Assignment 5

Pegah Khazaei

12 April 2020

Chapter 3

Problem 2

Part 2-1

In order for the algorithm to have minimum empirical risk, in other words, it is an ERM algorithm, we define the latter as follows: Noting the labeling function f, if the training set containts a positive instance x, the algorithm returns h_x. Otherwise, it returns h^- which is an ERM.

Part 2-2

Let $\mathcal{D}$ be a probability distribution over X and $\epsilon \in (0,1)$. If the labeling function equals h^-, the algorithm return the true hypothesis. Suppose there exist a x_+ in X that is a unique positive instance. If $x_+ \in S$ then the algorithm returns the true hypothesis. As such $L_{(\mathcal{D},f)}(S) = 0$.
Assume that $x_+ \notin S$, if $\mathcal{D}(\{x_+\}) \leq \epsilon$, then $L_{(\mathcal{D},f)}(h) \leq \epsilon$ for all h in H.
Suppose that $\mathcal{D}(\{x_+\}) > \epsilon$, then for all $x' \in X$ that $x' \neq x$, $\mathcal{D}(\{x'\}) \leq 1 - \epsilon$.
Then:

$$\begin{aligned}\{S|_X : L_{(\mathcal{D},f)}(h_s) > \epsilon\} &= \{S|_X : x_+ \notin S|_X \text{ and } \mathcal{D}(\{x_+\}) > \epsilon\} \\ &= \{S|_X : \forall x' \in S|_X \ \mathcal{D}(\{x'\}) \leq 1 - \epsilon\}\end{aligned}$$

As a result

$$\begin{aligned}\mathcal{D}^m\left(\{S|_X : L_{(\mathcal{D},f)}(h_s) > \epsilon\}\right) &= \mathcal{D}^m\left(\{S|_X : \forall x' \in S|_X \ \mathcal{D}(\{x'\}) \leq 1 - \epsilon\}\right) \\ &\leq (1-\epsilon)^m \leq e^{-\epsilon m}\end{aligned}$$

Now let $\delta \in (0,1)$ such that $e^{-\epsilon m} \leq \delta$. Therefore $m \geq \frac{log(1/\delta)}{\epsilon}$. As a result, H is PAC learnable with $m_H \leq \frac{log(1/\delta)}{\epsilon}$.

Problem 3

Suppose the ERM algorithm A takes S as training set and it returns the tightest circle that contains all positive instances. The output function is denoted by h_s and its radius by r_s. Having the realizability assumption, i.e. there exists a h^* in H such that the generilization error vanishes. Its radius is denoted by r^*.

Let $\epsilon, \delta \in (0,1)$, suppose that there is a scalar r less than r^* $(r \leq r^*)$:

$$\mathcal{D}_X\left(\{x : r \leq ||x|| \leq r^*\}\right) = \epsilon$$

Suppose that $E = \{x \in \mathbb{R}^2 : r \leq ||x|| \leq r^*\}$, then

$$P\left(L_{\mathcal{D}}(h_s) \geq \epsilon\right) = P\left(x_i \in S \text{ s.t. } x \notin E\right) = \Pi_i \left(1 - P(x_i \in E)\right)^m = (1-\epsilon)^m \leq e^{m\epsilon}$$

Let $\delta \in (0,1)$, we have $e^{m\epsilon} \leq \delta$ or $m \geq \frac{log(1/\delta)}{\epsilon}$. Therefore H is PAC learnable and $m_H(\epsilon, \delta) \leq \frac{log(1/\delta)}{\epsilon}$.

Problem 4

Unfortunately I did not have enough time to solve this problem.

Problem 5

Suppose that there is a $h \in H$ such that $L_{(\overline{\mathcal{D}}_m, f)} > \epsilon$. Then based on the definition of the generalization error we have:

$$\frac{P_{x\sim\mathcal{D}_1}[h(x) \neq f(x)] + ... + P_{x\sim\mathcal{D}_m}[h(x) \neq f(x)]}{m} > \epsilon$$

then we have

$$\frac{P_{x\sim\mathcal{D}_1}[h(x) = f(x)] + ... + P_{x\sim\mathcal{D}_m}[h(x) = f(x)]}{m} \leq 1 - \epsilon$$

and

$$P\left[S|_X : L_{(S,f)}(h) = 0\right] = \Pi_{i=1}^m P_{x\sim\mathcal{D}_i}\left[h(x) = f(x)\right] = \left((\Pi_{i=1}^m P_{x\sim\mathcal{D}_i})^{\frac{1}{m}}\right)^m$$

based on geometric-arithmetic mean inequality, we have:

$$P\left[S|_X : L_{(S,f)}(h) = 0\right] \leq \left(\frac{P_{x\sim\mathcal{D}_1}[h(x) = f(x)] + ... + P_{x\sim\mathcal{D}_m}[h(x) = f(x)]}{m}\right)^m$$
$$\leq (1-\epsilon)^m \leq e^{-\epsilon m}$$

