STA 250: Theoretical Foundations for Machine Learning

Lecture 1: Introduction and Overview

Dogyoon Song

Spring 2025, UC Davis

Agenda

- Course overview
- Logistics
- Supervised learning[1]

[1] Suggested reading: Bach, Chapter 2 & Ma, Chapter 1

Course objectives

Goal: ~~Fully explain how and why machine learning/deep learning work~~

Modest/realistic goals:

- Learn about fundamental tools and frameworks for reasoning about ML & Optimization
- Learn about what these can say about DL, and where they fall short
- Gain experience and strengthen ability to
  - critically read and assess (recent) research publications
  - identify and formulate research questions/approaches to pursue throughout the quarter

Course logistics

- Prerequisites
- Texts and resources
- Online plaforms
- Course contents & organization
- Grading criteria
- Course policies

See syllabus for details and additional information

Supervised learning

Goal: make good prediction on *new, unseen* future data ("test data")

Setup: We are given the following in the usual setup

- An unknown distribution μ on $\mathcal{X} \times \mathcal{Y}$
- A training sample $\mathcal{D}_n(\mu) = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $(x_i, y_i) \sim \mu$
- A loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

Given these,

- we want to design a learning algorithm $\texttt{Alg} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$; $\texttt{Alg} : \mathcal{D}_n \mapsto f$
- we care about the *population risk* (=expected risk)

$$R_\mu(f) := \mathbb{E}_{(x,y)\sim\mu}\left[\ell(f(x), y)\right]$$

Want: Design $\texttt{Alg}$ that learns from a "small" amout of data and achieves low risk

Bayes predictor and Bayes risk

For now, suppose we have access to μ

Q: What is the best f we can hope for?

By the law of total expectation,

$$R(f) = \mathbb{E}\left[\ell(f(x), y)\right] = \mathbb{E}\left[\mathbb{E}\left[\ell(f(x), y) \mid x\right]\right]$$

Thus, the minimizer of $R(f)$ can be obtained by minimizing $\mathbb{E}\left[\ell(f(x), y) \mid x\right]$ pointwisely

Definition

A map $f_* : \mathcal{X} \to \mathcal{Y}$ is a *Bayes predictor* if

$$f_*(x') \in \arg\min_{z \in \mathcal{Y}} \mathbb{E}\left[\ell(z, y) \mid x = x'\right], \quad \forall x' \in \mathcal{X}.$$

The *Bayes risk* R^* is the risk of any Bayes predictor, and is equal to

$$R^* = \mathbb{E}_{x'}\left[\inf_{z \in \mathcal{Y}} \mathbb{E}\left[\ell(z, y) \mid x = x'\right]\right].$$

Examples of Bayes predictors and excess risk

Examples

- Regression with square loss: $f_*(x') = \mathbb{E}[y \mid x = x']$
- Classification with 0-1 loss: $f_*(x') = \arg\max_z \Pr(y = z \mid x')$

Definition

The *excess risk* of $f : \mathcal{X} \to \mathcal{Y}$ is $R(f) - R^*$.

Goal (formally restated): We want to find `Alg` such that the excess risk

$$R\left(\texttt{Alg}(\mathcal{D}_n)\right) - R^*$$

is "small," where $\mathcal{D}_n$ is a *random* training dataset. However, "small" in what sense?

Measures of performance

Suppose μ is fixed for now

- $\texttt{Alg}$ is consistent in expectation (w.r.t. μ) if

$$\mathbb{E}\left[R\left(\texttt{Alg}(\mathcal{D}_n)\right)\right] - R^* \to 0 \quad \text{as } n \to \infty.$$

- $\texttt{Alg}$ is probably approximately correctly (PAC) consistent (w.r.t. μ) if for any $\epsilon > 0$, there exists a sequence δ_n ($\to 0$ as $n \to \infty$) such that

$$\Pr\left(R\left(\texttt{Alg}(\mathcal{D}_n)\right) - R^* \leq \epsilon\right) \geq 1 - \delta_n.$$

We may want consistency over a class of problems (not for a single μ, but all $\mu \in \mathcal{M}$):

- $\texttt{Alg}$ is universally consistent (over $\mathcal{M}$) if[2]

$$\sup_{\mu \in \mathcal{M}} \left\{\mathbb{E}\left[R\left(\texttt{Alg}(\mathcal{D}_n)\right)\right] - R^*\right\} \to 0 \quad \text{as } n \to \infty.$$

[2] Be careful with the order of quantifiers in universal consistency; also, see "no free lunch theorem"

Until next lecture

- Complete the “Homework 0” for your self-assessment ASAP if you haven’t yet
- Start exploring project ideas
- Suggested reading for next lecture: empirical risk minimization
  - Bach, Chapter 4
  - Ma, Chapters 2 & 4
  - For mathematical preliminaries, see also Bach, Chapter 1 & Ma, Chapter 3