Learnability of a Finite Hypothesis Class in Learning Problems with Noisy Training-Set Labels

Behrad Moniri
University of Pennsylvania
bemoniri@seas.upenn.edu

Abstract

In this note, the PAC-Learnability of a finite hypothesis class is proved for a learning problem in which the training set labels are flipped with a certain probability.

Let $\mathcal{H}$ be a finite hypothesis class, $h^* \in \mathcal{H}$ be the target concept and $S = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ be the training set. The training set labels are corrupted with noise, i.e., the label y_i is reported as $h^*(\mathbf{x}_i) \oplus \zeta$ in which $\zeta \sim \text{Bern}\ (\eta)$. We will prove that the algorithm which chooses the function h_S with the least number of errors on training points, has a low generalization error with high probability.

Definition 1. *For each $h \in \mathcal{H}$ and sample set S, Define $d(h)$ as the probability that the label y of a training sample $\mathbf{x}$ is not not equal to $h(\mathbf{x})$. Also define $\hat{d}(h)$ as the empirical loss of h on the training set, $\hat{d}(h) = \frac{\sum_{i=1}^{m} \mathbf{1}[y_i \neq h(\mathbf{x}_i)]}{m}$.*

Remark 2. *Trivially we have $d(h^*) = \eta$.*

We will now prove a very useful lemma linking $d(h)$ and the generalization error of h.

Lemma 3. *Let $\epsilon > 0$ be an arbitrary given number. If $L_D(h) > \epsilon$, then we have $d(h) - d(h^*) \geq \epsilon'$ in which $\epsilon' = \epsilon(1 - 2\eta)$.*

Proof. The label of a training set is flipped with probability η. Hence, we have

$$\begin{aligned} d(h) &= \eta\, \mathbb{P}[h^*(\mathbf{x}) = h(\mathbf{x})] + (1 - \eta)\, \mathbb{P}[h^*(\mathbf{x}) \neq h(\mathbf{x})] && (1)\\ &= \eta\big(1 - L_D(h)\big) + \big(1 - \eta\big) L_D(h) && (2)\\ &= \eta + (1 - 2\eta) L_D(h). && (3) \end{aligned}$$

For a given h, if $L_D(h) > \epsilon$, we have $d(h) - \eta = d(h) - d(h^*) > (1 - 2\eta)\epsilon = \epsilon'$. □

We will use the well known Hoeffding's inequality (Theorem (4)) in our proof.

Theorem 4. *Let $X_1, X_2, \dots, X_m$ be independent random variables. Also assume that each X_i takes values in $[a_i, b_i]$ with probability 1. For any $\epsilon > 0$, the following inequalities hold for $S_m = \sum_{i=1}^{m} X_i$:*

$$\mathbb{P}\big[S_m - \mathbb{E}[S_m] \geq \epsilon\big] \leq e^{-2\epsilon^2 / \sum_{i=1}^{m} (b_i - a_i)^2}, \quad (4)$$

$$\mathbb{P}\big[S_m - \mathbb{E}[S_m] \leq -\epsilon\big] \leq e^{-2\epsilon^2 / \sum_{i=1}^{m} (b_i - a_i)^2}. \quad (5)$$

Proof. See Appendix (D) of [MRT19]. □

We are now ready to state and prove our main claim.

Theorem 5. *In the noisy label setting with noise parameters η and η with any distribution D, let $\epsilon, \delta > 0$ be arbitrary given numbers. Given $m \geq \frac{2}{\epsilon^2 (1-2\eta)^2} \Big(\log(|\mathcal{H}|) + \log(\frac{2}{\delta}) \Big)$ training samples, for any $h \in \mathcal{H}$ with $L_D(h) > \epsilon$, we have*

$$\hat{d}(h) - \hat{d}(h^*) \geq 0 \quad (6)$$

with probability at least δ.

Proof. Let $h \in \mathcal{H}$ be any hypothesis with $L_D(h) > \epsilon$. We can decompose $\hat{d}(h) - \hat{d}(h^*)$ as follows:

$$\hat{d}(h) - \hat{d}(h^*) = [\hat{d}(h) - d(h)] + [d(h) - d(h^*)] + [d(h^*) - \hat{d}(h^*)]. \tag{7}$$

Based on lemma (3), we know that for any h with $L_D(h) > \epsilon$, we have $d(h) - d(h^*) > \epsilon'$. We will now show that given enough samples, each of the remaining terms are greater than $\frac{-\epsilon'}{2}$ with high probability.

