Chapter 3. A Formal Learning Model

Exercise 3.1.

Call a hypothesis h ϵ-approximate if $L_{\mathcal{D}}(h) \leq \inf_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon$.

By definition, $m_{\mathcal{H}}(\epsilon_2, \delta)$ samples suffice to yield an ϵ_2-approximate solution with probability at least $1 - \delta$. Since, $\epsilon_1 \leq \epsilon_2$, this solution is also ϵ_1-approximate. Thus $m_{\mathcal{H}}(\epsilon_1, \delta) \leq m_{\mathcal{H}}(\epsilon_2, \delta)$.

Similarly, an ϵ-approximate solution that fails with probability at most δ_1 also fails with probability at most $\delta_2 \geq \delta_1$.

Exercise 3.2.

3.2.1. Under realizability, the data set will have either zero or one positive examples (there may be multiple copies of the one positive example too). In the zero case, h_- clearly achieves zero empirical risk (hence, is an ERM). In the one positive example case, where z is the positive example, h_z agrees with the data, and so achieves zero empirical risk.

3.2.2. We will show that the ERM procedure (that we just described) PAC learns this hypothesis class. Therefore, we need to determine the sample complexity function.

If the true labeling function agrees with h_- then we never get a positive example and we always choose h_- which has zero risk. In this case, one example is sufficient! That is, for all distributions, when true labeling function is h_-, we require $m(\epsilon, \delta) \geq 1$ for all ϵ, δ. To show $\mathcal{H}$ is PAC learnable, we need to derive bounds for all labeling functions, which includes also h_z for some z.

Fix $z \in \mathcal{X}$ and let $p = \mathcal{D}(\{z\}) > 0$. If $p \leq \epsilon$, then h_- has risk less than ϵ, and so one sample suffices. If $p > \epsilon$, then in order to obtain a hypothesis with risk less than ϵ, we must choose h_z. We will do so exactly when the data set contains z. By independence, there is no z among m data points with probability $(1-p)^m \leq (1-\epsilon)^m \leq e^{-\epsilon m}$. We thus want to choose m such that $\delta \geq e^{-\epsilon m}$, hence $m \geq \frac{\log \frac{1}{\delta}}{\epsilon}$ samples suffice to obtain an ϵ-approximate solution with probability at least $1 - \delta$.

Thus the minimal sample complexity function $m_{\mathcal{H}}$ satisfies $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{\log \frac{1}{\delta}}{\epsilon} \rceil$.

Exercise 3.3.

As an algorithm, consider choosing the smallest concentric circle that contains all positive instances.

This is an ERM because, if it weren't, that would imply that there was a negative example closer to the origin than some positive example, but this is impossible by realizability.

Now consider the sample complexity. Pick $\epsilon, \delta \in (0, 1)$. Assume the true circle is C^*, let $C(S)$ be the smallest concentric circle containing the data.

If $\mathcal{D}_x$ assigns ϵ or more mass to its boundary, then we get no sample on the boundary with probability at most $(1-\epsilon)^m \leq e^{-\epsilon m}$. Thus with probability at least $1 - e^{-\epsilon m}$, we get at least one sample on the boundary and find a zero *risk* classifier.

Otherwise, let $C' \subseteq C^*$ such that $\mathcal{D}_x(C^* \setminus C) \in (\epsilon/2, \epsilon)$. (The existence of this set follows from the fact that probability is continuous: $\mathcal{D}_x(C') \uparrow \mathcal{D}_x(C^*)$ as $C' \uparrow C^*$.) Note that if we get a single sample *inside* $C^* \setminus C$ then we are golden because the risk of $C(S)$ would then be bounded by ϵ. On the other hand, if we get no sample in this region, then it's possible that our risk is larger and so it is sufficient (but not necessary) to prove that we get a sample in this region to obtain risk less than ϵ.

We get no sample in this region with probability $(1-\mathcal{D}_x(C^* \setminus C))^m \leq (1-\epsilon/2)^m \leq e^{-\epsilon m/2}$. If we want this probability bounded by δ, then we require $m \geq 2\log(1/\delta)/\epsilon$ samples.

Thus the minimal sample complexity function satisfies $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil 2\log(1/\delta)/\epsilon \rceil$.

Exercise 3.6.

Suppose $\mathcal{H}$ is agnostic PAC learnable by algorithm A. This implies that there is a sample complexity function $m_{\mathcal{H}}(\cdot, \cdot)$ such that, for all $\epsilon, \delta \in (0,1)$, all distributions $\mathcal{D}$ on $\mathcal{Z}$, with probability at least $1-\delta$ over $S \sim \mathcal{D}^m$, for $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, we have $L_{\mathcal{D}}(A(S)) \leq \inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$. Let P be the set of all distributions $\mathcal{D}$ on $\mathcal{Z}$ such that $\inf_{h \in \mathcal{H}} L_{\mathcal{D}}(h) = 0$. That is, P are the set of distributions satisfying the realizability criterion. Then it follows immediately that for all $\epsilon, \delta \in (0,1)$, all distributions $\mathcal{D} \in P$, with probability at least $1-\delta$ over $S \sim \mathcal{D}^m$, for $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, we have $L_{\mathcal{D}}(A(S)) \leq \epsilon$. This is precisely the statement that $\mathcal{H}$ is PAC learnable by A, proving both claims.

Exercise 3.7. Optimality of the Bayes classifier

Let $\mathcal{X}$ denote the marginal distribution of x when $(x,y) \sim \mathcal{D}$. For $z \in \mathcal{X}$, let $\mathcal{D}_{\mathcal{Y}|\mathcal{X}=z}$ denote the conds distribution of y given $x = c$ when $(x,y) \sim \mathcal{D}$.

Fix $x \in \mathcal{X}$. If $g(x) = f_{\mathcal{D}}(x)$, then

$$\mathcal{D}_{\mathcal{Y}|\mathcal{X}=x}(\{y : g(x) \neq y\}) < 1/2 \tag{1}$$

by definition of $f_{\mathcal{D}}$. Otherwise

$$\mathcal{D}_{\mathcal{Y}|\mathcal{X}=x}(\{y : g(x) \neq y\}) \geq 1/2. \tag{2}$$

Hence, $\mathcal{D}_{\mathcal{Y}|\mathcal{X}=x}(\{y : f_{\mathcal{D}}(x) \neq y\}) \leq \mathcal{D}_{\mathcal{Y}|\mathcal{X}=x}(\{y : g(x) \neq y\})$ for all x and all $g \in \mathcal{H}$. By the linearity of expectation and the identity,

$$L_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{x \in \mathcal{X}} \left[\mathcal{D}_{\mathcal{Y}|\mathcal{X}=x}(\{y : h(x) \neq y\})\right] \tag{3}$$

this completes the proof.