

POLI 175

Class 01 - Introduction

Dr. Umberto Mignozzetti
UCSD

Introduction

- Welcome! This is POLI 175 - Machine Learning for Social Scientists
- We are in the age of data-driven decisions.
- Every day, we produce 2.5 quintillion bytes of data.
- But how much is that?

Introduction

- My computer is competent.
- It is new, fast, and reliable. It is capable of storing 1.0 Terabytes.
- But one quintillion bytes is equal to 1,000,000 Terabytes.
- Every day, we produce the equivalent of 2.5 million computers like mine of data!

Introduction

- This is popularly known as Big Data.
- However, data per se means nothing! Big data is just a passive description of the world we live in now.
- To prove that, try to take a large dataset and learn something from it.
- If you want inspiration, take all your pictures and try to create coherent slides of moments of your life.

Introduction

- This is hard and time-consuming. You would spend days doing it.
- But note that, interestingly, your phone does that to you every day!
- Almost every week, I open up my iPhone, and it shows me a slide show with music and pictures of my family and me.
- How does it do that? Machine Learning!

Introduction

- Machine Learning is a branch of Artificial Intelligence that uses data and algorithms to imitate how humans learn (IBM).
- Algorithm: short for *recipe*.
- Data: can be anything.
- **And note the intent:** *Learn* here means both make sense of things, discover patterns, and predict things.

Introduction

- How can we use this as Social Scientists?
- Many applications in Political Science, Economics, Public Policy, etc.
- In this class, we will learn how to use the most popular Machine Learning techniques.
- We will have an *applied focus*, meaning that we will talk about theory, but the focus will be on generating results.

Introduction

- We will use three books:
 1. [ISL] James et al. (second edition, 2021) *Introduction to Statistical Learning with Applications in R*. Springer.
<https://www.statlearning.com>
 2. [MG] Müller & Guido (2017) *Introduction to Machine Learning with Python*. O'Reilly.
 3. [PDA] McKinney (2013) *Python for Data Analysis*. O'Reilly.

Introduction

Grading:

Assignment	Points	Weight	Due Date
Participation x 28	10	10 %	Every lecture
Quizzes x 28	10	20 %	Before next lecture start
Problem Sets x 4	10	50 %	PS 1 – Jan 27 PS 2 – Feb 10 PS 3 – Feb 24 PS 4 – Mar 10 All PSs will be live two weeks before their due date.
Final Exam	10	20 %	March 24, 3-6 pm
100%			

Introduction

Communication expectations:

- Office hours: every day, from 11 AM to 12 PM
- Office hours scheduling: **Calendly**
- Email: **umbertomig@ucsd.edu**
- I may miss your email. It happens. If this is the case, resend it.

Introduction

Syllabus:

- Check it on Canvas.
- It is subject to change (most likely, every other week).
- But I'll tell you what I'm changing there.

Introduction

Tentative schedule:

1. Intro
2. Python Refresh
3. Regression
4. Classification

Introduction

1. Tweaking (resampling, model selection, regularization, non-linear models)
2. Tree-based models
3. SVM
4. Deep Learning
5. PCA and Clusters

Introduction

About me:

- I hold a Ph.D. in Political Science from New York University.
- I work on how political institutions affect welfare in developing economies.
- I am from Brazil, and I like to watch soccer in my free time.

Introduction

About you:

- What is your name?
- What is your major / minor / field of study?
- What do you like to do in your free time?

Introduction

To do:

1. Check your first quiz! It is about the syllabus and this lecture.
2. Make sure you have access to Canvas.
3. Reach out to me if you are not fluent in Python.

Questions?

See you in the next class!