COMS 4771 Machine Learning (Spring 2020) Problem Set #5

Joseph High - jph2185@columbia.edu

May 3, 2020

Problem 1: Improving the Confidence

Proof. Using the hint, one can construct an algorithm (referred to as algorithm $\mathcal{B}$, or just $\mathcal{B}$, henceforth) that calls algorithm $\mathcal{A}$ on k independent sets of training samples where each run will be done over a draw of $n = O\left(\frac{1}{\epsilon^2}\right)$ samples. In total, $\mathcal{B}$ is run over a draw of n' samples such that the error of the resulting model is within ϵ-distance of the model with minimum error over all $f \in \mathcal{F}$. Algorithm $\mathcal{B}$ is constructed below.

Fix an $\epsilon_1 = \dfrac{\epsilon}{2}$ such that with probability 0.55 we have the following for each run of $\mathcal{A}$ (over a draw of n samples in each run):

$$\operatorname{err}(f_n^{\mathcal{A}}) - \inf_{f \in \mathcal{F}} \operatorname{err}(f) \;\leq\; \epsilon_1 \quad \iff \quad \operatorname{err}(f_n^{\mathcal{A}}) \;\leq\; \inf_{f \in \mathcal{F}} \operatorname{err}(f) + \epsilon_1 \tag{1}$$

Fix a $\delta > 0$ and choose some $\gamma_1 \leq \frac{\delta}{2}$ such that with probability $1 - \gamma_1 \;\geq\; 1 - \frac{\delta}{2}$ (i.e., with probability at least $1 - \frac{\delta}{2}$), the k iterations of algorithm $\mathcal{A}$ will return at least one model with a generalization error within $\frac{\epsilon}{2}$ of $\inf_{f \in \mathcal{F}} \operatorname{err}(f)$. Henceforth, such a model will be referred to as a "good" model. The number of iterations of $\mathcal{A}$ should be sufficiently large so that the probability of at least one "good" model being returned is at least $1 - \frac{\delta}{2}$. To determine a sufficient k, let X be a r.v. that denotes the number of "good" models that are returned in the k iterations of $\mathcal{A}$. Then, the probability that at least one "good" model is returned is

$$\begin{aligned}
P(X \geq 1) = \sum_{x=1}^{k} (1 - 0.55)^{k-1}(0.55) \;&=\; (0.55) \sum_{x=1}^{k} (0.45)^{k-1} \\
&=\; (0.55) \sum_{x=0}^{k-1} (0.45)^{k} \\
&=\; (0.55) \left(\frac{1 - (0.45)^k}{1 - 0.45} \right) \\
&=\; 1 - (0.45)^k
\end{aligned}$$

Therefore, we need to choose a k such that

$$1-(0.45)^k \;\geq\; 1-\frac{\delta}{2} \quad\Longleftrightarrow\quad (0.45)^k \;\leq\; \frac{\delta}{2} \quad\Longleftrightarrow\quad k\ln(0.45) \;\leq\; \ln\left(\frac{\delta}{2}\right)$$

$$\Longleftrightarrow\quad -k\ln\left(\frac{1}{0.45}\right) \;\leq\; -\ln\left(\frac{2}{\delta}\right) \quad\Longleftrightarrow\quad k \;\geq\; \frac{\ln\left(\frac{2}{\delta}\right)}{\ln\left(\frac{1}{0.45}\right)}$$

While this holds for $\forall k \;\geq\; \frac{\ln(2/\delta)}{\ln(1/0.45)}$, it suffices to set $k \;=\; \left\lceil \frac{\ln(2/\delta)}{\ln(1/0.45)} \right\rceil$. Indeed, k is necessarily an integer, so the ceiling function should be applied to the right-hand side.

Let $\mathcal{F}_{\mathcal{B}}$ be the collection of the k models returned by algorithm $\mathcal{A}$. Since $\mathcal{F}_{\mathcal{B}}$ is finite, the ERM algorithm can be applied to $\mathcal{F}_{\mathcal{B}}$. Then, for a fixed confidence level $\gamma_2 \leq \frac{\delta}{2}$ and tolerance $\epsilon_2 = \frac{\epsilon}{2}$, apply the ERM algorithm over $\mathcal{F}_{\mathcal{B}}$ using a test set of sufficient size m. Such an m should be such that[1]

$$m \;\geq\; \frac{1}{2\epsilon_2^2}\ln\left(\frac{2|\mathcal{F}_{\mathcal{B}}|}{\gamma_2}\right) \;\geq\; \frac{1}{2(\epsilon/2)^2}\ln\left(\frac{2k}{\delta/2}\right) \;=\; \frac{2}{\epsilon^2}\ln\left(\frac{4k}{\delta}\right)$$

and since m is necessarily a positive integer, we can set $m = \left\lceil \frac{2}{\epsilon^2}\ln\left(4k/\delta\right)\right\rceil$.
Let f_m^{ERM} be the model the ERM algorithm returns. By Occam's Razor, with probability at least $1-\frac{\delta}{2}$, we have

$$\operatorname{err}(f_m^{ERM}) - \inf_{f\in\mathcal{F}_{\mathcal{B}}}\operatorname{err}(f) \;\leq\; \epsilon_2 \tag{2}$$

