An Introduction to Machine Learning for Social Science

Class 1: Introduction

Rochelle Terman

Postdoctoral Fellow
Center for International Security & Cooperation
Stanford University

January 9th, 2018

Why Machine Learning for Social Science?

Examples of Learning Problems

- Predict who will win the 2020 Presidential Election, based on public opinion polls and economic data.

Examples of Learning Problems

- Predict who will win the 2020 Presidential Election, based on public opinion polls and economic data.
- Estimate a person's wage based on his or her age, education, and gender.

Examples of Learning Problems

- Predict who will win the 2020 Presidential Election, based on public opinion polls and economic data.
- Estimate a person's wage based on his or her age, education, and gender.
- Classify articles as either "fake news" or "real news" based on the words in the title.

Examples of Learning Problems

- Predict who will win the 2020 Presidential Election, based on public opinion polls and economic data.
- Estimate a person's wage based on his or her age, education, and gender.
- Classify articles as either "fake news" or "real news" based on the words in the title.
- Identify substantive topics or themes in a collection of documents.

Machine learning refers to a vast set of tools that can learn from and make predictions on data.

- Supervised learning: Predict or estimate an *output*, usually quantitative (wage) or categorical (Republican/Democrat), based on a set of *inputs*.
- Unsupervised learning: We observe only the inputs, but no measure for the outputs. Our task is to learn relationships and structures from such data.

Machine Learning Today

- 1800s–1980s: linear models

Machine Learning Today

- 1800s–1980s: linear models
- Since 1980s:

$$\begin{aligned} & \text{More computational power} \\ + \; & \text{More data} \\ + \; & \text{New techniques} \\ = \; & \text{Broader applications, bigger audience} \end{aligned}$$

Machine Learning Applications

Industry

- Measure consumer opinion
- Deliver engaging content to users

Machine Learning Applications

Industry

- Measure consumer opinion
- Deliver engaging content to users

Public Sector

- Predict disease onset
- Assist criminal sentencing

Machine Learning Applications

Industry

- Measure consumer opinion
- Deliver engaging content to users

Public Sector

- Predict disease onset
- Assist criminal sentencing

Campaigns

- Classify voters based on likely voting, using consumer information
- Identify ideology based on social media behavior

Machine Learning Applications

Industry

- Measure consumer opinion
- Deliver engaging content to users

Public Sector

- Predict disease onset
- Assist criminal sentencing

Campaigns

- Classify voters based on likely voting, using consumer information
- Identify ideology based on social media behavior

Social Science

- Infer extent and strategy of Chinese censorship: King, Pan, and Roberts (2014):
- Measure polarization in political institutions: Clinton, Jackman, and Rivers (2004):

Presumptions for this Course

1) Machine learning is relevant and useful in a wide range of academic and non-academic fields, beyond just statistics.

Presumptions for this Course

1) Machine learning is relevant and useful in a wide range of academic and non-academic fields, beyond just statistics.
2) This diverse group should be able to understand the models, intuitions, and strengths and weaknesses of the various approaches.

Presumptions for this Course

1) Machine learning is relevant and useful in a wide range of academic and non-academic fields, beyond just statistics.
2) This diverse group should be able to understand the models, intuitions, and strengths and weaknesses of the various approaches.
3) While it is important to know what job is performed by each cog, it is not necessary to have the skills to construct the machine inside the box.

Presumptions for this Course

1) Machine learning is relevant and useful in a wide range of academic and non-academic fields, beyond just statistics.
2) This diverse group should be able to understand the models, intuitions, and strengths and weaknesses of the various approaches.
3) While it is important to know what job is performed by each cog, it is not necessary to have the skills to construct the machine inside the box.
4) Applying machine learning methods to “real-world problems” requires both quantitative skills + social science reasoning.

