Statistical Learning Theory

Jayadev Naram

Contents

1 Learning Framework

- ***(Input).*** A finite sequence $S = ((x_1, y_1) \dots (x_m, y_m))$ called as *training data* is provided to the learner, where each instance called as *example* or *sample* $(x_i, y_i) \in X \times Y$ for some domain set X of data points and label set Y.
- ***(Output).*** A map $h : X \to Y$ called as a *predictor*, a *hypothesis*, or a *classifier* is the output of the learner. We use the notation $A(S)$ to denote the hypothesis that a learning algorithm A returns upon receiving the training sequence S.
- ***(Data-generation model).*** The instances are generated by some probability distribution $\mathcal{D}$ over X and the labels are assigned by some labeling function $f : X \to Y$, i.e., the label y of a randomly sampled data point $x \sim \mathcal{D}$ is $y = f(x)$.
- ***(Measure of success).*** The error of a hypothesis is the probability that it does not predict the correct label on a random data point generated by $\mathcal{D}$, i.e., for a hypothesis $h : X \to Y$, the *prediction error*, the *generalization error*, or the *risk* is defined to be

$$L_{\mathcal{D},f}(h) = \mathop{\mathbb{P}}_{x \sim \mathcal{D}}[h(x) \neq f(x)] = \mathcal{D}(\{x : h(x) \neq f(x)\}). \tag{1}$$

Remark 1.1. *The learner is blind to the underlying distribution $\mathcal{D}$ over X and to the labeling function f. The only way the learner can interact is through observing the training set.*

2 ERM with Finite Hypothesis Classes

Since the learner does not know $\mathcal{D}$ and f, the true error is not directly available. Instead one can compute the *training error* or *empirical risk* which is defined to be

$$L_S(h) = \frac{|\{(x_i, y_i) \in S : h(x_i) \neq y_i\}|}{m}. \tag{2}$$

Then, the ***Empirical Risk Minimization*** rule selects a hypothesis h_S that minimizes the empirical risk over the training samples S, i.e.,

$$h_S \in \operatorname*{argmin}_{h:X \to Y} L_S(h).$$

We show that ERM rule might lead to undesirable hypotheses.

Remark 2.1 (***Overfitting***). *Assume that $X = [0,1]^2$ and $Y = \{0,1\}$. Let the probability distribution $\mathcal{D}$ be uniform distribution over X and the labeling function f determines the label to be 1 for all data points in $[0, 1/2] \times [0, 1]$. Then note that area of region with labels 1 is $1/2$ and the total area of domain set is 1. Consider the following hypothesis:*

$$h_S(x) = \begin{cases} y_i & \text{if } (x_i, y_i) \in S \text{ and } x_i = x \\ 0 & \text{otherwise.} \end{cases}$$

Then we see that the empirical risk $L_S(h_S) = 0$, but the generalization error $L_{\mathcal{D},f}(h_S) = 1/2$ because h_S correctly labels all the negative (i.e., label 0) examples, but incorrectly labels all but finite positive (i.e., label 1) examples. The ERM rule can choose this hypothesis which performs no better than random guess.

To overcome overfitting, we provide the learner with some prior knowledge. We restrict the search space to a finite hypothesis class denoted by $\mathcal{H}$ and we apply ERM rule on it as before. Formally,

$$h_S \in \operatorname*{argmin}_{h \in \mathcal{H}} L_S(h),$$

where h_S is the ERM hypothesis. Additionally, we make the following simplifying assumption:

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

By this assumption, we have with probability 1 over random samples S that $L_S(h^*) = 0$. This also implies that for every ERM hypothesis h_S we have that $L_S(h_S) = 0$.

Proposition 2.2. *Let $\mathcal{H}$ denote a finite hypothesis class. Let $\delta \in (0,1)$ and $\epsilon > 0$ and let m be an integer that satisfies $m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$. If for a labeling function f, a distribution $\mathcal{D}$, and the hypothesis class $\mathcal{H}$, the realization assumption holds. Then 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, $L_{\mathcal{D},f}(h_S) \leq \epsilon$.*

