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 **2**
- The Statistical Learning Framework 2
- Empirical Risk Minimization 2
  - Overfitting 2
- Empirical Risk Minimization with Inductive Bias 2
  - Finite hypothesis classes 2

Lecture 2: PAC Learnability - 2/6 **3**
- PAC learnability 3
- Agnostic PAC learnability 3
- Prove learnability via uniform convergence . . . 3

Lecture 1: Prelude - 1/30

The Statistical Learning Framework

Learner's input:

- **Domain set**: Set $\mathcal{X}$ that we wish to label. Represented by a vector of features. Domain points: instances, $\mathcal{X}$: instance space.
- **Label set**: Set $\mathcal{Y}$ of possible labels
- **Training data**: $S = ((x_1, y_1) \dots (x_m, y_m))$, finite sequence of pairs in $\mathcal{X} \times \mathcal{Y}$. Training examples / training set.
- **The learner's output**: prediction rule, $h : \mathcal{X} \to \mathcal{Y}$. Predictor, hypothesis, classifier.
- **A simple data-generation model**: each pair in the training data S is generated by sampling a point x_i according to $\mathcal{D}$ (probability distribution over $\mathcal{X}$ by $\mathcal{D}$) and then labeling it by f.
- **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 $h(x) \neq f(x)$:

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

Generalization error, the risk, the true error of h.

Empirical Risk Minimization

Training error / empirical error / empircal risk - error the classifier incurs over the training sample:

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

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

Overfitting

Overfitting: h fits training data "too well"

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

Empirical Risk Minimization with Inductive Bias

Apply ERM over a restricted search space (hypothesis class $\mathcal{H}$), thus biasing it towards a particular set of predictors. Such restrictions are called an inductive bias.

$$\text{ERM}_{\mathcal{H}}(S) \in \underset{h \in \mathcal{H}}{\arg\min}\, L_s(h)$$

Finite hypothesis classes

$$h_S \in \underset{h \in \mathcal{H}}{\arg\min}\, L_S(h)$$

Definition 2.1: The Realizability Assumption
There exists $h^* \in \mathcal{H}$ s.t. $L_{(\mathcal{D},f)}(h^*) = 0$.

This assumption implies that with probability 1 over random samples, S, where the instances are sampled according to D and are labeled by f, we have $L_S(h^*) = 0$.

The i.i.d. assumption: $S \sim \mathcal{D}^m$, where m is the size of S, and $\mathcal{D}^m$ denotes the probability over m-tuples induced by applying $\mathcal{D}$ to pick each element of the tuple independently of the other members of the tuple.

Theorem 2.3:
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 $\text{ERM}_{\mathcal{H}}$ rule over a finite hypothesis will be *probably* (with confidence $1-\delta$) *approximately* (up to an error of ϵ) correct.

Proof:
δ is probability of getting a nonrepresentative sample, and $(1-\delta)$ is the confidence parameter of our prediction.
ϵ is the accuracy parameter. Event $L_{(\mathcal{D},f)}(h_S) > \epsilon$ is failure of the learner, while if $L_{(\mathcal{D},f)}(h_S) \leq \epsilon$ the output of the algorithm is an approximately correct predictor.

Let $S|_x = (x_1, \dots, x_m)$ be the instances of the training set.
We would like to upper bound $\mathcal{D}^m(S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon)$.
Set of "bad" hypotheses: $\mathcal{H}_B = \{h \in \mathcal{H} : L_{(\mathcal{D},f)}(h) > \epsilon\}$.
Set of misleading examples: $M = \{S|_x : \exists h \in \mathcal{H}_B, L_S(h) = 0\}$.
For every $S|_x \in M$, there is a "bad" hypothesis, $h \in \mathcal{H}_B$ that looks like a "good" hypothesis on $S|_x$.
The event $L_{(\mathcal{D},f)}(h_S) > \epsilon$ can only happen if our sample is in the set of misleading samples, M:

$$\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\} \subseteq M$$

We can rewrite M as $M = \bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}$.
$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}) \leq \mathcal{D}^m(M) = \mathcal{D}^m(\cup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\})$.
Upper bound right-hand side using union bound.

Lemma 2.2: Union Bound
For any two sets A, B and a distribution $\mathcal{D}$ we have

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

$$\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon) \leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S|_x : L_S(h) = 0\})$$

$$\begin{aligned} \mathcal{D}^m(\{S|_x : L_S(h) = 0\}) &= \mathcal{D}^m(\{S|_x : \forall i, h(x_i) = f(x_i)\}) \\ &= \prod_{i=1}^{m} \mathcal{D}(\{x_i : h(x_i) = f(x_i)\}) \end{aligned}$$

etc.

Lecture 2: PAC Learnability - 2/6

PAC learnability

Agnostic PAC learnability

Prove learnability via uniform convergence