Understanding Machine Learning: Exercises

Somnath Sikdar

July 18, 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 $\Pr_{x \sim \mathcal{D}}\{f(x) \neq h_S(x)\} = \Pr_{\mathcal{D}}\{z\} := p_z$.

The true loss of h_S will be 0 if $(z, 1) \in S$. Therefore the probability of getting a "bad" sample is $\Pr_{S \sim \mathcal{D}^m}\{(z, 1) \notin S\}$. 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. Hence the sample complexity is upper bounded by $\lceil \log(1/\delta)/\epsilon \rceil$.

Exercise 3.4

In this example, $\mathcal{X} = \{0, 1\}^d$, $\mathcal{Y} = \{0, 1\}$ and the hypothesis class $\mathcal{H}$ is the set of all conjunctions over d Boolean variables. Since there are $\sum_{i=0}^{d} \binom{d}{i} 2^i = 3^d$ Boolean conjunctions over d Boolean variables, the hypothesis class is finite. Hence the sample complexity is

$$\begin{aligned} m_{\mathcal{H}}(\epsilon, \delta) &= \left\lceil \frac{\log(\mathcal{H}/\delta)}{\epsilon} \right\rceil \\ &= \left\lceil \frac{d \cdot \log 3 + \log(1/\delta)}{\epsilon} \right\rceil \end{aligned}$$

```
procedure PACBoolean($S$)        ▷ $S$ is the sample set with elements $\langle(a_1,\ldots,a_d),b\rangle$, where $(a_1,\ldots,a_d)\in\{0,1\}^d$ and $b\in\{0,1\}$
    $f \leftarrow x_1 \wedge \bar{x}_1 \wedge \cdots \wedge x_d \wedge \bar{x}_d$
    for each $\langle(a_1,\ldots,a_d),b\rangle \in S$ with $b=1$ do
        for $j$ in $[1,\ldots,d]$ do
            if $a_j = 1$ then
                Delete $\bar{x}_j$ from $f$, if it exists in the formula
            else
                Delete $x_j$ from $f$, if it exists in the formula
            end if
        end for
    end for
    return $f$
end procedure
```

Figure 1: Learning Boolean conjunctions

To prove that the class $\mathcal{H}$ is PAC-learnable, it suffices to exhibit a polynomial-time algorithm that implements the ERM rule. The algorithm outlined in Figure 1 starts with the formula $x_1 \wedge \bar{x}_1 \wedge \cdots \wedge x_d \wedge \bar{x}_d$. It runs through the positive examples in the sample S and for each such example, it adjusts the formula so that it satisfies the assignment given in the example. At the end of this procedure, the modified formula satisfies all positive examples of S. The time taken is $O(d \cdot |S|)$.

What may not be immediately apparent is that the formula returned by the algorithm satisfies all negative examples too. This is clear when the sample S has *no* positive examples to begin with as every assignment to $x_1 \wedge \bar{x}_1 \wedge \cdots \wedge x_d \wedge \bar{x}_d$ results in a 0. The point is that if there is even *one* positive example, for each $1 \leq i \leq d$, the algorithm eliminates either x_i or $\bar{x}_i$ depending on the assignment. That is, it eliminates half of the literals on seeing that one example and the modified formula f contains the literals of the true labeling function along with possibly others. Now the literals of the true labeling function produce a 0 on all negative examples and so does f. Hence the sampling error of the function returned by the algorithm is 0.

Exercise 3.5

The first thing to verify is that $\bar{\mathcal{D}}_m$ is a distribution. This is easy since for all $x \in \mathcal{X}$, $\bar{\mathcal{D}}_m(x) \geq 0$ and

$$\begin{aligned}\int_{x\in\mathcal{X}} \bar{\mathcal{D}}_m(x)\mathrm{d}x &= \frac{1}{m}\sum_{i=1}^{m}\int_{x\in\mathcal{X}} \mathcal{D}_i(x)\mathrm{d}x \\ &= \frac{1}{m}\sum_{i=1}^{m} 1 \\ &= 1.\end{aligned}$$

Fix an accuracy parameter $\epsilon > 0$. As in the text, define the set of bad hypotheses to be $\mathcal{H}_B = \{h \in \mathcal{H} \colon L_{\bar{\mathcal{D}}_m,f}(h) > \epsilon\}$ and let $\mathcal{M} = \{S|_x \colon \exists h \in \mathcal{H}_B, L_S(h) = 0\}$ be the set of misleading samples. Since we assume realizability, any hypothesis h_S output by the ERM procedure has $L_S(h_S) = 0$. Thus the event $L_{\bar{\mathcal{D}}_m,f}(h) > \epsilon$ and $L_S(h) = 0$ happens only when $S|_x \in \mathcal{M}$. Hence,

$$\begin{aligned}
\Pr_{\forall i \colon x_i \sim \mathcal{D}_i} \{S|_x \in \mathcal{M}\} &= \Pr_{\forall i \colon x_i \sim \mathcal{D}_i} \left\{ \bigcup_{h \in \mathcal{H}_B} \{S|_x \colon L_S(h) = 0\} \right\} \\
&\leq \sum_{h \in \mathcal{H}_B} \Pr_{\forall i \colon x_i \sim \mathcal{D}_i} \{S|_x \colon L_S(h) = 0\} \\
&= \sum_{h \in \mathcal{H}_B} \prod_{i=1}^{m} \Pr_{x_i \sim \mathcal{D}_i} \{f(x_i) = h(x_i)\} \\
&\leq \sum_{h \in \mathcal{H}_B} \Pr_{S \sim \bar{\mathcal{D}}_m} \{\forall i \colon f(x_i) = h(x_i)\}
\end{aligned}$$

This last inequality follows from the fact that the geomteric mean of a set of numbers is at most their arithmetic mean. In our specific case, we have m probabilities $p_1, \ldots, p_m$ and the inequality says:

$$\begin{aligned}
\left(\prod_{i=1}^{m} p_i\right)^{1/m} &\leq \frac{\sum_{i=1}^{m} p_i}{m} \\
\prod_{i=1}^{m} p_i &\leq \left(\frac{\sum_{i=1}^{m} p_i}{m}\right)^{m} \\
&\leq \frac{\sum_{i=1}^{m} p_i}{m}
\end{aligned}$$

Again the expression $\sum_{h \in \mathcal{H}_B} \Pr_{S \sim \bar{\mathcal{D}}_m} \{\forall i \colon f(x_i) = h(x_i)\}$ can be bounded from above by $|\mathcal{H}| \cdot (1 - \epsilon)^m \leq |\mathcal{H}| \cdot e^{-\epsilon m}$, as in the text.