STAT 542: Statistical Learning

Introduction

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

Course Website: https://teazrq.github.io/stat542/

January 19, 2022

Department of Statistics
University of Illinois at Urbana-Champaign

Welcome to STAT 542

- M/W/F 3 - 3:50PM, 1002 Lincoln Hall
- Ruoqing Zhu, Ph.D <rqzhu@illinois.edu>
  - Office hour: M 4 - 4:50PM, R 1:30 - 2:20PM or by appointment
  - Zoom: 82244845695, password: 542
- Teaching Assistant: Tianning Xu <tx8@illinois.edu>
  - Office hour: T/W 7 - 8PM
  - Zoom: 669767288, password: 638309

About Me

- Research Interest
  - Personalized Medicine
  - Random Forests
  - Survival Analysis
  - Reinforcement Learning
  - ...
- Computational
  - Constrained Optimization
  - R packages
- Real world problems!

Overview

- Basic course information
  - Textbook
  - Course website
  - Homework
  - Project
- Topics and objectives

Textbook

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

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

- **Required** [free PDF]

ISL An Introduction to Statistical Learning: With Applications in R

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

- **Supplemental** [free PDF]

SMLR Statistical Learning and Machine Learning with R

by Zhu, R.

- **Supplemental** [online]

Course material goes beyond just a few textbooks!

Course Website

- Main website: https://teazrq.github.io/stat542/
  - post course material, homework and information
- Canvas: https://canvas.illinois.edu/courses/18369
  - Announcements
  - Discussion board
- Gradescope https://www.gradescope.com/courses/352176
  - Submit HW and project
  - Entry code: **KY457N**

Homework

- We have approximately 12 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 penalty: 5% per day, up to 4 days
- The lowest score can be dropped
- Submit to gradescope (.pdf, with all code chunks visible)

Discussion Board

- Canvas discussion board as the primarily platform of communication
- For email communications, start with "Stat 542" in your email title.

Midterm Exam

- Midterm Exam during the week of Apr 4th
- 15-20 multiple choices / filling-the-blank type of questions
- In-class, 50 minutes, closed-book
- A sample exam can be found at our course website
- No derivation or extensive calculation
- Will be curved if the median falls below 85%

Final Project

- Two options:
  - **[Option 1]**: Default project; Dataset and objectives provided.
    - Submit a 12-pages final report.
  - **[Option 2]**: Propose your own project
    - Complex data and goals
    - Setup a meeting with me (before Mar 31)
    - Final report
    - In-class 15-min presentation
- Up to 3 members per team
- Previous projects and presentations can be found at the project page

Final Score

- Homework 55%
- Midterm 15%
- Project 30%

Topics and Objectives

Overview

- Algorithm driven course that focus on
  - How to formulate a learning algorithm
  - How to solve them using various numerical optimization approaches
  - What are the statistical properties
  - How to interpret them
  - Use them in practice

- This course is **NOT** about learning how to use R packages to fit these models — you will need to hand code many models yourself

Statistical Learning Problems

Examples in Statistical Learning

Statistical learning is the process of extracting statistical regularities from datasets. They are motivated from real world problems. A few examples from HTF:

- Prostate Cancer (regression)
- Email Spam (binary classification)
- Handwritten Digits (multiclass classification)
- DNA microarray (clustering)

