MSDS 534: STATISTICAL LEARNING FOR DATA SCIENCE 16:954:534:01

FALL 2022, WEDNESDAY 6-9:00 PM, SEC203

1. Course Information

- Instructor: Han Xiao
- Office: Hill Center 451
- Office Hours: **Wednesday, 2:00–3:00pm**, or by appointment (on Zoom).
- Email: `hxiao@stat.rutgers.edu` **This is the only email account I check regularly!**
- Teaching Assistants: Zebang Li
  - Office hour: **Tuesday, 2:00–3:00pm, Hill Center 260**, or by appointment (on Zoom).
  - email: `zl326@stat.rutgers.edu`
- Prerequisites. FSRM588 or equivalent. Specifically, a comprehensive knowledge of the following: linear regression, shrinkage methods (lasso, ridge), model assessment and selection (information criterion, cross validation, boostrap), linear methods for classification (logistic regression, LDA), nonparametric methods (basis expansion, splines, smoothing). Basics of trees, additive models, model averaging, random forest and support vector machines.
- Texts.
  - Text 1 (ESL2). *The Elements of Statistical Learning*, by Trevor Hastie, Robert Tibshirani and Jerome Friedman. Springer, 2009, 2ed. Book website: `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`.
  - Text 2 (ISL2). *An Introduction to Statistical Learning with Applications in R*, by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Springer, 2ed 2021. Book website: `https://www.statlearning.com/`.
  - Text 3 (PRML). *Pattern Recognition and Machine Learning*, by Christopher Bishop. Springer, 2006. Book website: `https://www.microsoft.com/en-us/research/people/cmbishop/prml-book/`.
  - Text 4 (CO). *Convex Optimization*, by Stephen Boyd and Lieven Vandenberghe. Cambridge University Press, 2004. Book website: `https://web.stanford.edu/~boyd/cvxbook/`.
- Software: R. Free software available at `http://www.r-project.org/`. If you go to `Manuals` on the left panel of the webpage, you will find a good introduction *An Introduction to R*. A more advanced reference is *Modern Applied Statistics with S*, by Venables and Ripley. Springer, 2002, 4ed.
- Coursework and grades

  - Homework (50%): There will be 4 or 5 homework assignments. Submit your solution in **PDF** format (you can scan the handwritten pages into PDF) on Canvas.
  - Project (50%): proposal (5%), intermediate report (10%), presentation (15%), final report (20%).
- Rules of the homework
  - No late homework will be accepted without pre-approval of the instructor.
  - Students are encouraged to discuss the homework with classmates, the TA and the instructor. But each student needs to hand in an independent homework by himself/herself.
  - Computer generated output without detailed explanations and remarks will not receive any credit. Make sure to use different fonts to distinguish your own words from the computer output.
  - When you send emails about this course, please use the title "MSDS 534: ". This allows the instructor and the TA to respond to them with priority.
  - **Only** emails sent to `hxiao@stat.rutgers.edu` are guaranteed to be read.
- Notes
  - The lectures will be based on the combination of the textbook, notes, recommended references and additional materials prepared by the instructor.
  - All students are required to read the textbook, required notes and additional materials.

2. Syllabus (tentative)

Here is a tentative syllabus. The topics may not be covered in exactly the same order as listed. Adjustments will be made depending on the progress.

- Introduction and Review (Chapter 1, 2 of ESL2, Chapter 2, 3 of ISL2)
- Convex and non-convex optimization (Various places of CO)
- Support vector machine (Chapter 12 of ESL2)
- Kernel methods (Chapter 14 of ESL2, Chapter 12 of PRML)
- Graphical models (Chapter 17 of ESL2, Chapter 8 of PRML)
- Boosting and additive trees (Chapter 10 of ESL2)
- Deep Learning (Chapter 11 of ESL2, Chapter 10 of ISL2)

* Multiple testing (Chapter 13 of ISL2)

* Analysis of matrix and tensor data (Additional materials will be provided)

3. Project Guideline

Project is to be carried out by a team of no more than **three** investigators. You can choose to do (but not limited to) one of the following: (i) finding an interesting dataset, raising and answering meaningful questions; (ii) reading a journal/conference article, and reproducing its numerical results; and (iii) summarize and report the methodology/algorithms of some learning method (e.g. ADMM, XGBoost, matrix completion, transformer).

The **project proposal** needs to include what you plan to do, why it is important and meaningful, what kind of data you are going to use and a list of possible methodologies you plan to use. The **intermediate report** needs to include data description, preliminary analysis, the methodologies you are using, and the results you expect to get. The **presentation** is limited to 10–15 minutes (depending on how many groups there will be), describing the background, major methods, main findings, and a discussion about their limitations or what else can be done. The **final report** is limited to 5 pages (not including additional tables, figures, codes, outputs, references etc). It should be written in the format of a scientific paper, with an abstract, an introduction, main sections on methodologies and findings, and a conclusion.

Important dates: proposal due on **November 02 2022**, intermediate report due on **Novermber 30**, final report due on **TBA**. Please submit the proposal and intermediate/final reports in **PDF** format on Canvas, and the following additional items for the final report **through Canvas**: (i) a **PDF** file of the final report whose name should be "LastName1_LastName2_LastName3_report.pdf", (ii) data set used for the project, which can be directly read by `R`, (iii) `R` code (please make sure that by running the code, all the results in the report can be reproduced), and (iv) **PDF** slides of your presentation (with the name "LastName1_LastName2_LastName3_slides.pdf").