Proof. We will show that, over the choice of i.i.d. sample S of size m, with probability not more than δ, there exists a hypothesis h_S that the ERM might select for which $L_{\mathcal{D},f}(h_S) > \epsilon$. Formally, we would like to upper bound

$$\mathcal{D}^m(\{S|_X : L_{\mathcal{D},f}(h_S) > \epsilon\}), \tag{3}$$

where $S|_X$ is the restriction of S to the domain X and $\mathcal{D}^m$ is the probability of choosing i.i.d. sample $S \sim \mathcal{D}$ of size m. Let $\mathcal{H}_B \subseteq \mathcal{H}$ be the set of all bad hypothesis of $\mathcal{H}$, i.e.,

$$\mathcal{H}_B = \{h \in \mathcal{H} : L_{\mathcal{D},f}(h) > \epsilon\}.$$

In addition, let M be the set of all misleading samples:

$$M = \{S|_X : \text{for some } h \in \mathcal{H}_B, L_S(h) = 0\},$$

in other words, for every $S|_X \in M$, there is a bad hypothesis $h \in \mathcal{H}_B$, that looks like a good hypothesis on $S|_X$. By realizability assumption, we have that $L_S(h_S) = 0$. Then the event that $L_{\mathcal{D},f}(h_S) > \epsilon$ for a sample S implies that $h_S \in \mathcal{H}_B$ and also $L_S(h_S) = 0$ so that $S \in M$. Therefore we have shown that

$$\{S|_X : L_{\mathcal{D},f}(h_S) > \epsilon\} \subseteq M = \bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}. \tag{4}$$

Hence,

$$\begin{aligned}
\mathcal{D}^m(\{S|_X : L_{\mathcal{D},f}(h_S) > \epsilon\}) &\leq \mathcal{D}^m(\cup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}) \\
&\overset{(*)}{\leq} \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S|_x : L_S(h) = 0\}) \\
&= \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S|_x : \forall i, h(x_i) = f(x_i)\}) \\
&\overset{(\dagger)}{=} \sum_{h \in \mathcal{H}_B} \prod_{i=1}^{m} \mathcal{D}(\{x_i : h(x_i) = f(x_i)\}) = \sum_{h \in \mathcal{H}_B} \prod_{i=1}^{m} (1 - L_{\mathcal{D},f}(h)) \\
&\leq \sum_{h \in \mathcal{H}_B} \prod_{i=1}^{m} (1-\epsilon) \leq \sum_{h \in \mathcal{H}_B} \prod_{i=1}^{m} e^{-\epsilon} = |\mathcal{H}_B| e^{-\epsilon m} \leq |\mathcal{H}| e^{-\epsilon m},
\end{aligned}$$

where $(*)$ is obtained by applying union bound and $(\dagger)$ is obtained by using i.i.d. assumption. Using the fact that $m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$, we have

$$\mathcal{D}^m(\{S|_X : L_{\mathcal{D},f}(h_S) > \epsilon\}) \leq |\mathcal{H}| e^{-\epsilon m} \leq \delta$$

which is as required. □

3 PAC Learnability

Definition 3.1 (***PAC Learnability***)**.** A hypothesis class $\mathcal{H}$ is ***P***robably ***A***pproximately ***C***orrect ***Learnable*** if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$, for every distribution $\mathcal{D}$ over X, and for every labeling function $f : X \to \{0,1\}$, if the realizablilty assumption holds w.r.t. $\mathcal{H}, \mathcal{D}, f$, then when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis h such that with probability of at least $1 - \delta$ (over the choice of the examples), we have $L_{\mathcal{D},f}(h) \leq \epsilon$.

The definition contains two approximation parameters which are justified by the following propositions.

Proposition 3.2. *No learner can guarantee to find h such that $L_{\mathcal{D},f}(h) = 0$ with probability 1 over random samples S.*

Proof. For $\epsilon \in (0,1)$ take $X = \{x_1, x_2\}$ and $\mathcal{D}(x_1) = 1 - \epsilon$, $\mathcal{D}(x_2) = \epsilon$. Then the probability not to see x_2 at all among m i.i.d. examples is $(1-\epsilon)^m \approx e^{-\epsilon m}$. So, if $\epsilon \ll 1/m$, then the learner is likely to not see x_2 at all, therefore making error in predicting x_2. □

Thus we need the accuracy parameter ϵ which determines how far output classifier can be from the optimal one (this corresponds to the "approximately correct"), i.e., for $\epsilon \in (0,1)$, the learner tries to find a classifier such that $L_{\mathcal{D},f}(h) \leq \epsilon$.

Proposition 3.3. *No learner can guarantee to find h such that $L_{\mathcal{D},f}(h) \leq \epsilon$, for $\epsilon \in (0,1)$ with probability 1 over random samples S.*

Proof. The input to the learner is randomly generated samples. Then there is always a (very small) chance to see the same example again and again. □

Thus, the confidence parameter δ indicates how likely the classifier is to meet that accuracy requirement (corresponds to the "probably" part of "PAC").

