

Course Overview

Ravid Shwartz-Ziv

CDS, NYU

Jan 24, 2023

Contents

1 Logistics

2 Course Overview and Goals

Logistics

Course Staff

- Instructors:
 - Ravid Shwartz Ziv
 - Mengye Ren
- Section leaders:
 - Colin Wan
 - Ying Wang
 - Yanlai Yang
- Graders:
 - Xiaojing Fan • Richard-John Lin
 - Junze Li • Richard-John Lin
 - Ying Wang • Jerry Xue
 - Frances Yuan

- Class webpage: <https://nyu-ds1003.github.io/spring2023>
 - Course materials (lecture slides, homework assignments) will be made available on the website
- Announcements via Brightspace
- Discussion / questions on CampusWire: <https://campuswire.com/c/G0F20206F/feed>
- Sign up to Gradescope to submit homework assignments (entry code **DJZ86D**)
- Office Hours: By appointment
 - Ravid: Tuesday 9:00-10:00 am; Mengye: Tuesday 2:00-3:00 pm
 - Colin: Monday 1:00-2:00 pm; Ying :Wednesday 6:00 pm - 7:00 pm; Yanlai : Wednesday 1:00pm-2:00 pm; Room 204, 60 5th Ave;

Assessment

- 7-8 assignments (40%)
- Two tests (60%)
 - Midterm Exam (30%)
 - Final Exam (30%)
- Extra credits (2%) answer other students' questions in a substantial and helpful way on Campuswire

- Submit through Gradescope as a **PDF document**
- Late policy: You have seven late days in total which can be used throughout the semester without penalty (see more details on website).
- You can collaborate with other students on the homework assignments, but please:
 - Write up the solutions and code on your own;
 - And list the names of the students you discussed each problem with.

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
- More detailed schedule on the course website (still subject to change)
- Certain applications and practical algorithms may be covered in the labs

The high level goals of the class

- Our focus will be on the fundamental building blocks of machine learning
- ML methods have a lot of names; our goal is for you to notice that
 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 forests “are just” bagging with trees, with a different approach to choosing splitting variables

The level of the class

- We will learn how to implement each ML algorithm **from scratch** using numpy alone, without any ML libraries.
- Once we have implemented an algorithm from scratch once, we will use the sklearn version.