CSC 482A: Problem set 3: Due by 7:00pm Tuesday, November 12

Swapnil Daxini (V00861672)

November 13, 2019

1. For this question, we have that our concept class C is weakly-learn-able as we have a weak learning algorithm A that output a hypothesis $\hat{f}$, given a sample size $n(\epsilon)$, for any $\epsilon > 0$, with a probability of $\delta_0 = \frac{1}{2}$, for which:

$$Pr_{X\ P}(\hat{f}(X) \neq c(X)) < \epsilon \tag{1}$$

In order to devise a learning algorithm for $\delta \in (0, \frac{1}{2})$, we can use that algorithm that boosts the confidence. We can do this by running the algorithm A k times to obtain a set of weak hypotheses. We can then choose k such that atleast one of our weak hypothesis achieves a risk of at most ϵ. Thus, we first have to select a value for k, such that we are sure a good hypothesis is present. Let $h_1, h_2, ..., h_k$ be the hypotheses produced using our algorithm A. The probability that none of our hypothesis is good is $(1 - \delta_0)^k = \frac{1}{2}^k$. We can then choose k such that this is equal $\frac{\delta}{2}$:

$$(\frac{1}{2})^k = \delta/2 \implies k = 2\log(\frac{2}{\delta}) \tag{2}$$

Now that we have a set of hypothesis that contains a good hypothesis, we can ERM to find a hypothesis in our set that minimizes our risk. By Theorem 3 in Lecture 5, we have that for an $\epsilon' > 0$, if we have

$$n \geq \frac{2\log(\frac{2k}{\delta})}{\epsilon'^2} \tag{3}$$

where $k = 2\log(\frac{2}{\delta})$ then with probability at least $1 - \delta/2$, we have that

$$R(\hat{f}) \leq R(f^*) + \epsilon' \tag{4}$$

We can now choose ϵ' such that $\epsilon = R(f^*) + \epsilon'$ and we are done! Our training sample size is not quite linear in $\frac{1}{\epsilon}$ but it is polynomial.

Applying the union bound with the previous step we get the required learning algorithm which with probability $1 - \delta$ will output a hypothesis with risk at most ϵ.

2. For Adaboost, we have that:

$$D_{t+1}(j) = \frac{D_t(j)e^{-\alpha_t y_j h_t(x_j)}}{Z_t} \tag{5}$$

where $\alpha_t = \frac{1}{2}\log(\frac{1-\epsilon_t}{\epsilon_t})$, $Z_t = 2(\epsilon_t(1-\epsilon_t))^{\frac{1}{2}}$ and $\epsilon_t = Pr_{j\ D_{t+1}}(h_t(X_j) \neq Y_j)$

Given that we want to find the empirical risk of h_t for the distribution D_{t+1}. This risk will be equal to sum of the weights of each sample that is predicted incorrectly by h_t:

$$\begin{aligned} R_{D_{t+1}}(h_t) &= \mathbb{1}_{y_j h_t(x_j)<0}[\sum_{j=1}^{n} \frac{D_t(j)e^{-\alpha_t y_j h_t(x_j)}}{Z_t}] \\ &= \sum_{y_j h_t(x_j)<0}^{n} \frac{D_t(j)e^{-\alpha_t}}{Z_t} \\ &= \frac{e^{-\alpha_t}}{Z_t} \sum_{y_j h_t(x_j)<0}^{n} D_t(j) \end{aligned} \tag{6}$$

The sum of weights for which h_t is wrong in D_t is simply the training error for h_t. Thus, substituting for α_t and Z_t, we have that:

$$\begin{aligned} R_{D_{t+1}}(h_t) &= \frac{(\frac{1-\epsilon_t}{\epsilon_t})^{\frac{1}{2}}}{2(\epsilon_t(1-\epsilon_t))^{\frac{1}{2}}}\epsilon_t \\ &= \frac{1}{2} \end{aligned} \tag{7}$$

Thus we have proved that the risk of h_t for the distribution D_{t+1} is $\frac{1}{2}$.

3. a) We have that $\mathcal{F}$ is the set of hypothesis used by Adaboost. Let H be the output of Adaboost (i.e. $\mathcal{F}$) after T iterations. Then, we have:

$$H(X) = sgn(\sum_{t=1}^{T} \alpha_t h_t(X)) \tag{8}$$

The VCdim of H is T. This is true because H(X) is a homogeneous linear threshold function with T variables. The VCdim of homogeneous linear threshold functions is given by the dimensions of its space, which in this case is T. I remember learning this in class at one point but I am not too sure about the proof.

Thus, by Sauer Lemma and its corollary, the growth function of H(X) is bounded by:

$$\Pi_{H(x)(n)} \leq (\frac{en}{T})^T \tag{9}$$

We have that the maximum number of choices of H(X) is equal $|\mathcal{H}|^T$, since that is number of combinations for $(h_1, h_2, ..., h_T)$, we thus have that:

$$\Pi_{\mathcal{F}(n)} \leq |\mathcal{H}|^T (\frac{en}{T})^T \tag{10}$$

b) Using Sauer's Lemma again, we have that:

$$\Pi_{\mathcal{H}} \leq (\frac{en}{V})^V \tag{11}$$

Although the answer is simply substituting the above expression into the part a, I am not exactly sure of the proof of why it is we can do that:

$$\Pi_{\mathcal{F}(n)} \leq (\frac{en}{V})^{VT} (\frac{en}{T})^T \tag{12}$$

c) From class (Lecture notes 7, Theorem 2), we learned that for an ERM classifier, we have that with probability $1 - \delta$:

$$R(\hat{f}) \leq C \frac{\log(\Pi_F) + log(\frac{1}{\delta})}{n} \tag{13}$$

I know I am supposed to use this but I am unsure how to continue.