Remark 3.4. *The function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ determines the **sample complexity** of learning $\mathcal{H}$, i.e., how many examples are required to guarantee a probably approximately correct solution. Further note that if $\mathcal{H}$ is PAC learnable, there are many functions $m_{\mathcal{H}}$ that satisfy the requirements in Def. 3.1. Therefore, to be precise we define the sample complexity of learning $\mathcal{H}$ to be the "minimal function", in the sense that for any ϵ, δ, the number $m_{\mathcal{H}}(\epsilon, \delta)$ is minimal that satisfies the requirements of PAC learning with accuracy ϵ and confidence δ.*

Corollary 3.5. *Every finite hypothesis class is PAC learnable with sample complexity*

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(\mathcal{H}/\delta)}{\epsilon} \right\rceil.$$

The next example shows that there are infinite hypothesis classes that are PAC learnable.

Example 3.6 (***Axis aligned rectangles***)**.** Given $a_1 \leq b_1$, $a_2 \leq b_2$, define the axis aligned rectangle classifier $h_{(a_1,b_1,a_2,b_2)}$ by

$$h_{(a_1,b_1,a_2,b_2)}(x_1, x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq b_1 \text{ and } a_2 \leq x_2 \leq b_2 \\ 0 & \text{otherwise.} \end{cases}$$

Let $R(a_1, b_1, a_2, b_2)$ denote the rectangle corresponding to $h_{(a_1,b_1,a_2,b_2)}$. The class of all axis aligned rectangles in $\mathbb{R}^2$ is defined as

$$\mathcal{H}^2_{rec} = \{h_{(a_1,b_1,a_2,b_2)} : a_1 \leq b_1, a_2 \leq b_2\}.$$

Let $\mathcal{D}$ be some distribution over $\mathbb{R}^2$ and $R^* = R(a_1^*, b_1^*, a_2^*, b_2^*)$ be the rectangle that generates the labels such that $\mathcal{D}(R^*) = 1$. Let f be the labeling function corresponding to the rectangle R^*.

We show that $\mathcal{H}_{rec}^2$ is PAC learnable. Define a simple ERM algorithm as follows: return the smallest rectangle enclosing all the positive examples in the training set. It is easy to note that this is indeed an ERM because the empirical risk of the selected rectangle over the training sample is 0.

Let $R(S)$ be rectangle returned by ERM rule. Choose $\epsilon \in (0,1)$ and choose $a_1 \geq a_1^*, b_1 \leq b_1^*, a_2 \geq a_2^*, b_2 \leq b_2^*$ such that $\mathcal{D}(R_1) = \mathcal{D}(R_2) = \mathcal{D}(R_3) = \mathcal{D}(R_4) = \epsilon/4$ where $R_1 = R(a_1^*, a_1, a_2^*, b_2^*)$, $R_2 = R(b_1, b_1^*, a_2^*, b_2^*)$, $R_3 = R(a_1^*, b_1^*, a_2^*, a_2)$ and $R_4 = R(a_1^*, b_1^*, b_2, b_2^*)$.

Note that $L_{\mathcal{D},f}(h_S) = \mathcal{D}(R^* - R(S))$. If S contains (positive) examples in all of the rectangles R_1, R_2, R_3, R_4, then $R^* - R(S) \subseteq R_1 \cup R_2 \cup R_3 \cup R_4$ and consequently $L_{\mathcal{D},f}(h_S) \leq \epsilon$. Then we have

$$\begin{aligned}
\mathcal{D}^m(\{S|_X : L_{\mathcal{D},f}(h_S) > \epsilon\}) &\leq \mathcal{D}^m(\{S|_X : S|_X \cap R_i = \emptyset \text{ for some } i \in \{1,2,3,4\}\}) \\
&\leq \sum_{i=1}^{4} \mathcal{D}^m(\{S|_X : S|_X \cap R_i = \emptyset\}) \\
&\leq \sum_{i=1}^{4} (1-\epsilon/4)^m = 4(1-\epsilon/4)^m \leq 4e^{-m\epsilon/4}.
\end{aligned}$$

Setting $m \geq \frac{4 \log 4/\delta}{\epsilon}$ for some $\delta \in (0,1)$ we have

$$\mathcal{D}^m(\{S|_X : L_{\mathcal{D},f}(h_S) > \epsilon\}) \leq 4(1-\epsilon/4)^m \leq \delta.$$

