CSCI699: Topics in Learning and Game Theory

Lecture 1

Lecturer: Ilias Diakonikolas *Scribes: Li Han*

Learning problems usually come in two types.

1. Unsupervised: Discover hidden structure in a set of unlabeled data points, e.g., a set of points in $\mathbb{R}^n$.
2. Supervised: Make predictions based on labeled data

Machine learning in some sense is the automatic extraction of useful information from raw data. Examples include:

1. Classification, clustering;
2. Text categorization, fraud detection;
3. Web search, prediction(weather, finance, etc);
4. Automatically writing complex software;

1 Basics of Learning Theory

In learning theory, we specify a mathematical model for learning and prove formal results within that model. To do so, we need to answer the following questions.

1. **Who is learning:** Since machine learning is automatic, we usually assume the learner is a computer program, whose goal is to minimize sample complexity and running time;
2. **What are we learning:** A hypothesis or function, e.g., a classfication rule.
3. **How does learning obtain raw data**: The learner obtains information from given examples, which are data-label pairs $(x, f(x))$. Such example could be given to the learner in a random or malicious way.
4. **How to measure accuracy**: We often assume the data is labeled according to some ground truth (or target hypothesis) and then measure the distance between the output hypothesis and the ground truth.

Each of these questions have multiple answers and that give raise to different learning models. We will start with simplest setting: statistical learning framework where we want to learn a binary classification rule with random samples..

2 Statistical Learning Framework

Let $\mathcal{X}$ be the domain or instance space (i.e., set of object to label). For example, $\mathcal{X}$ could be $\{0,1\}^d$ or $\mathbb{R}^d$. Let $\mathcal{Y}$ be the label set. In binary classification, $\mathcal{Y} = \{0,1\}$.

The learner receives its training data $S = \{(x_1, y_1), (x_2, y_2), \cdots, , (x_m, y_m)\} \subseteq (\mathcal{X} \times \mathcal{Y})^m$. Note that $m = |S|$ is the size of the training data or the number of samples the learner receives.

Once the learner performs some computation on the training data, it needs to output a prediction rule $h : \mathcal{X} \to \mathcal{Y}$ (also called predictor, hypothesis, classifier, etc.), which predicts the labeling of unseen domain points.

2.1 Data Generation Model

In the most basic setting, we assume each x_i in the training data is drawn i.i.d. from some distribution $\mathcal{D}$ over $\mathcal{X}$ and $y_i = f(x_i)$ where f is the *unknown* target hypothesis (or ground truth). Note that both $\mathcal{D}$ and f are unknown to the learner as it only interacts with the training data sampled from $\mathcal{D}$.

The error of a hypothesis h (with respect to $\mathcal{D}$ and f) is defined as

$$L_{\mathcal{D},f} = \Pr_{x \sim \mathcal{D}}[h(x) \neq f(x)].$$

Given the training data $S = \{(x_i, y_i)\}_{i=1}^m$, one very natural algorithm is called Empirical Risk Minimization (ERM). The empirical risk (or training error) of a hypothesis h is defined as

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

Let $\mathcal{U}$ be the uniform distribution over S (also called the empirical distribution). Then $L_S(h)$ can also be written as $Pr_{x \sim \mathcal{U}(S)}[h(x) \neq f(x)]$.

ERM just minimizes the empirical risk. So ERM will output the following hypothesis h_S

$$h_S = \arg\min_h L_S(h) \tag{1}$$

2.2 Overfitting

Although very natural, ERM might fail miserably without being careful. Consider the following example of learning rectangles. Let B be a big rectangle with area 2 and C be a small rectangle contained inside B with area 1. Let $\mathcal{D}$ be the uniform distribution over rectangle B. Let the target hypothesis f be the indicator function of rectangle C.

Given any training data S, the following hypothesis is consistent with ERM [1]:

$$h_S(x) = \begin{cases} y_i \text{ if there exists } i \text{ such that } x_i = x \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

No matter how large m is, the error of h_S is always $\frac{1}{2}$ as it *almost surely* labels every point in B as 0, while the target hypothesis f labels half of them as 1. Intuitively, overfitting occurs when the hypothesis fits the training data too well. We need to apply ERM over a restricted search space to obtain any meaningful results.

2.3 Hypothesis Class

Let H be a family of hypotheses. The restricted ERM rule ERM_H now finds $h \in H$ with smallest empirical error. By restricting the learner to choose h from H, we bias the learner towards a particular set of solution. The choice of H is usually based on domain knowledge of the learning problem.

3 Finite Hypothesis Class

In this section, we will prove that ERM works on a finite hypothesis class once it receives enough training samples. Recall in ERM_H, we have $h_S = \arg\min_{h \in H} L_S(h)$ where L_S is the empirical loss.

Realizability assumption: There exists $h^* \in H$ such that $L_{\mathcal{D},f}(h^*) = 0$. We call h^* the consistent hypothesis.

