Course Overview

He He

CDS, NYU

January 28, 2021

Contents

Logistics

Course Staff

- Instructors:
  - He He (CS/CDS): most lectures
  - Marylou Gabrie (CDS): 2 lectures and assignments
  - Ming Liao (NYUSH): in-person sections in Shanghai
- Section leaders:
  - Xiangyun Chu, Jatin Khilnani, Xiaocheng Li, Haresh Rengaraj Rajamohan, Daeyoung Kim
- Graders:
  - Artie Shen, Daeyoung Kim, Xiangyun Chu, Xiaocheng Li, Aniket Bhatnagar, Udit Arora

Logistics

- Class webpage: https://nyu-ds1003.github.io/spring2021
  - Course material will be distributed on the website
- Piazza: https://piazza.com/class/ki0j413ace3bn
  - **All class announcements via Piazza**
  - Ask all questions on Piazza
- Gradescope: entry code **ZRWBGD**
  - Sign up yourself for assignment submission
- Office Hours:
  - All office hours will be on Zoom
  - See course calender for details: https://tinyurl.com/y23zkucf

Evaluation

- 7 assignments ($1 \times 4\% + 6 \times 6\% = 40\%$)
- Two tests (60%)
  - Midterm Exam (30%) in Week 8 (March 23rd), covering material up to Week 7
  - Final Exam (30%), time TBD, covering all material
- These scores determine "class rank".
- Typical grade distribution: A (40%), A- (20%), B+ (20%), B (10%), B- (5%), <B- (5%)

Homework

- Assignment 0: Help you get farmiliar with the format (not submitted or graded)
- First assignment out now – due on **Feb 10**
- Submit through Gradescope as a **PDF document**
- Late policy: Assignments are accepted up to **48 hours** late (see more details on website)
- Collaboration is fine, but
  - Write up solutions and code on your own
  - List names of who you talked to about each problem

Exams (60%)

- Exams will be submitted through Gradescope (similar to assignments)
- Start within **24 hours** once it's released
- Submit in **2 hours** once started
- Typesetting or handwritten, but must be submitted in **PDF**
- No collaboration is allowed
- Exams from previous years will be posted

Prerequisites

- DS-GA 1001: Introduction to Data Science
- DS-GA 1002: Statistical and Mathematical Methods
- Math
  - Multivariate Calculus
  - Linear Algebra
  - Probability Theory
  - Statistics
  - [Preferred] Proof-based linear algebra or real analysis
- Python programming (numpy)

Course Overview and Goals

Syllabus (Tentative)

13 weeks of instruction + 1 week midterm exam

- 2 weeks: introduction to **statistical learning theory**, **optimization**
- 2-3 weeks: **Linear** methods for binary classification and regression (also **kernel methods)**
- 2 weeks: **Probabilistic models**, **Bayesian** methods
- 1 week: **Multiclass** classification and introduction to **structured prediction**
- 3-4 weeks: **Nonlinear** methods (**trees**, **ensemble** methods, and **neural networks**)
- 2 weeks: **Unsupervised** learning: **clustering** and **latent variable** models
- See tentative schedule on the webpage
- Applications and practical algorithms may be covered in labs

High Level Goals of the Class

- Learn fundamental building blocks of machine learning
- Goal is to start seeing

  fancy new method A "is just" familiar thing B + familiar thing C + tweak D

  - SVM "**is just**" ERM with hinge loss with ℓ_2 regularization
  - Pegasos "**is just**" SVM with SGD with a particular step size rule
  - Random forest "**is just**" bagging with trees, with an interesting tweak on choosing splitting variables

Level of the Class

- We will learn how to build all ML algorithms **from scratch** – no ML libraries, just numpy.
- Once we have built it from scratch once, we can use the sklearn version.