4 General Learning Framework

In many practical problems the realizability assumption does not hold. Furthermore, it is more realistic not to assume that the labels are fully determined by features we measure on input elements. Following these lines, we generalize the learner framework:

- ***(Realistic Data-generation model).*** Let $\mathcal{D}$ be a joint probability distribution over $X \times Y$. The distribution $\mathcal{D}$ can be seen to be composed of two components: (i) a distribution $\mathcal{D}_X$ over unlabeled domain points (called the marginal distribution) and (ii) a conditional probability over labels for each domain point $\mathcal{D}((x,y)|x)$.
- ***(Revised True Error).*** We redefine the true error (or risk) of a hypothesis h to be

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

Proposition 4.1 (***Bayes optimal predictor***)**.** *Given any probability distribution $\mathcal{D}$ over $X \times \{0,1\}$, the Bayes predictor $f_{\mathcal{D}}$ defined as*

$$f_{\mathcal{D}}(x) = \begin{cases} 1, & \text{if } \mathbb{P}[y=1|x=x] \geq 1/2 \\ 0 & \text{otherwise,} \end{cases}$$

is optimal, in the sense that for every predictor $g : X \to \{0,1\}$, $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$.

Proof. Fix $x \in X$ and $\mathbb{P}[y=1|x=x] = \alpha_x$. Note that $\mathbb{P}[f_{\mathcal{D}}(x)=1|x=x]$ is 1 if $\alpha_x \leq 1/2$ and 0 if $\alpha_x < 1/2$, i.e., $\mathbb{P}[f_{\mathcal{D}}(x)=1|x=x] = \mathbb{I}_{[\alpha_x \geq 1/2]}$ and $\mathbb{P}[f_{\mathcal{D}}(x)=0|x=x] = \mathbb{I}_{[\alpha_x < 1/2]}$. Then

$$\begin{aligned}
\mathbb{P}[f_{\mathcal{D}}(x) \neq y|x=x] &= \mathbb{P}[f_{\mathcal{D}}(x)=1|x=x]\mathbb{P}[y=0|x=x] + \mathbb{P}[f_{\mathcal{D}}(x)=0|x=x]\mathbb{P}[y=1|x=x] \\
&= \mathbb{I}_{[\alpha_x \geq 1/2]}\mathbb{P}[y=0|x=x] + \mathbb{I}_{[\alpha_x < 1/2]}\mathbb{P}[y=1|x=x] \\
&= \mathbb{I}_{[\alpha_x \geq 1/2]}(1-\alpha_x) + \mathbb{I}_{[\alpha_x < 1/2]}\alpha_x \\
&= \min\{\alpha_x, 1-\alpha_x\}.
\end{aligned}$$

And similarly for g, we have

$$\begin{aligned}\mathbb{P}[g(x) \neq y | x = x] &= \mathbb{P}[g(x) = 1 | x = x]\mathbb{P}[y = 0 | x = x] + \mathbb{P}[g(x) = 0 | x = x]\mathbb{P}[y = 1 | x = x] \\ &= \mathbb{P}[g(x) = 1 | x = x](1 - \alpha_x) + \mathbb{P}[g(x) = 0 | x = x]\alpha_x \\ &\geq \min\{\alpha_x, 1 - \alpha_x\}(\mathbb{P}[g(x) = 1 | x = x] + \mathbb{P}[g(x) = 0 | x = x]) \\ &= \min\{\alpha_x, 1 - \alpha_x\}.\end{aligned}$$

The statement follows now due to the fact that the above is true for every $x \in X$. □

We make a further generalization in the measure of success.

- ***(Generalized Loss function).*** Given any set $\mathcal{H}$ (that plays the role of hypotheses) and some domain Z, the loss functions are functions $l : \mathcal{H} \times Z \to \mathbb{R}_+$. The true error or risk of a hypothesis $h \in \mathcal{H}$ is defined as

  $$L_{\mathcal{D}}(h) \equiv \mathbb{E}_{z \in \mathcal{D}}[l(h, z)].$$

  Similarly, we define the empirical risk to be the expected loss over a given sample $S = (z_1, \ldots, z_m) \in Z^m$, i.e.,

  $$L_S(h) \equiv \frac{1}{m} \sum_{i=1}^{m} l(h, z_i).$$

Example 4.2 ($\mathbf{0-1}$ *loss*). The random variable z ranges over the set of pairs $X \times Y$ and the loss function is

$$l_{0-1}(h, (x, y)) = \begin{cases} 0 & \text{if } h(x) = y \\ 1 & \text{if } h(x) \neq y. \end{cases}$$

