STAT 432: Basics of Statistical Learning

Introduction

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

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

August 21, 2023

University of Illinois at Urbana-Champaign

Welcome to STAT 432

- Course Website
  - https://teazrq.github.io/stat432/
- Instructor: Ruoqing Zhu, Ph.D <rqzhu@illinois.edu>
- Teaching Assistant: Zexuan Zhang <zexuanz4@illinois.edu>
- Office hour: Mon through Thur, 7 - 8 PM
- Office Hour Zoom
  - Zoom: 89153753457, password: 432

About Myself

- Machine learning in general
- Reinforcement learning, random forests, survival analysis, etc.
- Personalized medicine, decision making, computational challenges ...
- Applications to real world problems: sepsis, infectious diseases, nutrition and food, cancer, genetics ...
- more at sites.google.com/site/teazrq/

Overview

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

Textbook

SMLR **Statistical Learning and Machine Learning with R**

by Zhu, R.

[online]

ISL **An Introduction to Statistical Learning: With Applications in R**

by James, G., Witten, D., Hastie, T. & Tibshirani, R.

[free PDF]

ESL **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**

by Hastie, T., Tibshirani, R. & Friedman, J.

[free PDF]

Course material goes beyond just a few textbooks!

Course Website

- Main website: https://teazrq.github.io/stat432/
  - post course material, homework, project and other info
- Canvas: https://canvas.illinois.edu/
  - Announcements
  - Discussion board
- Gradescope https://www.gradescope.com/courses/570816
  - Submit HW and project
  - Entry code: **WV7ZDP**

Discussion Board

- Canvas discussion board as the primarily platform of communication
- For email communications, start with "Stat 432" in your email title.

Homework

- We have approximately 10 sets of homework (1 per week), depending on the course progression
- Assigned on Monday and due at Thursday (11:59PM) of the following week
- Late submission allowed: up to 4 days, 5% penalty per day
- The lowest score dropped
- Submit to gradescope (.pdf, with all code chunks visible)

Midterm Quizzes

- Two in-class midterm quizzes
- 10 - 15 multiple choices or ture/false questions, cumulative
- Each determines 5% of the total score
- Dates: Oct 10 and Nov 16
- Examples on course website

Final Project

- Two options:
  - **[Option 1]**: Default project; Dataset and objectives provided
  - **[Option 2]**: Self-proposed project
    - Required for graduate students
    - Complex data and goals. Cannot be a straightforward classification or regression problem.
    - Setup a meeting with me no later than Nov 3rd. Update project progression before presentation.
    - In-class 15-min presentation
- Both options need to submit a 12-pages final report
- Maximum 3 members per team
- Previous projects and presentations can be found at the project page

Topics and Objectives

Data Analytic

Typical requirements for a data analytic:

- Domain knowledge, understand the data and the goal
- Know what model(s) to use and how to evaluate them
- Interpret results and communicate with others

