Advanced Machine Learning: Homework Problem Set I Solutions

Guidelines: You have to submit hardcopy of the solutions (printed or handwritten) by February 14, 2018 begining of lecture class. Write your name and roll number clearly on top of the solution. Be clear and precise in your solution.

Problem 1:

We consider a two distribution variant of the PAC model in which the learning algorithm may explicitly request positive examples and negative examples, but must find a hypothesis that performs well on both the distributions of positive and negative examples.

We say that an algorithm A PAC learns a hypothesis class $\mathcal{H}$ in the two distribution variant of the PAC model if for any target concept $c \in \mathcal{H}$, for any distribution $\mathcal{D}^+$ over the positively labeled instances, distribution $\mathcal{D}^-$ over the negatively labeled instances, and for any given (ϵ, δ), if A is given access sufficiently large number (finite) of positive and negative examples that are i.i.d. in $\mathcal{D}^+$, $\mathcal{D}^-$, then A outputs a hypothesis $h \in \mathcal{H}$ such that with probability at least $1 - \delta$, $Pr_{x \sim \mathcal{D}^+}[h(x) = 0] \leq \epsilon$ and $Pr_{x \sim \mathcal{D}^-}[h(x) = 1] \leq \epsilon$.

(a) Prove that if $\mathcal{H}$ is PAC learnable using the basic (one distribution) model, then it is also PAC learnable using the two distribution model.

Proof. Let us assume that concept class $\mathcal{H}$ is PAC learnable using the one distribution PAC model using algorithm L. Consider the distribution $\mathcal{D} = \frac{1}{2}(\mathcal{D}^+ + \mathcal{D}^-)$. Let h be the hypothesis output by L. Choose s such that

$$Pr\left[erro_{\mathcal{D}}(h) \leq \frac{\epsilon}{2}\right] \geq 1 - \delta \tag{1}$$

$$\begin{aligned} erro_{\mathcal{D}}(h) &= Pr_{x \sim \mathcal{D}}[h(x) \neq c(x)] \\ &= \frac{1}{2}\left[Pr_{x \sim \mathcal{D}^+}[h(x) \neq (x)] + Pr_{x \sim \mathcal{D}^-}[h(x) \neq c(x)]\right] \\ &= \frac{1}{2}[error_{\mathcal{D}^+}(h) + error_{\mathcal{D}^-}(h)] \end{aligned}$$

From (1) $Pr[error_{(D^+)}(h) \leq \epsilon] \geq 1 - \delta$ and $Pr[error_{(D^-)}(h) \leq \epsilon] \geq 1 - \delta$

Hence $\mathcal{H}$ is also PAC learnable using the two distribution model $\square$

(b) Let h_0 be a function that always outputs 0, and h_1 be a function that always outputs 1. Prove that if a hypothesis class $\mathcal{H}$ is PAC learnable using the two distribution model, then the hypothesis class $\mathcal{H} \cup \{h_0, h_1\}$ is PAC learnable in the basic one distribution model.

Proof. Since $\mathcal{H}$ is PAC learnable using the two distribution model, there exists a learning algorithm L, such that for $h \in \mathcal{H}$, $\epsilon > 0$ and $\delta > 0$, $\exists m_+$ and m_- such that for samples of size greater than m_+ and m_- we have for h output L that $Pr[error_{\mathcal{D}^-(h)}] \leq \epsilon$ and $Pr[error_{\mathcal{D}^+(h)}] \leq \epsilon$. Suppose $\mathcal{D}$ is a probability distribution over both +ve and -ve examples. If m examples are drawn from D such that $m \geq \max\{m_+, m_-\}$ then

$$\begin{aligned} Pr[error_{\mathcal{D}}(h)] &\leq Pr[error_{\mathcal{D}}(h)|c(x) = 0] + Pr[error_{\mathcal{D}}(h)|c(x) = 1] \\ &\leq \mathcal{E}[Pr[c(x) = 0] + Pr[c(x) = 1]] \end{aligned}$$

Let $\mathcal{S}_m$ be a m-sample. Then by chernoff bounds $Pr[\mathcal{S}_m \leq (1-\alpha)m\epsilon] \leq e^{-m\epsilon\alpha^2/2}$. We want to ensure that atleast m_+ examples are found with $\alpha = \frac{1}{2}$, $m = \frac{2m_+}{\epsilon}$, $Pr[\mathcal{S}_m > m_+] \leq e^{-\frac{m_+}{4}}$. Setting the bound to be less than or equal to $\frac{\delta}{2}$, we have $m \geq \min\{\frac{2m_+}{\epsilon}, \frac{\delta}{\epsilon}\log\frac{2}{\delta}\}$ and similarly for -ve examples. We will find atleast m_+ and m_- examples, if we draw m examples if $m \geq \min\{\frac{2m_+}{\epsilon}, \frac{2m_-}{\epsilon}, \frac{\delta}{\epsilon}log\frac{2}{\delta}\}$ Otherwise if D is biased to -ve examples returns $h = h_0$ if D is biased to +ve examples returns $h = h_1$ both these guarantee $Pr[error_{\mathcal{D}}(h)] \leq \epsilon$. Hence $\mathcal{H} \cup \{h_0, h_1\}$ is PAC learnable in the basic one distribution model □

Problem 2:

Let $X = \mathbb{R}^2$ be the domain and $Y = \{0, 1\}$ be the label set of a learning problem. Let $\mathcal{H} = \{h_r, r \in \mathbb{R}_+\}$ be the set of hypothesis corresponding to all concentric circles in the plane that classify as

$$h_r(x) = \left\{ \begin{array}{ll} 1 & ||x||_2 \leq r \\ 0 & \text{otherwise} \end{array} \right\}$$

Prove that under realizability assumption $\mathcal{H}$ is PAC learnable with sample complexity

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \frac{1}{\epsilon}\log\frac{1}{\delta}$$

.

Proof. Our training dataset is $\mathcal{T}$ and learned concept $L(T)$ is the tightest circle which is consistent with T. Suppose our target concept C is the circle around origin with radius r, we will choose slightly smaller radius s by $s = \inf\{s' : P(s' \leq ||x|| \leq r) < \epsilon\}$. Let A denote the annulus between radii s and r, i.e., $A = \{x : s \leq ||x|| \leq r\}$, by definition of s,

$$P(A) \geq \epsilon \tag{2}$$

In addition, generalization error $P(C\Delta L(T))$ must be small if T intersects A. We can state this as $P(C\Delta L(T)) > \epsilon \rightarrow T \cap A = \phi$. From 2, any point in T chosen according to P will miss region A with probability at most $1-\epsilon$. Defining $error = P(C\Delta L(T))$ we get $P(error > \epsilon) \leq P(T \cap A = \phi) \leq (1-\epsilon)^m \leq e^{-m\epsilon}$ thus $m \geq \frac{1}{\epsilon} \log \frac{1}{\delta}$ $\square$