5 Agnostic PAC Learnability

Definition 5.1 (*Agnostic PAC Learnability*). A hypothesis class $\mathcal{H}$ is agnostic PAC learnable w.r.t. a set Z and a loss function $l : \mathcal{H} \times Z \to \mathbb{R}_+$, if there exists a function $m_{\mathcal{H}} : (0, 1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0, 1)$ and for every distribution $\mathcal{D}$ over Z, when running the algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns $h \in \mathcal{H}$ such that, with probability of at least $1 - \delta$, over the choice of the m training examples, we have

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon,$$

where $L_{\mathcal{D}}(h) = \mathbb{E}_{z \in \mathcal{D}}[l(h, z)]$.

Remark 5.2. *In the above definition, for every $h \in \mathcal{H}$, we view the function $l(h, \cdot) : Z \to \mathbb{R}_+$ as a random variable and define $L_{\mathcal{D}}(h)$ to be the expectation of this random variable.*

Remark 5.3. *Agnostic PAC learnability $\implies$ PAC learnability. So, agnostic PAC learnability is a stronger requirement than PAC learnability.*

Definition 5.4 (ϵ-*representative sample*). A training set S is called ϵ-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function l, and distribution $\mathcal{D}$) if

$$|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon, \quad \forall\, h \in \mathcal{H}.$$

Proposition 5.5. *Assume that a training set S is $\frac{\epsilon}{2}$-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function l, and distribution $\mathcal{D}$). Then, any output of $ERM_{\mathcal{H}}(S)$, namely any $h_S \in argmin_{h \in \mathcal{H}} L_S(h)$, satisfies*

$$L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon.$$

Proof. For every $h \in \mathcal{H}$,

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \epsilon/2 \leq L_S(h) + \epsilon/2 \leq L_{\mathcal{D}}(h) + \epsilon/2 + \epsilon/2 = L_{\mathcal{D}}(h) + \epsilon.$$

□

Definition 5.6 (***Uniform Convergence***)**.** We say that a hypothesis class $\mathcal{H}$ has the uniform convergence property (w.r.t. a domain Z and a loss function l) if there exists a function $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and for every probability distribution $\mathcal{D}$ over Z, if S is a sample of size $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then with probability of at least $1 - \delta$ we have S is ϵ-representative.

Remark 5.7. *The term uniform in the above definition refers to having a fixed sample size that works for all members of $\mathcal{H}$ and over all possible probability distributions over the domain.*

Corollary 5.8. *If a class $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}^{UC}$, then the class is agnostic PAC learnable with the sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$. Furthermore, in that case, the $ERM_{\mathcal{H}}$ paradigm is a successful agnostic PAC learner for $\mathcal{H}$.*

Lemma 5.9 (***Hoeffding's Inequality***)**.** *Let $\theta_1, \ldots, \theta_m$ be a sequence of i.i.d. random variables and let $\bar{\theta} = \frac{1}{m}\sum_{i=1}^{m} \theta_i$. Assume that $\mathbb{E}(\bar{\theta}) = \mu$ and $\mathbb{P}[a \leq \theta_i \leq b] = 1$. Then for any $\epsilon > 0$, we have*

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m} \theta_i - \mu\right| > \epsilon\right] \leq 2\exp\left(\frac{-2m\epsilon^2}{(b-a)^2}\right).$$

Proposition 5.10. *Let $\mathcal{H}$ be a finite hypothesis class, Z be a domain, and $l : \mathcal{H} \times Z \to [0,1]$ be a loss function. Then, $\mathcal{H}$ enjoys the uniform convergence property with sample complexity*

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \left\lceil \frac{\log(2\mathcal{H}/\delta)}{2\epsilon^2} \right\rceil.$$

Furthermore, the class $\mathcal{H}$ is agnostic PAC learnable using the ERM algorithm with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta) \leq \left\lceil \frac{2\log(2\mathcal{H}/\delta)}{\epsilon^2} \right\rceil.$$