On the other hand if we have $H' = \{h \in H : L_{(\overline{\mathcal{D}}_m, f)}(h) > \epsilon\}$ and $B = \{S|_X : \exists h \in H' \text{ s.t. } L_{(S,f)}(h) = 0\}$, then:

$$P[B] = P\left[\exists h \in H \text{ s.t. } L_{(\overline{\mathcal{D}}_m, f)}(h) > \epsilon \text{ and } L_{(S,f)}(h) = 0\right]$$
$$= P\left[\bigcup_{h\in H'} \{S|_X : L_{(S,f)}(h) = 0\}\right]$$

Hence, we have:

$$P\left[\exists h \in H \text{ s.t. } L_{(\overline{\mathcal{D}}_m, f)}(h) > \epsilon \text{ and } L_{(S,f)}(h) = 0\right]$$
$$\leq |H| \times P\left[\bigcup_{h\in H'} \{S|_X : L_{(S,f)}(h) = 0\}\right]$$

The proof is complete based on the two previous equations.

Problem 6

We know that H is an agnostic PAC learnable. The objective is to show that H is also PAC learnable. Based on the assumption, there is an algorithm A and a function $m_H : (0,1)^2 \to \mathbb{N}$ such that for all $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over $X \times Y$ when running A on $m \geq m_H(\epsilon, \delta)$ iid examples generated by $\mathcal{D}$, A returns hypothesis h such that, with a probability of at least $1-\delta$, $L_{\mathcal{D}}(h) \leq \min_{h' \in H} L_{\mathcal{D}}(h') + \epsilon$.

Now we should assume that given the realizability assumption, H is PAC learnable using A.

Suppose that $\mathcal{D}$ is the probability distribution over X and f is the labeling function on H. We consider the distribution $\mathcal{D}'$ over $X \times \{0,1\}$ by drawing $x \in X$ according to $\mathcal{D}$. Based on the realizability assumption, $\min_{h' \in H} L_{\mathcal{D}}(h') = 0$. Let $\epsilon, \delta \in (0,1)$, therefore when running A on $m \geq m_H(\epsilon, \delta)$ iid examples which are labeled by f, A returns a hypothesis h such that with a probability of at least $1-\delta$ we have :

$$L_{\mathcal{D}}(h) \leq \min_{h' \in H} L_{\mathcal{D}}(h') + \epsilon = \epsilon$$

Problem 7

Suppose that $x \in X$ and P be the conditional probability of a positive label given x. We have:

$$\begin{aligned} P\left[f_{\mathcal{D}}(X) \neq y | X = x\right] &= \mathbb{1}_{[P \geq \frac{1}{2}]} P\left[Y = 0 | X = x\right] + \mathbb{1}_{[P < \frac{1}{2}]} P\left[Y = 1 | X = x\right] \\ &= \mathbb{1}_{[P \geq \frac{1}{2}]}(1-P) + \mathbb{1}_{[P < \frac{1}{2}]} P = \min\{P, 1-P\} \end{aligned}$$

Suppose that g is a classifier $g : X \to \{0,1\}$:

$$\begin{aligned} &P\left[g(X) \neq Y | X = x\right] = P\left[g(X) = 0 | X = x\right] P\left[Y = 1 | X = x\right] \\ &+ P\left[g(X) = 1 | X = x\right] P\left[Y = 0 | X = x\right] = P\left[X = x\right] P + P\left[g(X) = 1 | X = x\right](1-P) \\ &\geq P\left[g(X) = 0 | X = x\right] \min\{P, 1-P\} + P\left[g(X) = 1 | X = x\right] \min\{P, 1-P\} \\ &= \min\{P, 1-P\} \end{aligned}$$

So

$$\begin{aligned} L_{\mathcal{D}}(g) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}_{[g(x) \neq y]}\right] &= \mathbb{E}_{x \sim \mathcal{D}_x}\left[\mathbb{E}_{y \sim \mathcal{D}_{Y|x}}\left[\mathbb{1}_{[g(x) \neq y]} | X = x\right]\right] \\ &\geq \mathbb{E}_{x \sim \mathcal{D}_x} \min\{P, 1-P\} \end{aligned}$$