Chapter 3

Somnath Sikdar

July 9, 2019

Exercise 3.1

Let $m_{\mathcal{H}}(\epsilon, \delta)$ be the sample complexity of a PAC-learnable hypothesis class $\mathcal{H}$ for a binary classification task. For a fixed δ, let $0 < \epsilon_1 \leq \epsilon_2 < 1$ and suppose that $m_{\mathcal{H}}(\epsilon_1, \delta) < m_{\mathcal{H}}(\epsilon_2, \delta)$. Then when running the learning algorithm on $m_{\mathcal{H}}(\epsilon_1, \delta)$ i.i.d examples, we obtain a hypothesis h, which with probability at least $1 - \delta$ has a true error $L_{\mathcal{D},f}(h) \leq \epsilon_1 \leq \epsilon_2$. This implies that for the (ϵ_2, δ) combination of parameters, we can bound the true error of h by ϵ_2 by using a smaller number of i.i.d examples than $m_{\mathcal{H}}(\epsilon_2, \delta)$. This contradicts the minimality of the sample complexity function. Hence we must have $m_{\mathcal{H}}(\epsilon_1, \delta) \geq m_{\mathcal{H}}(\epsilon_2, \delta)$.

Next suppose that $0 < \delta_1 \leq \delta_2 < 1$ and that $m_{\mathcal{H}}(\epsilon, \delta_1) < m_{\mathcal{H}}(\epsilon, \delta_2)$, where ϵ is fixed in advance. Then with $m_{\mathcal{H}}(\epsilon, \delta_1)$ i.i.d examples, the learner outputs a hypothesis h which with probability at least $1 - \delta_1 \geq 1 - \delta_2$ has a true error of at most ϵ. This implies that for the (ϵ, δ_2) combination of parameters, we can bound the true error of h by ϵ by using a smaller number of i.i.d examples than $m_{\mathcal{H}}(\epsilon, \delta_2)$. This again contradicts the minimality of the sample complexity function. Hence we must have $m_{\mathcal{H}}(\epsilon, \delta_1) \geq m_{\mathcal{H}}(\epsilon, \delta_2)$.

Exercise 3.2

Given a sample S, we output a hypothesis h_S with the property that $\forall x \in \mathcal{X}$,

$$h_S(x) = \begin{cases} 1, & \text{if } (x, 1) \in S \\ 0, & \text{otherwise} \end{cases}$$

For any sample S, this hypothesis has an empirical loss of 0. Note that h_S disagrees with the true labeling function f in at most one point $z \in \mathcal{X}$. It's true loss is therefore $\Pr_{S \sim \mathcal{D}^m}\{(z, 1) \notin S\}$. This is at most $(1 - p_z)^m$, where $p_z = \Pr_{\mathcal{D}}\{z\}$.

The true loss of h_S will be 0 if $(z, 1) \in S$. Therefore the probability of getting a "bad" sample is $\Pr_{S \sim \mathcal{D}^m}\{(z, 1) \notin S\}$. In this situation, the true loss and the probability of getting a bad sample are one and the same. Let $z^* \in \mathcal{X}$ be a point at which $(1 - p_z)^m$ is maximized. Since $(1 - p_{z^*})^m \leq e^{-mp_{z^*}}$ and since we want the probability of picking a bad sample to be at most

δ, we want:

$$e^{-mp_{z^*}} < \delta \tag{1}$$

$$m > \frac{\log(1/\delta)}{p_{z^*}} \tag{2}$$

If $\epsilon \geq (1 - p_{z^*})^m$, then even a sample of size one will guarantee that the true error of h_s is at most ϵ. However if $\epsilon < (1 - p_{z^*})^m$ then $\epsilon < 1 - p_{z^*}$ and hence $p_{z^*} < 1 + \epsilon$. We can then use this upper bound in (2) to obtain:

$$m > \frac{\log(1/\delta)}{1 + \epsilon}.$$

Thus the sample complexity is $m_{\mathcal{H}}(\epsilon, \delta) = \max\{1, \frac{\log(1/\delta)}{1+\epsilon}\}$.