% Midterm II
- 25% for the Final Project



Homework Policies

- Homework assigned bi-weekly, it is due before midnight on Friday.
- 20% of final grade.
- Late homework will have penalties (TA is full in charge.).
 - 24 hours late: 5% penalty
 - 48 hours late: 15% penalty
 - > 48 hours: No late work accepted
- Allowed to work with other students but homework should be in your own words.
- ~~Each homework carries equal weight.~~
- Code needs to follow Google Style

Quiz Policies

- Once a week, there will be a 10 minutes Quiz at the beginning of the class.
- 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
 - There will be some 3-minutes bonus-quizzes (1 credit) at the end of some classes. These credits can be added to your formal quizzes but your total Quiz-score is topped.



Graduate School

- 1. Research experience ↪
- 2. Letters ↪
- 3. Fellowship (NSF)
- 4. Course work ↪
- 5. Broader Impact ↪

Industrial

- 1. Research experience ↪
- 2. Internship ↪
- 3. Programming skill (Leetcode) ↪
- 4. Letters ↪

Make yourself different!



What is Machine Learning?

- Arthur Samuel (1959, IBM): *Field of study that gives computers the ability to learn without being explicitly programmed*



What is Statistical Learning?

- Subfield of Statistics
- Statistical learning emphasizes models and their interpretability, and precision and uncertainty
- Machine learning has a greater emphasis on large scale applications and prediction accuracy.
- The distinction between machine learning and statistical learning has become more and more blurred.



Why should you care?

- Nowadays, 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)
 - Sport team (MSU, Lions, etc)
- Interpreting results and prevent any **bad decisions.**



Example: Presidential election



<http://fivethirtyeight.com/>

On 2008, he correctly predicted the winner of 49 of the 50 states and all 35 U.S. Senate races .

On 2012, he correctly predicted the winner of all 50 states and the District of Columbia. That same year, Silver's predictions of U.S. Senate races were correct in 31 of 33 states.

All the data he used came from internet!

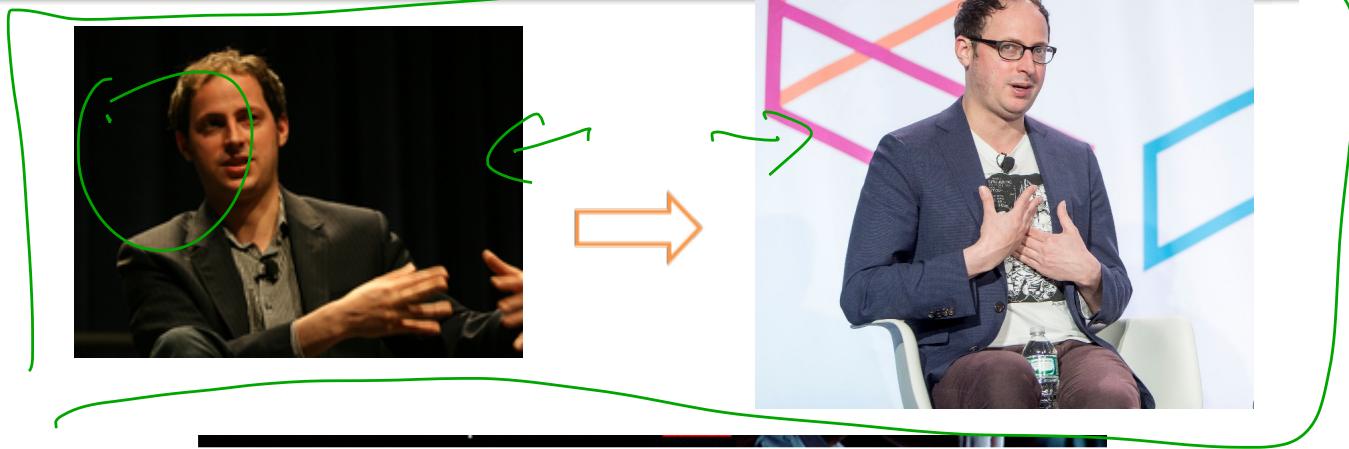
Dow down 313 points after Obama won

On 2016, oops.....