Core Learning Objectives

Ultimate Goal: Introduce students to modern machine learning techniques and provide the skills necessary to apply the methods widely

Core Learning Objectives

Ultimate Goal: Introduce students to modern machine learning techniques and provide the skills necessary to apply the methods widely

Proximate Goals

1) Learn about core concepts in machine learning and statistics, developing skills that are transferable to other types of data and inference problems.
2) Be introduced to substantive problems and apply the techniques from the course.
3) Develop their programming abilities in R.
4) Be able to learn independently and tackle more advanced topics and challenges in data analysis.

Course Outline

Supervised Learning:

- simple and multiple regression
- classification and logistic regression
- LASSO
- cross validation

Unsupervised Learning:

- clustering
- topic models
- principle component analysis

Other Stuff

- text as data
- how to assess performance
- the politics of machine learning (bias, transparency, etc)

This Course Will Not

- Go into the technical details behind machine learning methods, such as optimization algorithms and theoretical properties.
- Cover all machine learning tools, or even most of them.
- Teach you to be a professional programmer or software developer.

Prerequisites

150A or equivalent. This includes:

- A mechanical understanding of regression
- A brief introduction to statistical inference
- Experience in the R programming language

This course is geared towards a 150B audience.

Instructors

- Main Instructor: Dr. Rochelle Terman

Instructors

- Main Instructor: Dr. Rochelle Terman
- TA: Haemin Jee

Instructors

- Main Instructor: Dr. Rochelle Terman
- TA: Haemin Jee
- TA: Tongtong Zhang

Lecture & Sections

Semi flipped classroom

- 1/2 lecture, 1/2 coding in R.
- Bring your laptop, prepare to close it
- Install R, RStudio, and R markdown now!

Sections

- Review lecture materials, finish exercises
- Improve R programming
- Introduce Python***

Materials & Websites

Canvas

- Lectures Notes, Code, and Data
- Homework (Assigned and Returned)

Piazza

- Questions and discussion
- Ask question anonymously
- Communicate with instructors and each other
- Use Piazza first, before email!!!

Evaluation

- Homework: five assignments, 35% of final grade.
    - In general, assignments are assigned at the end of lecture, and due the following week. Exceptions will be noted.
    - Programming in R should be submitted in R markdown.
    - Submit on Canvas.
    - Collaboration is encouraged, write up your own.
- Group Project: 15% of final grade.
    - Teach the class about one ML topic we didn't cover.
- Midterm exam: One exam, 20% of final grade.
- Final Exam: 20% of final grade.
- Participation: 10%.
    - Attend class and ask questions.
    - Post on Piazza.
    - Actively participate in weekly sections.

Grading Policy and Accommodations

- All grades in this class are final.
- Extensions or incompletes will be given only to students with a documented emergency or illness.
- Let me know ASAP if you need special accommodations.

How to Be Successful in this Course

- Learning is 5% intelligence, 95% endurance.

How to Be Successful in this Course

- Learning is 5% intelligence, 95% endurance.
- Like learning to play an instrument or speak a foreign language, programming takes practice, practice, practice.

How to Be Successful in this Course

- Learning is 5% intelligence, 95% endurance.
- Like learning to play an instrument or speak a foreign language, programming takes practice, practice, practice.
- Program a little bit every day, preferably with others. Do the problem sets in pairs or groups.

How to Be Successful in this Course

- Learning is 5% intelligence, 95% endurance.
- Like learning to play an instrument or speak a foreign language, programming takes practice, practice, practice.
- Program a little bit every day, preferably with others. Do the problem sets in pairs or groups.
- Ask questions. Resist imposter syndrome.

How to Be Successful in this Course

- Learning is 5% intelligence, 95% endurance.
- Like learning to play an instrument or speak a foreign language, programming takes practice, practice, practice.
- Program a little bit every day, preferably with others. Do the problem sets in pairs or groups.
- Ask questions. Resist imposter syndrome.
- Stay organized.

How to Be Successful in this Course

- Learning is 5% intelligence, 95% endurance.
- Like learning to play an instrument or speak a foreign language, programming takes practice, practice, practice.
- Program a little bit every day, preferably with others. Do the problem sets in pairs or groups.
- Ask questions. Resist imposter syndrome.
- Stay organized.