Some examples

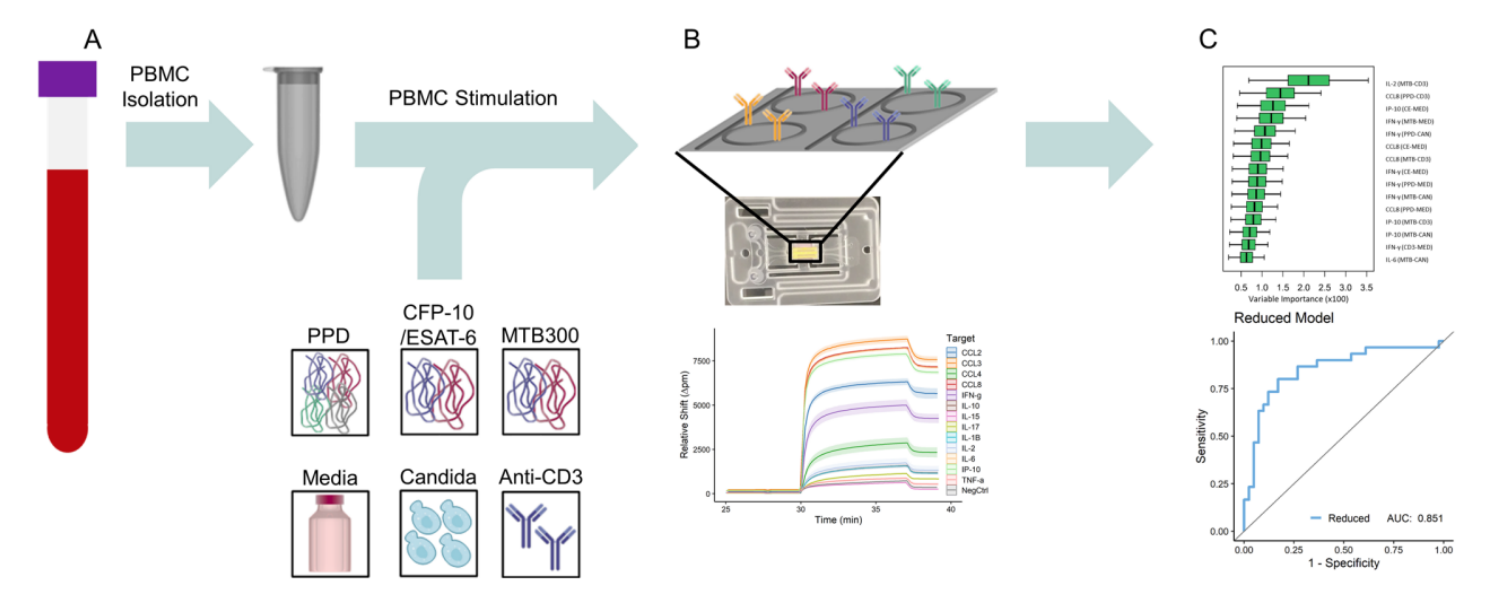

Figure 1: Identify latent tuberculosis infection with cytokine biomarkers, Robison et al., 2021

- challenges: low sample size, unbalanced group, ...