Example: Presidential election 2016

MICHIGAN STATE
UNIVERSITY



FiveThirtyEight

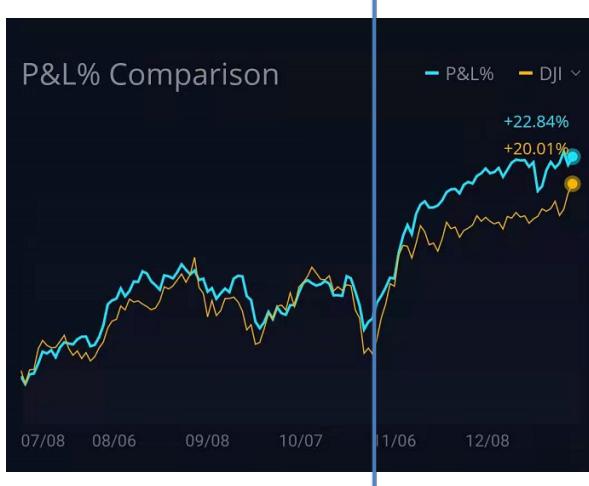
Politics Sports Science & Health Economics Culture

NOV. 11, 2016, AT 4:09 PM

Why FiveThirtyEight Gave Trump A Better Chance Than Almost Anyone Else



Example: Presidential election



Example: Big Data

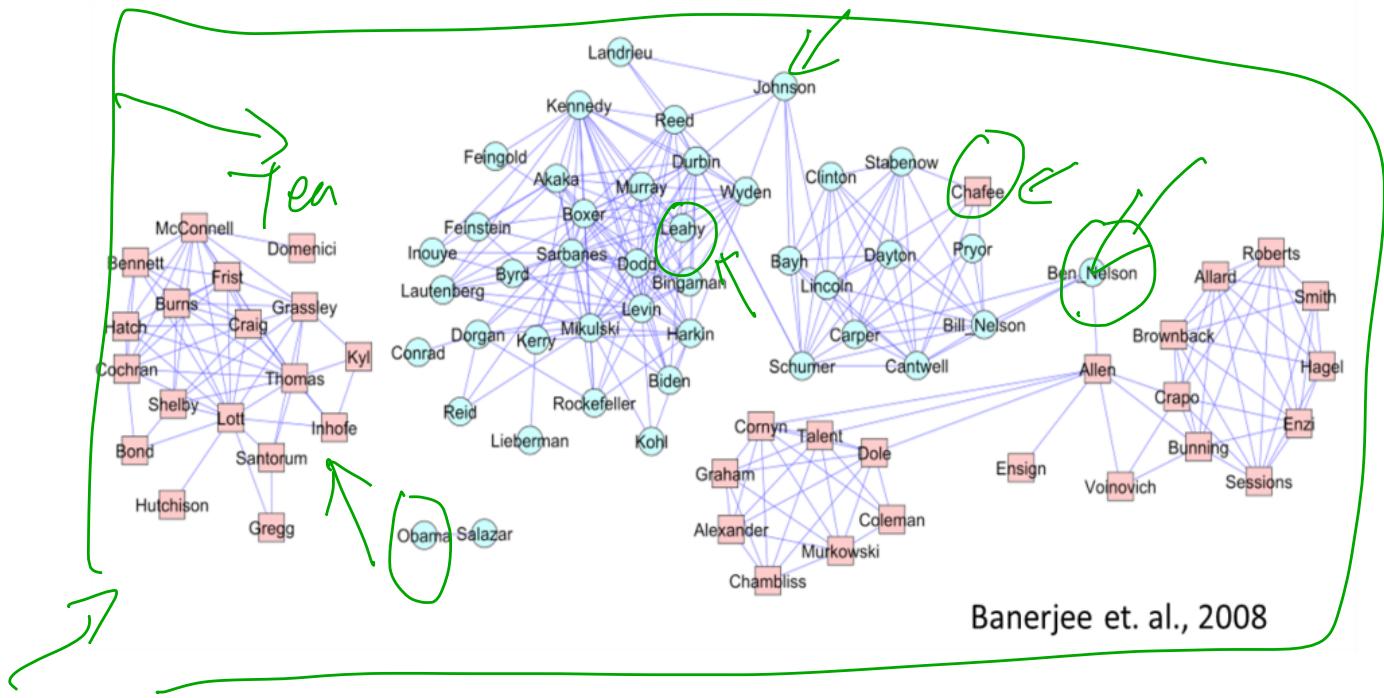
The screenshot shows the Facebook Marketplace interface. On the left, there's a sidebar with a search bar, 'Browse All' button, 'Stores' link, 'Your Account' dropdown, and a 'Create New Listing' button. Below these are 'Filters' (set to 'Okemos, Michigan - Within 40 miles') and 'Categories' (with a 'Vehicles' link). At the bottom of the sidebar is a URL: <https://www.facebook.com/marketplace/item/1046734842357256/>. The main area displays 'Today's Picks for You' with five items:

- \$550** iPhone X 256gb ATT Unlocked, SIM-Free
East Lansing, Michigan
- \$1** iPhone8 Plus 64GB Silver.
Please read info
East Lansing, Michigan
- \$200** Yamaha p95 digital piano
Spring Arbor, Michigan
- \$280** Brand new iPad 32GB 7th Gen. silver
Edgemont Park, Michigan

Annotations include:

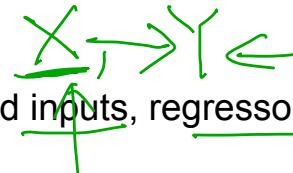
- A green circle highlights the 'Create New Listing' button in the sidebar.
- A green circle highlights the profile picture of a user in the top right corner, which has a red notification badge with the number '5'.
- A green arrow points from the bottom of the page towards the highlighted profile picture.

Graphical model



The 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).

$$Y = f$$

- In the regression problem, Y is quantitative (e.g price, blood pressure).

Continuous

- In the classification problem, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).

- We have training data $(x_1, y_1), \dots, (x_N, y_N)$. These are observations (examples, instances) of these measurements.



Objectives

On the basis of the training data we want to?

- Accurately predict unseen test cases.
- Understand which inputs affect the outcome, and how.
- Assess the quality of our predictions and inferences

A hand-drawn green diagram consisting of two 'Y' shaped symbols. A horizontal line connects their bases. To the right of this line is a square symbol, indicating a squared difference between the two outputs. This represents a loss function, such as mean squared error, used to measure the difference between predicted values and actual values.



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.



Marketplace

Search Marketplace

Browse All

Stores

Your Account

Create New Listing

Filters

Okemos, Michigan - Within 40 miles

Categories

Vehicles

<https://www.facebook.com/marketplace/item/1046734842357256/>

Today's Picks for You

Okemos · 40 mi

Item	Price	Description	Location
iPhone X 256gb ATT Unlocked, SIM-Free	\$550	iPhone8 Plus 64GB Silver. Please read info	East Lansing, Michigan
Yamaha p95 digital piano	\$1		East Lansing, Michigan
Brand new iPad 32GB 7th Gen. silver	\$200		Spring Arbor, Michigan
	\$280		Edgemont Park, Michigan

Today's Picks for You

Okemos · 40 mi

Item	Price	Description	Location
iPhone X 256gb ATT Unlocked, SIM-Free	\$550	iPhone8 Plus 64GB Silver. Please read info	East Lansing, Michigan
Yamaha p95 digital piano	\$1		East Lansing, Michigan
Brand new iPad 32GB 7th Gen. silver	\$200		Spring Arbor, Michigan
	\$280		Edgemont Park, Michigan

Today's Picks for You

Okemos · 40 mi

Item	Price	Description	Location
iPhone X 256gb ATT Unlocked, SIM-Free	\$550	iPhone8 Plus 64GB Silver. Please read info	East Lansing, Michigan
Yamaha p95 digital piano	\$1		East Lansing, Michigan
Brand new iPad 32GB 7th Gen. silver	\$200		Spring Arbor, Michigan
	\$280		Edgemont Park, Michigan



SAFEGRAPH

It obtains information related to customers via their mobile devices using 17 trillion location markers and various mobile applications that partner with SafeGraph.