Now, the probability that algorithm $\mathcal{B}$ fails is equal to the probability that either algorithm $\mathcal{A}$ fails to return a "good" model in k iterations or the ERM algorithm fails to select a "good" model from $\mathcal{F}_{\mathcal{B}}$. That is,

$$\begin{aligned}
\mathbb{P}(\mathcal{B}\text{ fails}) \;&=\; \mathbb{P}\left(\left\{\operatorname{err}(f_n^{\mathcal{A}}) - \inf_{f\in\mathcal{F}}\operatorname{err}(f) > \epsilon_1\right\}\bigcup\left\{\operatorname{err}(f_m^{ERM}) - \inf_{f\in\mathcal{F}_{\mathcal{B}}}\operatorname{err}(f) > \epsilon_2\right\}\right)\\
&\leq\; \mathbb{P}\left(\operatorname{err}(f_n^{\mathcal{A}}) - \inf_{f\in\mathcal{F}}\operatorname{err}(f) > \epsilon_1\right) \;+\; \mathbb{P}\left(\operatorname{err}(f_m^{ERM}) - \inf_{f\in\mathcal{F}_{\mathcal{B}}}\operatorname{err}(f) > \epsilon_2\right)\\
&=\; \gamma_1 \;+\; \gamma_2 \;\leq\; \frac{\delta}{2} \;+\; \frac{\delta}{2} \;=\; \delta
\end{aligned}$$

[1] The specific inequality used here was extracted from the proof of efficient PAC learnability for finite $\mathcal{F}$, using the the ERM algorithm, provided in the Learning Theory lecture notes (slide 8).

Then, by combining inequalities (1) and (2), we have, with probability at least $1-\delta$,

$$\begin{aligned}
\mathrm{err}(f_m^{ERM}) &\leq \epsilon_2 + \inf_{f\in\mathcal{F}_\mathcal{B}} \mathrm{err}(f) \\
&\leq \epsilon_2 + (\epsilon_1 + \inf_{f\in\mathcal{F}} \mathrm{err}(f)) \\
&= \frac{\epsilon}{2} + \frac{\epsilon}{2} + \inf_{f\in\mathcal{F}} \mathrm{err}(f)) \\
\implies \quad \mathrm{err}(f_m^{ERM}) - \inf_{f\in\mathcal{F}} \mathrm{err}(f)) &\leq \epsilon
\end{aligned}$$

where the second inquality is a result of Boole's inequality (a.k.a. subadditivity).

Since f_m^{ERM} is an arbitrary model returned by algorithm $\mathcal{B}$, the above inequality holds for any model returned by algorithm $\mathcal{B}$, with probability at least $1-\delta$. Note, that each run of algorithm $\mathcal{B}$ is run on $n' = kn + m$ samples. Then, denote any model returned by $\mathcal{B}$ as $f_{n'}^{\mathcal{B}}$. Hence, for a tolerance $\epsilon > 0$, with probability at least $1-\delta$, we have

$$\mathrm{err}(f_{n'}^{\mathcal{B}}) - \inf_{f\in\mathcal{F}} \mathrm{err}(f)) \ \leq \ \epsilon$$

Moreover,

$$k \ = \ \left\lceil \frac{\ln\left({}^{2}\!/\!_{\delta}\right)}{\ln\left({}^{1}\!/\!_{0.45}\right)} \right\rceil \quad \implies \quad k \ = \ O\left(\ln\left({}^{1}\!/\!_{\delta}\right)\right)$$

$$m \ = \ \left\lceil \frac{2}{\epsilon^2} \ln\left({}^{4k}\!/\!_{\delta}\right) \right\rceil \quad \implies \quad m \ = \ O\left(\frac{1}{\epsilon^2}\ln\left({}^{1}\!/\!_{\delta}\right)\right)$$

Recall that, $n \ = \ O\left(\frac{1}{\epsilon^2}\right)$. Therefore, since $n' = nk + m \quad \implies \quad n' \ = \ O\left(\frac{1}{\epsilon^2}\ln\left({}^{1}\!/\!_{\delta}\right)\right)$.
Since $\ln\left({}^{1}\!/\!_{\delta}\right)$ grows at a slower rate than ${}^{1}\!/\!_{\delta}$ (i.e., $\ln\left({}^{1}\!/\!_{\delta}\right) \leq {}^{1}\!/\!_{\delta}$), then n' is polynomial in ${}^{1}\!/\!_{\epsilon}$ and ${}^{1}\!/\!_{\delta}$, or equivalently

$$n' = \mathrm{poly}\left(\frac{1}{\epsilon}\ ,\ \frac{1}{\delta}\right)$$

$\square$

Problem 2: Non-linear Dimensionality Reduction

(i)

(ii)

(iii)

(iv)