Chapter 3

Somnath Sikdar

July 10, 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 $\forall x \in S_x$,

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

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 $\text{Pr}_{x \sim \mathcal{D}}\{f(x) \neq h_S(x)\} = \text{Pr}_{\mathcal{D}}\{z\} := p_z$.

The true loss of h_S will be 0 if $(z, 1) \in S$. Therefore the probability of getting a "bad" sample is $\text{Pr}_{S \sim \mathcal{D}^m}\{(z, 1) \notin S\}$. 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 $e^{-mp_{z^*}} < \delta$, which gives us the sample size to be:

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

Depending on the value of the error bound ϵ, there are two situations to consider. If $\epsilon \geq p_{z^*}$, then even a sample of size one will guarantee that the true error of h_s is at most ϵ. However if $\epsilon < p_{z^*}$ then we can then use this in (1) to obtain:

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

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

Exercise 3.3

Here $\mathcal{X} = \mathbf{R}^2$ and $\mathcal{Y} = \{0, 1\}$. The hypothesis class $\mathcal{H}$ is the set of concentric circles in $\mathbf{R}^2$ centered at the origin. Assuming realizability, this implies that the true labeling function $f = h_r$ for some $r \in \mathbf{R}_+$. Thus f assigns the label 1 to any point (x, y) that is within a distance of r from the origin and 0 otherwise.

Given any sample S, let $q \in \mathbf{R}_+$ be the minimum real number such that all $(x, y) \in S_x$ with a label of 1 are included in a circle centered at the origin with radius q. The output of the ERM procedure is h_q. The empirical error of h_q is zero, but it's true error is:

$$\Pr_{(x,y)\sim\mathcal{D}}\left\{(x, y) \in C_r \setminus C_q\right\}$$

where C_r and C_q are concentric circles centered at the origin with radius r and q respectively. Given an $\epsilon > 0$, let $t \in \mathbf{R}_+$ be such that

$$\epsilon = \Pr_{(x,y)\sim\mathcal{D}}\left\{(x, y) \in C_r \setminus C_t\right\}.$$

That is, we choose t so that the true error matches the probability of picking anything inside the ring described by the circles C_r and C_t. Then the probability that we fail to choose any point in this ring in an i.i.d sample of size m is $(1 - \epsilon)^m \leq e^{-\epsilon m}$. This is the probability that we are handed a "bad" sample. Upper bounding this by δ, we obtain that $m > \log(1/\delta)/\epsilon$.

Now a sample of size at least $\log(1/\delta)/\epsilon$ has with probability at least $1 - \delta$ a point from $C_r \setminus C_t$, and hence the true error of the resulting ERM hypothesis is at most ϵ.