[HOME](#) > [SCIENCE](#) > VOL. 360, NO. 6392 > THE EFFECT OF PARTISANSHIP AND POLITICAL ADVERTISING ON CLOSE FAMILY TIES

REPORT



The effect of partisanship and political advertising on close family ties

M. KEITH CHEN  AND RYNE ROHLA 

Curtailed conversations

Most articles written about U.S. politics in the past few years have mentioned the increasing polarization of the electorate. But is this real, or does it merely reflect the increasing polarization of the media? Chen and Rohla estimate that in 2016, Thanksgiving dinners in which the hosts and guests lived in oppositely voting precincts were up to 50 minutes shorter than same-party-precinct dinners. That is, family members, adjured to avoid talking about contentious subjects, may have simply talked less.





SAFEGRAPH

It obtains information related to customers via their mobile devices using 17 trillion location markers and various mobile applications that partner with SafeGraph.

How can you use this data?

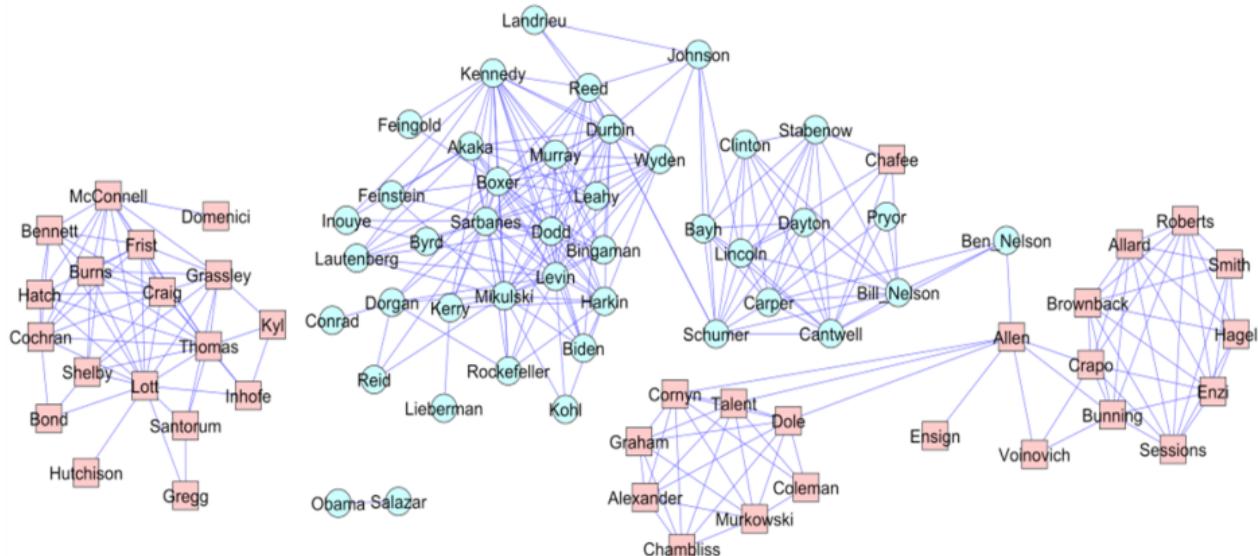


Plan for the lab

- Find a group of 4 or so.
- Download the jupyter notebook and the csv file from github.
- Get started!



Question



Banerjee et. al., 2008