Some examples

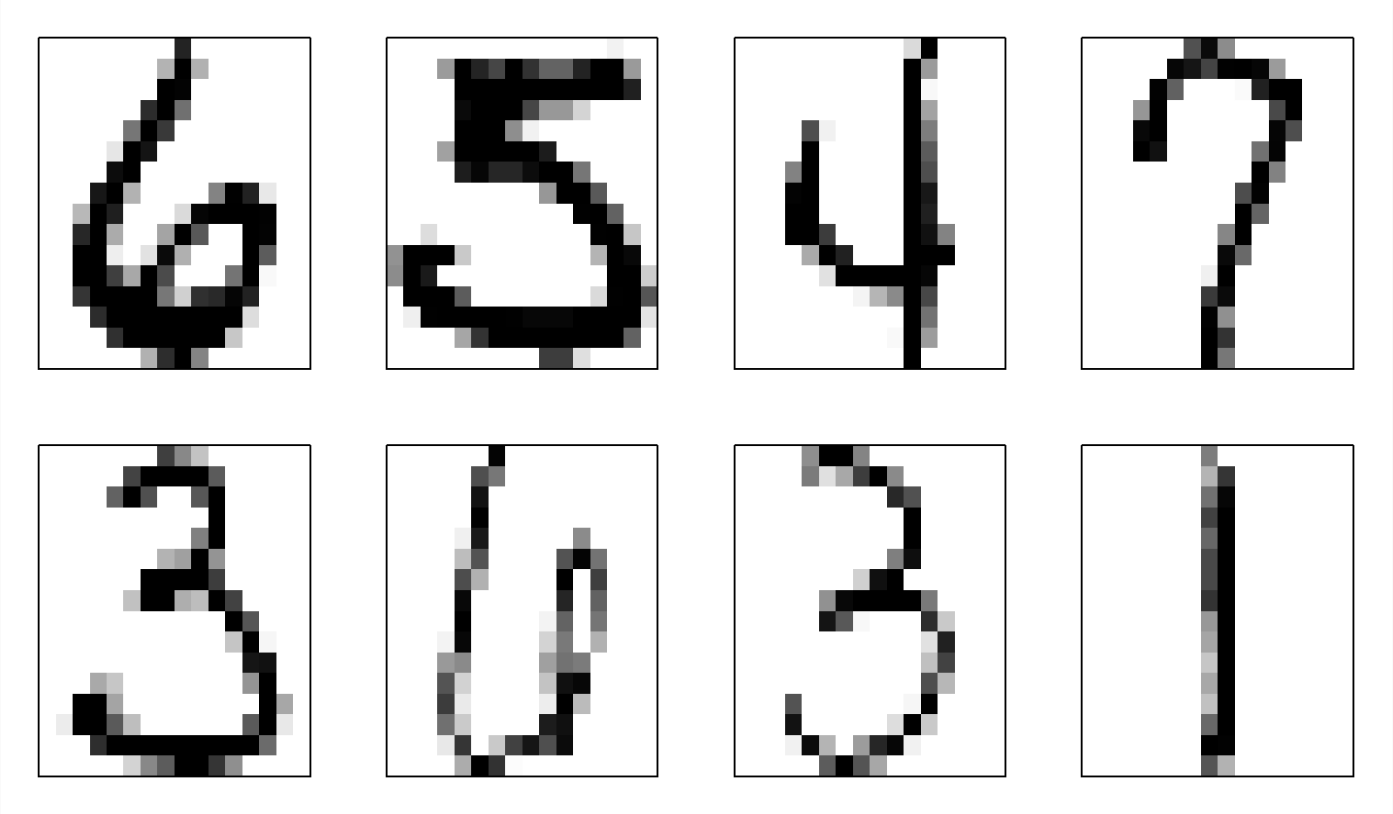

Figure 2: Hand written digit data from ElemStatLearn

- challenges: high-dimensionality, high correlation, non-linear

Some examples

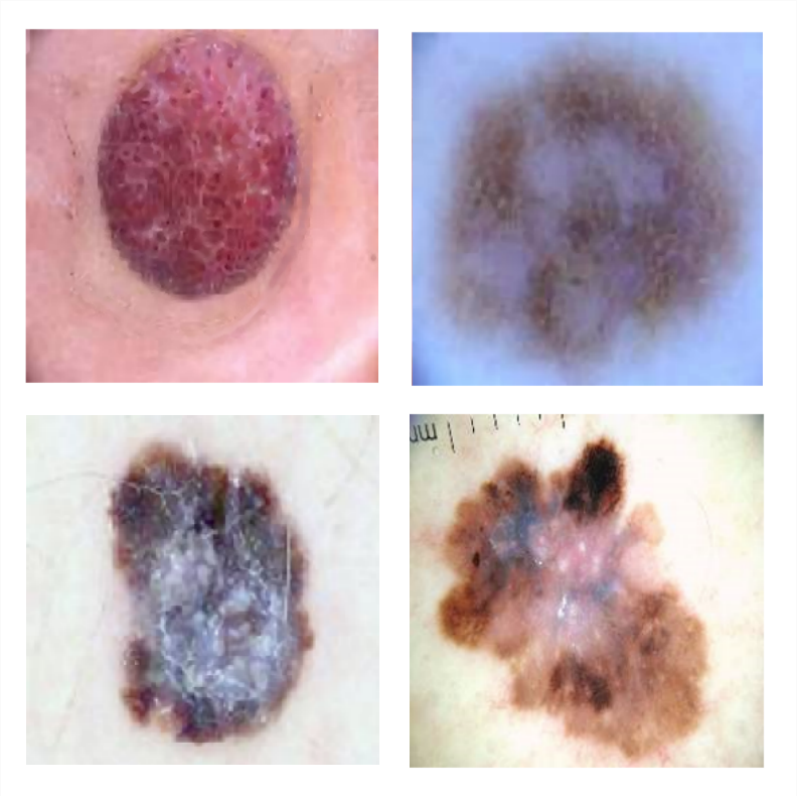

Figure 3: Dermoscopic Image Classification, Li et al., 2021

- challenges: no well-defined features

Some examples

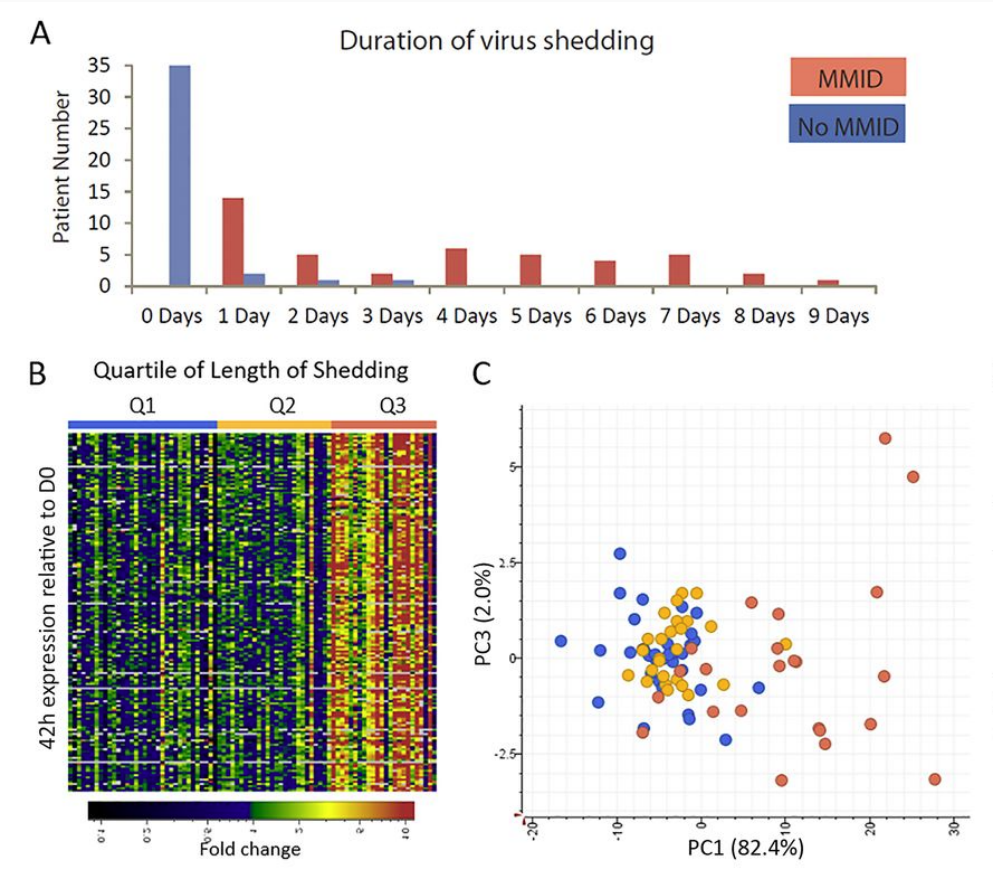

Figure 4: Gene expression changes after influenza infection, Walters et al., 2019

- challenges: longitudinal, small sample size, unsupervised

Some examples

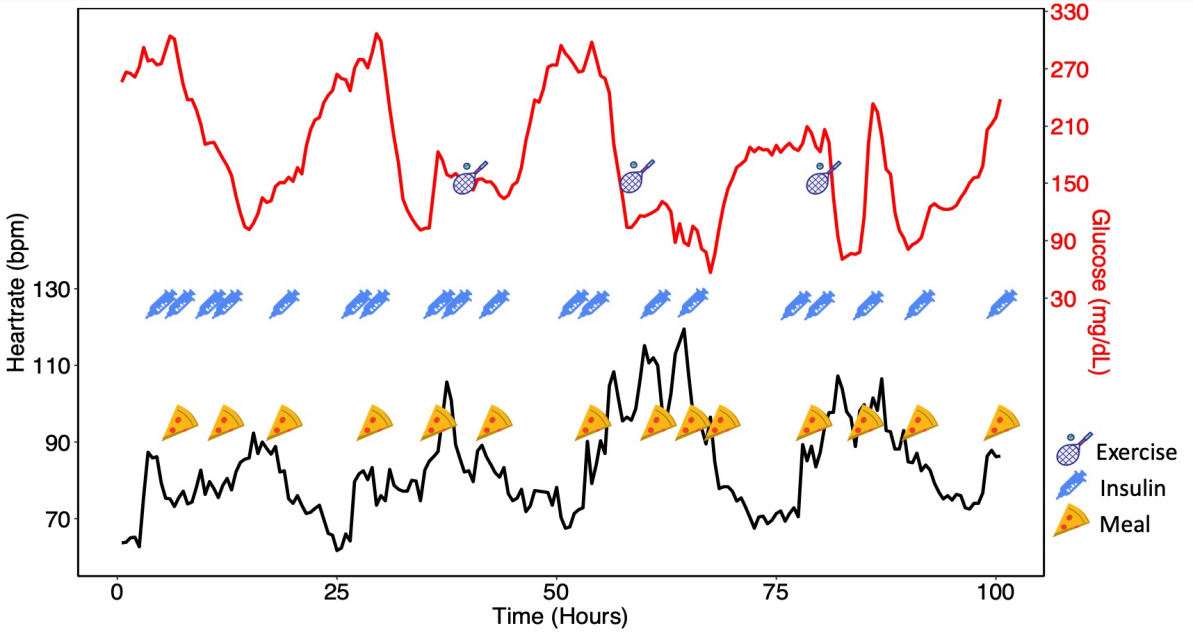

Figure 5: Dynamic treatment regime for diabetes, Zhou et al., 2021

- challenges: many decision points, robustness

ChatGPT

- GPT-4, Copilot, etc.
- Use them as much as you can
- How to write prompt
- Correctness
- Some other resources
  - Info from CITL
  - AI TA built by the Call team at NCSA [Talk]

What's covered?

- Understand the models
  - Supervised / unsupervised, Regression / classification
  - Suitable for high-dimensional, high correlated data?
  - How to tune parameters? Which one(s) should I tune?
  - Computational cost
  - Interpretation

- Understand the data...

What will we learn?

- Fundamental statistical concepts
  - Bias-variance trade-off
  - Cross-validation
  - Resampling
  - Statistical simulation
- Practical skills
  - Using R for implementation
  - Data processing
  - Using AI tools to assist learning and programming

Expectations

- Intense weekly schedule!
- Several homework assignments are very challenging
- Pay attention to weekly objectives
- Bring your laptop, practice and ask questions

Prerequisites

- Probability: probability and random variables, distributions
- Statistics: estimators, likelihood, linear regressions
- Mathematics: linear algebra (basic matrix operations) and calculus
- Some prior knowledge of R
- Be able to use ChatGPT

Next Step

- Install/update to the lasted version of R and R Studio (or VS Code)
  - Install packages: `devtools` and `tidyverse`
  - Update out-of-date packages, if any
- Read homework 1

Questions?