CS 183: Fundamentals of Machine Learning

Lectures by Yaron Singer
Notes by Maggie Wang

Harvard University, Spring 2020

Course textbook:
Understanding Machine Learning: From Theory to Algorithms by Shai Shalev-Shwartz and Shai Ben-David.

Contents

Lecture 1: Prelude - 1/30

The Statistical Learning Framework

Learner's input:

- **Domain set**: Arbitrary set $\mathcal{X}$ that we wish to label. Represented by a vector of features. Domain points: instances, $\mathcal{X}$: instance space.
- **Label set**
- **Training data**
- **The learner's output**
- **A simple data-generation model**
- **Measure of success**: error of a prediction rule, $h : \mathcal{X} \to \mathcal{Y}$ is the probability of randomly choosing an ex. x for which $h(x) \neq f(x)$:

$$L_{\mathcal{D},f}(h) = \mathbb{P}_{x\sim\mathcal{D}}[h(x) \neq f(x)] = \mathcal{D}(\{x : h(x) \neq f(x)\})$$

Empirical Risk Minimization

Training error:

$$L_S(h) = \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}$$

Empirical Risk Minimization (ERM): coming up with a predictor h that minimizes $L_S(h)$

Overfitting

Overfitting: h fits training data "too well"

$$h_S(x) = \begin{cases} y_i & \text{if } \exists i \in [m] \text{ s.t. } x_i = x \\ 0 & \text{otherwise.} \end{cases}$$

Learning with Inductive Bias

Finite hypothesis classes

Definition 1 The Realizability Assumption
There exists $h^* \in \mathcal{H}$ s.t. $L_{(\mathcal{D},f)}(h^*) = 0$

Lemma 1 Union Bound
For any two sets A, B and a distribution $\mathcal{D}$ we have

$$\mathcal{D}(A \cup B) \leq \mathcal{D}(A) + \mathcal{D}(B)$$

Corollary 1 Let $\mathcal{H}$ be a finite hypothesis class. Let $\delta \in (0, 1)$ and $\epsilon > 0$ and let m be an integer that satisfies $m \geq \frac{\log(|\mathcal{H}|\delta)}{\epsilon}$.
Then, for any labeling function, f, and for any distribution, $\mathcal{D}$, for which the realizability assumption holds (that is, for some $h \in \mathcal{H}, L_{(\mathcal{D},f)}(h) = 0$) with probability of at least $1-\delta$ over the choice of an i.i.d. sample S of size m, we have that for every ERM hypothesis, h_S, it holds that

$$L_{(\mathcal{D},f)}(h_S) \leq \epsilon$$

For a sufficiently large m, the ERM$_{\mathcal{H}}$ rule over a finite hypothesis will be *probably* (with confidence $1-\delta$) *approximately* (up to an error of ϵ) correct.

Lecture 2: PAC Learnability - 2/6

PAC learnability

Agnostic PAC learnability

Prove learnability via uniform convergence