Theorem 1. *If H is finite, then ERM_H achieve errors ϵ and success probability $1 - \delta$ with $m \geq \frac{\log \frac{|H|}{\delta}}{\epsilon}$ samples.*

Proof. This implies that with probability 1, $L_S(h^*) = 0$ and $L_S(h_s) \leq L_S(h^*) = 0$. We want to prove with probability at least $1 - \delta$, $L_{\mathcal{D},f}(h_S) \leq \epsilon$. This is equivalent to upper bounding $\Pr_{S|\mathcal{X} \sim \mathcal{D}^m}[L_{\mathcal{D},f}(h_S) > \epsilon]$.

[1] In some sense, this is memoization: just output what you see.

Let $H_B = \{h \in H : L_{\mathcal{D},f}(h) > \epsilon\}$ be the set of bad hypotheses. Let $M = \{S|\mathcal{X} : \exists h \in H_B \text{ s.t. } L_S(h) = 0\}$ be the set of misleading examples that fools ERM to output a bad hypothesis. Note that $M = \cup_{h \in H_B}\{S|\mathcal{X} : L_S(h) = 0\}$, so, by union bound, $\Pr[M] \leq \sum_{h \in H_B} \Pr[L_S(h) = 0]$.

Now fix any $h \in H_B$ with $L_{\mathcal{D},f}(h) > \epsilon$, $\Pr[L_S(h) = 0] = \prod_{i=1}^{m} \Pr[h(x_i) = f(x_i)] \leq (1-\epsilon)^m \leq e^{-m\epsilon}$. So $\Pr_S[L_{\mathcal{D},f}(h_S) > \epsilon] \leq |H|e^{-m\epsilon}$. By equating δ and $|H|e^{-m\epsilon}$, we have $m \geq \frac{\log(|H|/\epsilon)}{\epsilon}$. □

4 PAC Learnability

Definition 2 (PAC-Learnability)**.** *H is PAC-learnable if $\exists m_H : [0,1]^2 \to \mathbb{N}$ and learner such that for all ϵ, δ, for all $\mathcal{D}$ over $\mathcal{X}$ and target concept $f : \mathcal{X} \to \mathcal{Y}$, the learner outputs a hypothesis h with $L_{\mathcal{D},f}(h) \leq \epsilon$ with probability at least $1-\delta$ after at least $m_H(\epsilon, \delta)$ samples.*

To obtain a more general model of learning, we remove the realizability assumption (thus agnostic) and go beyong binary labels.

In particular, let $\mathcal{Z}$ be the new domain space and let $l : H \times \mathcal{Z}$ be the new loss function. For prediction problems, let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.

Definition 3 (Agnostic PAC-Learnability)**.** *H is agnositcally PAC-learnable w.r.t. set $\mathcal{Z}$ and loss function $l : H \times \mathcal{Z} \to \mathbb{R}$ if $\exists m_H : [0,1]^2 \to \mathbb{N}$ and learner such that for all ϵ, δ, for all $\mathcal{D}$ over $\mathcal{Z}$, the learner outputs a hypothesis h with $L_{\mathcal{D}}(h) \leq \min_{h' \in H} L_{\mathcal{D}}(h') + \epsilon$ with probability at least $1-\delta$ after at least $m_H(\epsilon, \delta)$ samples.*

5 Learning via Uniform Convergence

Theorem 4. *Let H be finite and $l : H \times \mathcal{Z} \to [0,1]$ be a* bounded *loss function, then ERM_H is an agnostic learner if $m \geq O(\frac{\log(|H|/\delta)}{\epsilon^2})$.*

Note that the dependence on ϵ is now $\frac{1}{\epsilon^2}$ instead of $\frac{1}{\epsilon}$ as in the realizable case. We will prove the above theorem via uniform convergence, but let us first define representative sample.

Definition 5 (Representative sample)**.** *A training set S is called ϵ-representative if for all $h \in H$ we have $|L_D(h) - L_S(h)| \leq \epsilon$.*

Lemma 6. *If S is $\epsilon/2$-representative, then ERM satisfies $L_D(h_S) \leq \min_{h \in H} L_D(h) + \epsilon$.*

Proof. Let $h^* = \arg\min_{h \in H} L_D(h)$.

$$\begin{aligned} L_D(h_S) &\leq L_S(h_S) + \epsilon/2 && (3) \\ &\leq L_S(h^*) + \epsilon/2 && (4) \\ &\leq L_D(h^*) + \epsilon && (5) \end{aligned}$$

The first and third inequality follows from S being ϵ-representative. The second inequality follows from the fact that h_S is ERM. □

Theorem 7 (Uniform convergence)**.** *With high probability, S is ϵ-representative if $m \geq O(\frac{\log(|H|/\delta)}{\epsilon^2})$.*

Proof. Fix any $h \in H$, with high probability $L_S(h)$ approximates $L_D(h)$ by Hoeffding bound. Then apply union bound over all hypotheses in H. □