Proof. Fix $\epsilon, \delta \in (0,1)$. We need to find a sample of size m that guarantees that for any $\mathcal{D}$, with probability of at least $1 - \delta$ of the choice of $S = (z_1, \ldots, z_m)$ sampled i.i.d. from $\mathcal{D}$ we have that for all $h \in \mathcal{H}$, $|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$, i.e.,

$$\mathcal{D}^m(\{S : \forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon\}) \geq 1 - \delta.$$

Equivalently, we need to show that

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) < \delta.$$

Writing

$$\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\} = \bigcup_{h \in \mathcal{H}} \{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\},$$

and applying the union bound we get

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq \sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}).$$

Let θ_i be the random variable $l(h, z_i)$. For a fixed h and $z_1, \ldots, z_m$ sampled i.i.d., it follows that $\theta_1, \ldots, \theta_m$ are also i.i.d. random variables. Furthermore, $L_S(h) = \frac{1}{m}\theta_i$ and $L_{\mathcal{D}}(h) = \mu$,

where $\mu = \mathbb{E}[\theta_i]$. Since the range of l is $[0,1]$, we have that $\theta_i \in [0,1]$ for $i = 1, \ldots, m$. Then by applying Hoeffding's inequality we get

$$\mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) = \mathbb{P}\left[\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu\right| > \epsilon\right] \leq 2\exp(-2m\epsilon^2).$$

Then

$$\begin{aligned}\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) &\leq \sum_{h\in\mathcal{H}} 2\exp(-2m\epsilon^2) \\ &= 2|\mathcal{H}|\exp(-2m\epsilon^2).\end{aligned}$$

Finally, if we choose $m \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2}$, we obtain

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq \delta.$$

□

Remark 5.11 (***"Discretization Trick"***). *While the preceding proposition only applies to finite hypothesis classes, there is a simple trick that allows us to get a very good estimate of the practical sample complexity of infinite hypothesis classes. Consider an infinite hypothesis class parameterized by d parameters, such as $h(x; w) = \langle x, w\rangle$, for all $x \in \mathbb{R}^d$ and parameter $w \in \mathbb{R}^d$. In practice we store real numbers in a computer using floating point representation, say of 64 bits. Then it follows that in practice, the actual size of the hypothesis class is 2^{64d}. Applying the previous proposition, we obtain that the sample complexity of such classes is bounded by $\frac{128d+2\log(2/\delta)}{\epsilon^2}$. This upper bound on the sample complexity has the deficiency of being dependent on the specific representation of real numbers used.*

6 The Bias-Complexity Trade-off

Proposition 6.1 (***No-Free-Lunch***). *Let A be a learning algorithm for the task of binary classification w.r.t. the $0-1$ loss over a domain X. Let m be any number smaller than $|X|/2$, representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $X \times \{0,1\}$ such that:*

(a) There exists a function $f : X \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$.

(b) With probability at least $1/7$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.

Proof. Let C be a subset of X of size $2m$. There are $T = 2^{2m}$ functions from C to $\{0,1\}$. Denote these functions by $f_1, \ldots, f_T$. For each such function, let $\mathcal{D}_i$ be a distribution over $C \times \{0,1\}$ defined by

$$\mathcal{D}_i(\{(x,y)\}) = \begin{cases} 1/2m & \text{if } y = f_i(x) \\ 0 & \text{otherwise.}\end{cases}$$

Clearly, $L_{\mathcal{D}_i}(f_i) = 0$.

We will show that for every algorithm A that receives a training set of m examples from $C \times \{0,1\}$ and returns a function $A(S) : C \to \{0,1\}$, it holds that

$$\max_{i=1,\ldots,T} \mathop{\mathbb{E}}_{S\sim\mathcal{D}_i^m}[L_{\mathcal{D}_i}(A(S))] \geq 1/4.$$

There are $k = (2m)^m$ possible sequences of m examples from C. Denote these sequences by $S_1, \ldots, S_k$. Also, if $S_j = (x_1, \ldots, x_m)$ we denote by S_j^i the sequence containing instances in S_j labeled by the function f_i, i.e., $S_j^i = ((x_1, f_i(x_1)), \ldots, (x_m, f_i(x_m)))$. If the distribution is $\mathcal{D}_i$

then the possible training sets A can receive are $S_1^i, \ldots, S_k^i$, and all these training sets have the same probability of being sampled. Therefore,

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}_i^m}[L_{\mathcal{D}_i}(A(S))] = \frac{1}{k} \sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i)).$$

□

References

[1] Shalev-Shwartz S., Ben-David S. (2014) Understanding Machine Learning: From Theory to Algorithms, Cambridge University Press