1. First, we will show that given $m \geq \frac{2}{\epsilon^2(1-2\eta)^2}\Big(\log(\frac{2}{\delta})\Big)$ samples, we have $\hat{d}(h^*) - d(h^*) \leq \frac{\epsilon'}{2}$ with probability at least $1 - \delta/2$. This is a direct consequence of the Hoeffding's inequality. Note that $\mathrm{E}[\hat{d}(h^*)] = d(h^*)$ and we have

$$\hat{d}(h) = \frac{\sum_{i=1}^{m} \mathbf{1}[y_i \neq h(\mathbf{x}_i)]}{m} = \sum_{i=1}^{m} X_i, \tag{8}$$

   in which $X_i \in [0, \frac{1}{m}]$ with probability 1 and X_is are jointly independent. The Hoeffding's inequality yields

$$\mathbb{P}\big[\hat{d}(h^*) - d(h^*) \geq \frac{\epsilon'}{2}\big] \leq \exp\Big[-\frac{m\epsilon'^2}{2}\Big] \leq \frac{\delta}{2}. \tag{9}$$

   Where the last inequality follows from $m \geq \frac{2}{\epsilon^2(1-2\eta)^2}\Big(\log(\frac{2}{\delta})\Big)$.

2. Second, we will prove a *Uniform Convergence* property for $\mathcal{H}$. We will prove that given $m \geq \frac{2}{\epsilon^2(1-2\eta)^2}\Big(\log(|\mathcal{H}|) + \log(\frac{2}{\delta})\Big)$ samples, the event

$$\forall h \in \mathcal{H},\ d(h) - \hat{d}(h) \leq \frac{\epsilon'}{2} \tag{10}$$

   occurs with probability at least $1 - \frac{\delta}{2}$.

   To prove it, we can write the following chain of inequalities:

$$\mathbb{P}\Big[\exists h \in \mathcal{H}, d(h) - \hat{d}(h) \geq \frac{\epsilon'}{2}\Big] = \mathbb{P}\Big[\bigcup_{h\in\mathcal{H}} \big\{d(h) - \hat{d}(h) \geq \frac{\epsilon'}{2}\big\}\Big] \tag{11}$$

$$\leq \sum_{h\in\mathcal{H}} \mathbb{P}\Big[d(h) - \hat{d}(h) \geq \frac{\epsilon'}{2}\Big] \tag{12}$$

$$\leq \sum_{h\in\mathcal{H}} \exp\Big[-\frac{m\epsilon'^2}{2}\Big] = |\mathcal{H}| \exp\Big[-\frac{m\epsilon'^2}{2}\Big] \tag{13}$$

   Given $m \geq \frac{2}{\epsilon^2(1-2\eta)^2}\Big(\log(\frac{2}{\delta}) + \log(|\mathcal{H}|)\Big)$, the right hand side is less than or equal to $\frac{\delta}{2}$.

Thus, given $m \geq \frac{2}{\epsilon^2(1-2\eta)^2}\Big(\log(\frac{2}{\delta}) + \log(|\mathcal{H}|)\Big)$, the first term and third terms of (7) are both greater than $-\frac{\epsilon'}{2}$ with probability at least $1 - \frac{\delta}{2}$. Hence, with probability at least $1 - \delta$,

$$\hat{d}(h) - \hat{d}(h^*) = [\hat{d}(h) - d(h)] + [d(h) - d(h^*)] + [d(h^*) - \hat{d}(h^*)] \tag{14}$$

$$\geq -\frac{\epsilon'}{2} + \epsilon' - \frac{\epsilon'}{2} = 0, \tag{15}$$

which proves the theorem. □

This theorem states that given $m \geq \frac{2}{\epsilon^2(1-2\eta)^2}\Big(\log(\frac{2}{\delta}) + \log(|\mathcal{H}|)\Big)$ samples, the probability of the algorithm $h_S = \arg\min_{h\in\mathcal{H}} \hat{d}(h)$ having $L_D(h_S) \leq \epsilon$ is at least than $1 - \delta$.

References

[MRT19] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, second edition, 2019.