Some examples

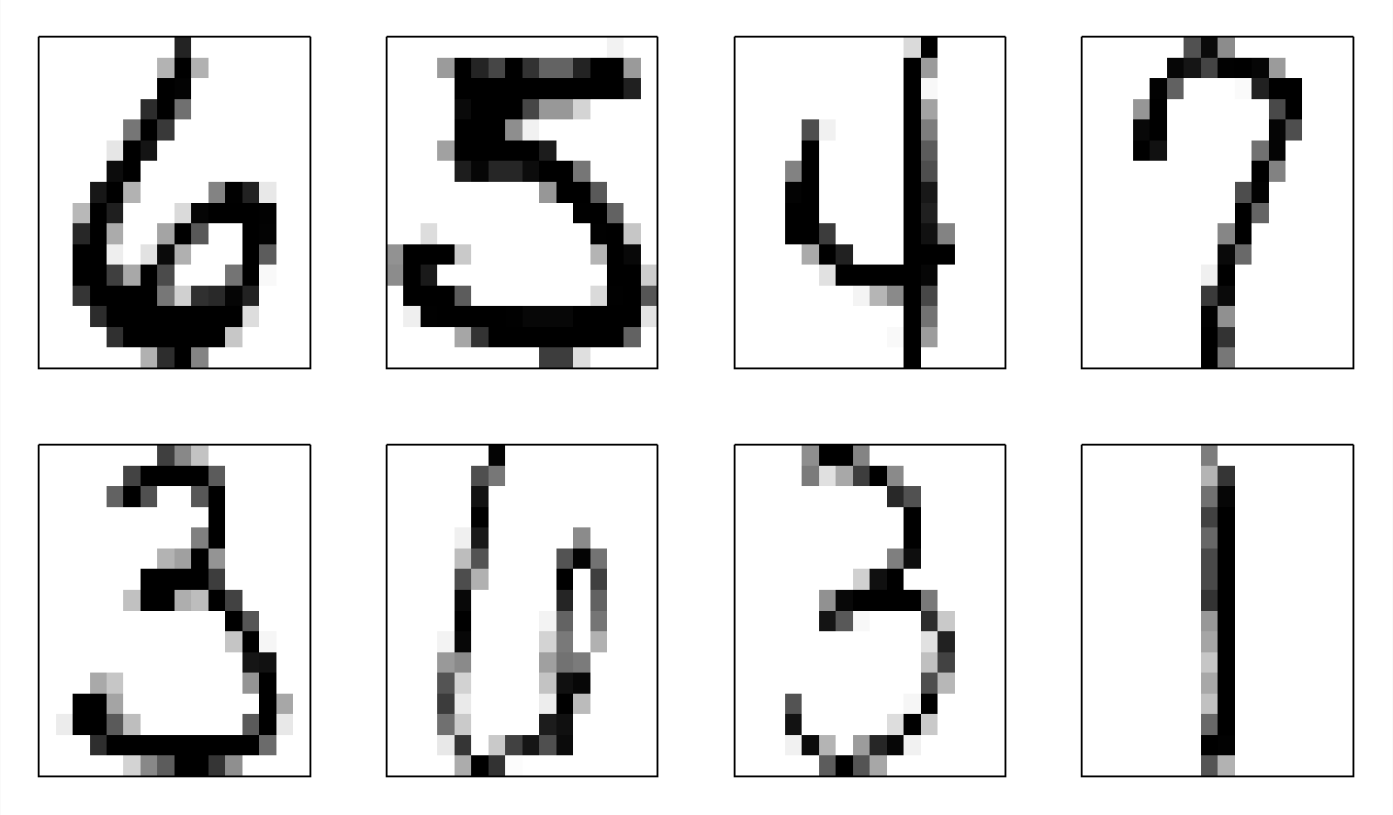

Figure 1: Hand written digit data from ElemStatLearn

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

Some examples

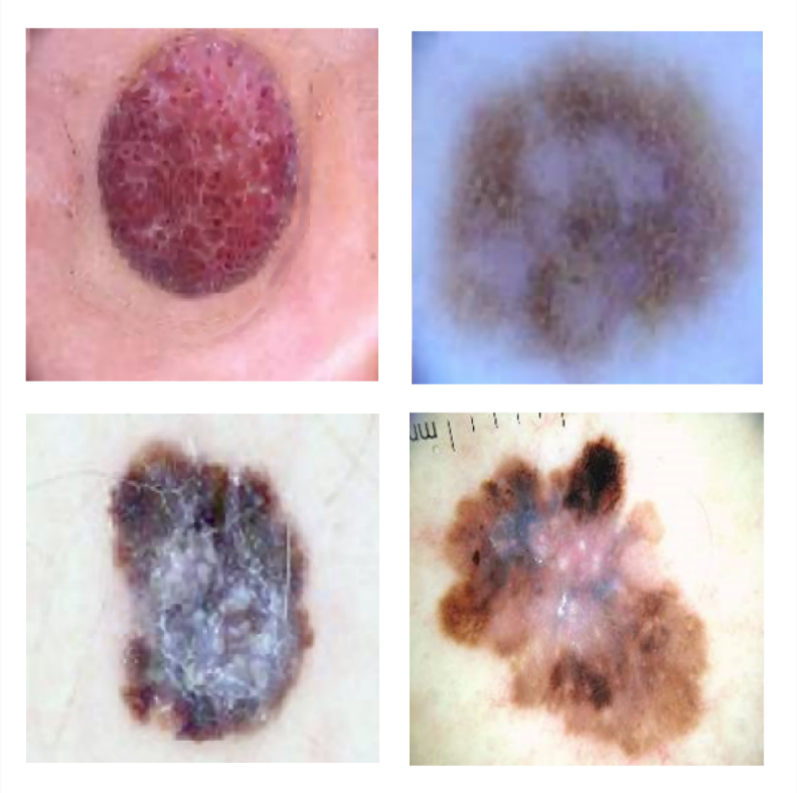

Figure 2: Dermoscopic Image Classification, Li et al., 2021

- challenges: no well-defined features

Some examples

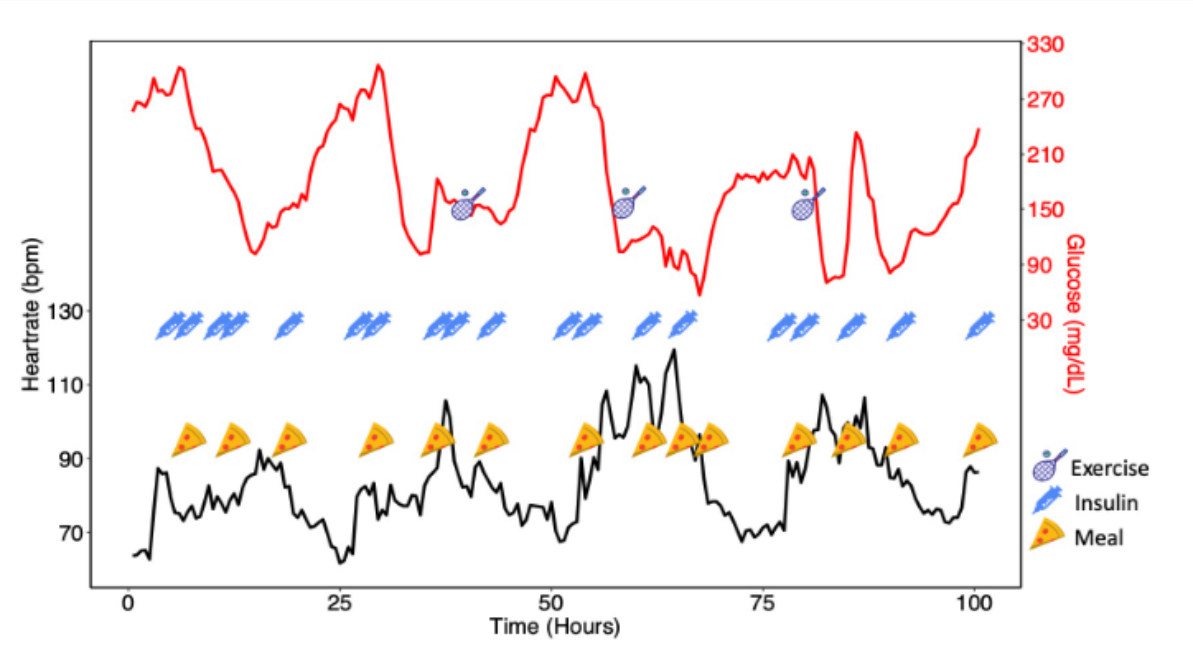

Figure 3: OhioT1DM study, Zhou et al., 2021

- challenges: longitudinal, dynamic changes

Course Overview

Course Overview

- Formulating, understanding and hand coding: Ridge, Lasso, KNN, splines, kernel methods, logistic regression, support vector machines, boosting, k-means, spectral clustering
- Optimization: Gradient descent, coordinate descent, primal-dual, general problems
- Concepts: bias-variance trade-off, simulation, local vs. global estimators
- Other skills: debugging, data processing, writing reports, visualization

Prerequisite

- Statistical/mathematical
  - Statistical concepts: random variables, samples, mean, variance, distributions, conditional variables and distributions, likelihood, estimators and linear regressions.
  - Linear algebra and multivariate calculus
- Programming skills
  - Programming in R or other equivalent
  - Basic optimization