MSRI Summer School on Mathematics of Machine Learning

Problem Set #2 Tuesday, July 30, 2019

Problem 2-1

Let the domain be $X = \mathbb{R}$, and let $\mathcal{C} = \mathcal{C}_s$ be the class of concepts defined by unions of s intervals. That is, each concept c is defined by real numbers $a_1 \le b_1 \le \cdots \le a_s \le b_s$ where $c(x) = 1$ if and only if $x \in [a_1, b_1] \cup \cdots \cup [a_s, b_s]$.

a. Compute the VC-dimension of $\mathcal{C}_s$ exactly.

b. Describe an efficient algorithm that learns the class $\mathcal{C}_s$ for every s, assuming that s is known ahead of time to the learner. You should describe a single algorithm that works for all $\mathcal{C}_s$, provided that s is known so that the learner can choose the number of examples needed as a function of ϵ, δ and s. Prove that your algorithm is PAC (i.e., produces a hypothesis with error at most ϵ with probability at least $1 - \delta$), and argue that both the running time and the required number of examples are polynomial in $1/\epsilon$, $1/\delta$ and s.

Problem 2-2

For this problem, you need not be concerned about computational efficiency. Throughout this problem, as usual, $\mathcal{C}$ and $\mathcal{H}$ are classes of concepts defined on the domain X.

a. Prove or disprove the following statement: For every *finite* domain X, and for all classes $\mathcal{C}$ and $\mathcal{H}$, if $\mathcal{C}$ is PAC learnable by $\mathcal{H}$, then $\mathcal{C} \subseteq \mathcal{H}$. (To prove the statement, you of course need to give a proof showing that it is always true. To disprove the statement, you can simply provide a counterexample showing that it is not true in general.)

b. Repeat part (a) *without* the assumption that X is finite. In other words, prove or disprove that: For every (not necessarily finite) domain X, and for all classes $\mathcal{C}$ and $\mathcal{H}$, if $\mathcal{C}$ is PAC learnable by $\mathcal{H}$, then $\mathcal{C} \subseteq \mathcal{H}$.

Problem 2-3

Let D be a distribution over $X \times \{0, 1\}$, and let $S = \langle (x_1, y_1), \ldots, (x_m, y_m) \rangle$ be a random sample from D. Let

$$\begin{aligned} \mathrm{err}(h) &= \Pr_{(x,y)\sim D}\left[h(x) \neq y\right] \\ \widehat{\mathrm{err}}(h) &= \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{h(x_i) \neq y_i\}. \end{aligned}$$

For simplicity, we will assume that $\mathcal{H}$ is finite, although the results of this problem can be carried over to the infinite case. Note that none of the results depend on $|\mathcal{H}|$.

Let $\hat{h}$ and h^* be the hypotheses in $\mathcal{H}$ with minimum training error and generalization error, respectively:

$$\begin{aligned} \hat{h} &= \arg\min_{h\in\mathcal{H}} \widehat{\mathrm{err}}(h) \\ h^* &= \arg\min_{h\in\mathcal{H}} \mathrm{err}(h). \end{aligned}$$

Be sure to keep in mind that, unlike h^*, $\hat{h}$ is a *random variable* that depends on the random sample S.

a. Prove that

$$\mathrm{E}\left[\widehat{\mathrm{err}}(\hat{h})\right] \leq \mathrm{err}(h^*) \leq \mathrm{E}\left[\mathrm{err}(\hat{h})\right].$$

b. Prove that, with probability at least $1 - \delta$,

$$\left|\widehat{\mathrm{err}}(\hat{h}) - \mathrm{E}\left[\widehat{\mathrm{err}}(\hat{h})\right]\right| \leq O\left(\sqrt{\frac{\ln(1/\delta)}{m}}\right).$$

Give explicit constants, and be sure to end up with a bound that is independent of $|\mathcal{H}|$.

c. Explain in words the meaning of what you proved in both of the preceding parts, and how we would expect training error to compare to test error when using a machine learning algorithm on actual data.