STAT 546: Machine Learning in Data Science

Introduction

Ruoqing Zhu, Ph.D. <rqzhu@illinois.edu>

https://teazrq.github.io/stat432/

March 7, 2024

University of Illinois at Urbana-Champaign

Welcome to Phase II

- Course Website
  - https://teazrq.github.io/stat546/
- Instructor: Ruoqing Zhu, Ph.D <rqzhu@illinois.edu>
- Teaching Assistant: Yuhan Li <yuhanli8@illinois.edu>
- Office hour:
  - Yuhan: Mon/Wed 7 - 8 PM, Zoom: 89040207215, pw: 260056
  - Ruoqing: Tue/Thu 2 - 3 PM, 137 CAB

Overview

- Basic course information
  - Textbook
  - Course website
  - Homework
  - Project
- Topics and objectives
- ChatGPT, GitHub Copilot and other tools

Textbook

- Most of the course material is based on published papers, textbooks, and open course notes
- You may start with the textbooks such as
  - Sutton, R.S. and Barto, A.G., 2018. Reinforcement learning: An introduction.
  - Imbens, G.W. and Rubin, D.B., 2015. Causal inference in statistics, social, and biomedical sciences.

Course Website

- Canvas: discussion, grades
- github.io/stat546/: course material, posting homework
- Gradescope: submit homework

Discussion Board

- Canvas discussion board as the primary platform for communication
- Each homework has a thread, post your question as a new reply to the main thread
- For email communications, start with “Stat 546” in your email title.

Homework

- We will have 5 or 6 sets of homework (approx. 1 per week)
- Assigned on Monday and due at Thursday (11:59PM) of the following week
- Late submission allowed: up to 4 days, 5% penalty per day
- Submit to gradescope (Entry code: VB756J)

Homework

- All homework reports should be submitted in PDF format with all code chunks visible
- Logic of the code should be clear, well structured (readable, comments can be helpful)
- Break down complex problems into smaller, manageable functions or modules and clearly state the intension of each modules
- Tables/Figures should have clear legend, caption
- Key results should be highlighted

Topics and Objectives

What will we learn?

- Causal Inference
  - Randomized Trial, observational study, propensity score weighting
  - Doubly robust estimators, instrumental variable
- Personalized Medicine
  - Conditional average treatment effects and optimal policy
  - Outcome weighted learning
  - Dynamic treatment rules
- Reinforcement Learning
  - Markov decision process
  - Bellman equation and properties
  - Policy evaluation, policy optimization, online and offline settings
  - Various algorithms in RL

Why these topics?

- How the field evolved
- Causal inference was used extensively in economics, business, political science, etc. Most of these are population conclusion/inference of the policy. Also used extensively in drug developments.
- After 2000, genomic sequencing becomes feasible and affordable, and individual level health data becomes digitalized, making individual leveling decisions possible
- Reinforcement learning seen a huge boost during 2010's, when AI start to beat human in some examples such as Atari 2600 and AlphaGo
- However, application of RL to the medical field still faces many challenges (data, signal/noise, ethical, etc.)

What should we learn?

- Ways of formulate the problem and ways of thinking
- Various algorithms, pros and cons
- Being able to implement some algorithms

Prerequisites

- Probability: probability and random variables, distributions
- Statistics: estimators, likelihood, linear regressions, a sense of statistical convergence
- Mathematics: linear algebra and calculus
- Some prior knowledge of R and Python

GPT and other AI tools

- Use them!
- But at your own risk

Questions?