Modern Topics on Statistical Learning Theory Spring 2023

Lecture 1 — Basics of Machine Learning

Prof. Qi Lei *Scribe: Qi Lei*

1 Overview

Theme. In this lecture, we will talk about the theoretical foundation of many machine learning tasks, with a concentration on weakly-supervised learning.

Importance. We want to understand how/why machine learning works.

- For instance, a common scenario: after you train a giant model, and see it doesn't transfer to a smaller dataset. How do you know what went wrong? After you learn the course, you get to know there are roughly three possibility:

$$\left.\begin{array}{l}\to \text{knowledge doesn't transfer}\\ \to \text{not enough samples}\end{array}\right\} \qquad \text{fundamental statistical issue}$$

$$\to \text{computational issue} \qquad \} \text{ optimization issue}$$

2 Structure and logic of this course

ML (AI in general) is important and we have seen lots of incredible achievements.

$$\boxed{\text{Chatbot}} \quad \boxed{\text{Medical Diagnose}} \quad \boxed{\text{Alpha-Go}}$$

As we learn this course, we no longer care much about the data modality as in these examples, while we will view data as high-dimensional samples that are generated from a certain distribution.

Basics of ML. Machine learning usually consists of the following elements:

- data (MNIST, CIFAR10, IMAGENET)
- loss function (measuring the difference between prediction and true labels, l2, cross-entropy, etc)
- model (linear/affine, kernel, neural network)
- optimization ((stochastic) gradient descent, Adam, Adagrad, etc)

(Many implementation details are omitted: cross-validation, hyper-parameter tuning, regularization, etc)

3 Supervised learning

Input data: $\{(\underbrace{x_i}_{\text{feature}}, \underbrace{y_i}_{\text{label}}\}_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$

Goal: Find prediction function $f_\theta : \mathbb{R}^d \to \mathbb{R}$, (Or $\mathbb{R}^d \to \mathbb{R}^c$) for multi label classification) such that

$$f_\theta(x_i) \approx y_i, \forall i.$$

In this course, we are interested in parametric models.

Empirical Risk Minimization (ERM).

$$\min_\theta \frac{1}{n} \sum_{i=1}^n \ell(f_\theta(x_i), y_i) \tag{1}$$

$$\xrightarrow{\text{concentrates to}} \mathop{\mathbb{E}}_{(x,y)\sim P_{X,Y}} \ell(f_\theta(x), y). \tag{2}$$

Hope: prediction function learned from (1) can perform well on (2).

4 Scope of this course

Traditional types of machine learning.

$$\begin{cases} \text{Unsupervised learning} \\ \text{Supervised learning} \\ \text{Reinforcement learning} \end{cases}$$

Spectrum between supervised and unsupervised learning.

More labeled data involved $\downarrow$

- Unsupervised learning
- Self-supervised learning
- Meta-learning
- Domain generalization
- Domain adaptation
- Semi-supervised learning
- Supervised learning

Table distinguishing different types of learning.

Notation. L : labeled data; U: unlabeled data; S: source data; T: target data; e: number of environments/tasks.

learning task	data that model is trained on	data that your model is tested on
supervised learning	L	U
semi-supervised learning	L, U	U
domain adaptation	L^S, U^T	U^T
domain generalization	$L_1^S, L_2^S, \cdots, L_e^S$	U^T
meta-learning	$L_1^S, L_2^S, \cdots, L_e^S, \underbrace{L^T}_{\text{few-shot}}$	U^T
self/un-supervised learning	$U^S, \underbrace{L^T}_{\text{few-shot}}$	U^T
reinforcement learning	source environment	target environment

5 Theoretical groundings of ML algorithms in general.

objective functions:

e.g. sup.L: $\sum_i \ell(f_\theta(x_i), y_i)$ or unsup.L: $\|M - XX^T\|_F^2$

Theoretical understanding requires studying the two aspects:

$\Leftarrow \Rightarrow$

statistics $\downarrow$ whether optimal solution generalizes

optimization $\downarrow$ whether our algorithms find global minima

$\Downarrow$

Together they form learning theory.