

# 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  $\epsilon$ -distance of the model with minimum error over all  $f \in \mathcal{F}$ . Algorithm  $\mathcal{B}$  is constructed below.

Fix an  $\epsilon_1 = \frac{\epsilon}{2}$  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):

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

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  $\inf_{f \in \mathcal{F}} \text{err}(f)$ . Henceforth, such a model will be referred to as a “good” model, while models that don’t exhibit this behavior will be referred to as “bad”. Let  $\mathcal{F}_{\mathcal{B}}$  be the collection of the  $k$  models returned by algorithm  $\mathcal{A}$ . That is, for  $i \in \{1, \dots, k\}$ , let  $f_{n_i}^{\mathcal{A}}$  denote the model returned by  $\mathcal{A}$  in the  $i^{\text{th}}$  run, then  $\mathcal{F}_{\mathcal{B}} = \{f_{n_1}^{\mathcal{A}}, \dots, f_{n_k}^{\mathcal{A}}\}$ . The number of models in  $\mathcal{F}_{\mathcal{B}}$  (i.e., 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  $k$ , consider the following:

Let  $E_i$  be the random variable that denotes the event that the  $i^{\text{th}}$  model,  $f_{n_i}^{\mathcal{A}}$ , is a “good” model. That is, let  $E_i = \{\text{err}(f_{n_i}^{\mathcal{A}}) - \inf_{f \in \mathcal{F}} \text{err}(f) \leq \frac{\epsilon}{2}\}$ . Then, the complement,  $E_i^c$ , is the event that  $f_{n_i}^{\mathcal{A}}$  is a “bad” model:  $E_i^c = \{\text{err}(f_{n_i}^{\mathcal{A}}) - \inf_{f \in \mathcal{F}} \text{err}(f) > \frac{\epsilon}{2}\}$ . From the supposition, the probability that the  $i^{\text{th}}$  run of  $\mathcal{A}$  returns a “bad” model is

$$\mathbb{P}(E_i^c) = 1 - \mathbb{P}(E_i) = 1 - 0.55 = 0.45$$

Then, the probability that all  $k$  models returned by  $\mathcal{A}$  are “bad” is

$$\mathbb{P}(\forall i \in \{1, \dots, k\} : f_{n_i}^{\mathcal{A}} \in \mathcal{F}_{\mathcal{B}} \text{ is “bad”}) = \mathbb{P}\left(\bigcap_{i=1}^k E_i^c\right) = \prod_{i=1}^k \mathbb{P}(E_i^c) = \prod_{i=1}^k 0.45 = (0.45)^k$$

Then, the probability that there is *at least one* “good” model in  $\mathcal{F}_{\mathcal{B}}$  is

$$\mathbb{P}(\exists i \in \{1, \dots, k\} : f_{n_i}^{\mathcal{A}} \in \mathcal{F}_{\mathcal{B}} \text{ is “good”}) = 1 - \mathbb{P}\left(\bigcap_{i=1}^k E_i^c\right) = 1 - (0.45)^k$$

Therefore, we need to choose a  $k$  such that

$$\begin{aligned} 1 - (0.45)^k &\geq 1 - \frac{\delta}{2} \iff (0.45)^k \leq \frac{\delta}{2} \iff k \ln(0.45) \leq \ln\left(\frac{\delta}{2}\right) \\ \iff -k \ln\left(\frac{1}{0.45}\right) &\leq -\ln\left(\frac{2}{\delta}\right) \iff k \geq \frac{\ln\left(\frac{2}{\delta}\right)}{\ln\left(\frac{1}{0.45}\right)} \end{aligned}$$

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.

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 size  $m$ . By Occam’s Razor, such an  $m$  should be such that<sup>1</sup>

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

and since  $m$  is necessarily a positive integer, we can set  $m = \left\lceil \frac{2}{\epsilon^2} \ln(4k/\delta) \right\rceil$ .

Let  $f_m^{ERM}$  be the model returned by the ERM algorithm. It follows from Occam’s Razor that, with probability at least  $1 - \frac{\delta}{2}$ ,

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

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,

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

<sup>1</sup>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).

where the second inequality is a result of Boole's inequality (a.k.a. subadditivity). Then, by combining inequalities (1) and (2), we have, with probability at least  $1 - \delta$ ,

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

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, each run of algorithm  $\mathcal{B}$  is over a draw of  $n' = kn + m$  samples. Indeed,  $k \cdot n$  samples are used for the  $k$  iterations of  $\mathcal{A}$  and  $m$  test samples are used to identify the model in  $\mathcal{F}_B$  with the minimum empirical error. Then, denote any model returned by  $\mathcal{B}$  as  $f_{n'}^B$ . Hence, for a tolerance  $\epsilon > 0$ , with probability at least  $1 - \delta$ , we have

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

**Joe: Revise the above paragraph. Not sure it's totally correct to say this. Go with safer language.**

Moreover,

$$\begin{aligned} k &= \left\lceil \frac{\ln(2/\delta)}{\ln(1/0.45)} \right\rceil \implies k = O(\ln(1/\delta)) \\ m &= \left\lceil \frac{2}{\epsilon^2} \ln(4k/\delta) \right\rceil \implies m = O\left(\frac{1}{\epsilon^2} \ln(1/\delta)\right) \end{aligned}$$

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

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

**Joe: Possibly revise or add to the above by showing that the sample size  $n'$  is bounded by a polynomial in these terms.**

Hence,  $\mathcal{F}$  is *efficiently* PAC-learnable.

□



**Problem 2: Non-linear Dimensionality Reduction**

- (i)  $\|y_i - y_j\|$
- (ii)
- (iii)
- (iv)