Assignment 3

Applied Machine Learning

Fatemeh Saberi Khomami | 400422114 | fatima.saberi@gmail.com

Question 3.2

Let X be a discrete domain, and let $H_{Singleton} = \{h_z : z \in X\} \bigcup \{h^-\}$, where for each $z \in X$, h_z is the function defined by $h_z(x) = 1$ if $x = z$ and $h_z(x) = 0$ if $x \neq z$. h^- is simply the all-negative hypothesis, namely, $\forall x \in X, h^-(x) = 0$. The realizability assumption here implies that the true hypothesis f labels negatively all examples in the domain, perhaps except one.

1. Describe an algorithm that implements the ERM rule for learning $H_{Singleton}$ in the realizable setup.
2. Show that $H_{Singleton}$ is PAC learnable. Provide an upper bound on the sample complexity.

Answer:
1. Since h^- is already covering data points with negative label, and the true hypothesis f labels most of the samples negative, we only need to introduce an algorithm that covers the positive samples.
We propose an algorithm that returns h^+ for positive instances. That is

$$\forall x \in X^+ \; , \; h^+(x) = 1$$

For non-positive samples, our proposed algorithm returns h^-.
This algorithm is ERM. As the question mentioned before, "the true hypothesis f labels negatively all examples in the domain, perhaps except one." So, majority of samples will be assigned label 0, the few remaining which are non-negative will receive label 1.
Therefore, By the definition of ERM, our proposed algorithm is $argmin_{h \in H}(L_s(h))$

Answer:
2. To show that $H_{Singleton}$ is PAC learnable, we must show that there exists a function $m_H : (0,1)^2 \longrightarrow \mathbb{N}$ and a learning algorithm with the following property: For every ϵ , $\delta \in (0,1)$, for every distribution D over X, and for every labeling function $f : X \longrightarrow \{0,1\}$, if the realizable assumption holds with respect to H, D, f , then when running the learning algorithm on $m \geq m_H(\epsilon,\delta)$ $i.i.d$ examples generated by D and labeled by f , the algorithm returns a hypothesis h such that,

$$\mathbb{P}(L_{(D,f)}(h) \leq \epsilon) \geq 1 - \delta$$

To prove that the error is at most ϵ, we'll show that the probability that learner fails is less than δ, that is,

$$\mathbb{P}(L_{(D,f)}(h_s) > \epsilon) < \delta$$

Assume the probability of facing a positive sample is ϵ, The worst case of learner's failure could be mislabeling every positive instance.

$$\mathbb{P}^m(h(x) \neq f(x) \,|\, x \in X^+) \;=\; (1-\epsilon)^m \leq e^{-\,m\,\epsilon}$$

Now we want the above statement to be less than δ.

$$\begin{aligned} e^{-\,m\,\epsilon} &\leq \delta \\ -m\epsilon &\leq \ln(\delta) \\ m &\geq \frac{\ln(\frac{1}{\delta})}{\epsilon} \end{aligned}$$

So, with sample complexity function $m_H(\epsilon,\delta) \leq \frac{\ln(\frac{1}{\delta})}{\epsilon}$, $H_{Singleton}$ is PAC learnable.

Question 3.3

Let $X = \mathbb{R}^2$, $Y = \{0,1\}$, and let H be the class of concentric circles in the plane, that is, $H = \{h_r : r \in \mathbb{R}_+\}$, where $h_r(x) = \mathbb{1}_{[\|x\| \leq r]}$. Prove that H is PAC learnable (assume realizability), and its sample complexity is bounded by

$$m_H(\epsilon,\delta) \leq \left\lceil \frac{\log \frac{1}{\delta}}{\epsilon} \right\rceil.$$

Answer: H is the class of oncentric circles and any hypothesis $h \in H$ is a circle which will label anything inside itself, positive, and everything outside of itself, negative. As we said in the previous question, H is PAC learnable if there is a $h \in H$ in which:

$$\mathbb{P}(L_{(D,f)}(h) \le \epsilon) \ge 1 - \delta$$

To prove that the error is at most ϵ, we'll show that the probability that learner fails is less than δ, that is,

$$\mathbb{P}(L_{(D,f)}(h_s) > \epsilon) < \delta$$

ERM can return a h^* which is the best circle with the least radius(r^*). Regarding realizability assumption, it can correctly label all samples inside itself, positive. If we choose a circle with radius less than h^*'s radius, it will fail to classify our samples correctly. Assume the worst case scenario, in which we chose a circle with radius r_f ($r_f \le r^*$) that all the positive samples are left out of the region, and the probability of facing a positive sample inside the circle is ϵ. So,

$$\mathbb{P}^m(h_r(x) = \mathbb{1}_{[\|x\| \le r_f]}) \; = \; (1-\epsilon)^m \le e^{-\,m\,\epsilon}$$

Now we want the above statement to be less than δ.

$$e^{-\,m\,\epsilon} \le \delta$$
$$-m\epsilon \le \ln(\delta)$$
$$m \ge \frac{\ln(\frac{1}{\delta})}{\epsilon}$$

So, with sample complexity function $m_H(\epsilon, \delta) \le \frac{\ln(\frac{1}{\delta})}{\epsilon}$, H is PAC learnable.

Question 3.4

In this question, we study the hypothesis class of *Boolean conjunctions* defined as follows. The instance space is $X = \{0,1\}^d$ and the label set is $Y = \{0,1\}$. A literal over the variables $x_1, \ldots, x_d$ is a simple Boolean function that takes the form $f(x) = x_i$, for some $i \in [d]$, or $f(x) = 1 - x_i$ for some $i \in [d]$. We use the notation $\overline{x_i}$ as a shorthand for $1 - x_i$. A conjunction is any product of literals. In Boolean logic, the product is denoted using the $\wedge$ sign. For example, the function $h(x) = x_1 \cdot (1 - x_2)$ is written as $x_1 \wedge \overline{x_2}$.
We consider the hypothesis class of all conjunctions of literals over the d variables. The empty conjunction is interpreted as the all-positive hypothesis (namely, the function that returns $h(x) = 1$ for all x). The conjunction $x_1 \wedge \overline{x_1}$ (and similarly any conjunction involving a literal and its negation) is allowed and interpreted as the all-negative hypothesis (namely, the conjunction that returns $h(x) = 0$ for all x).
We assume realizability: Namely, we assume that there exists a Boolean conjunction that generates the labels. Thus, each example $(x,y) \in X \times Y$ consists of an assignment to the d Boolean variables $x_1, \ldots, x_d$, and its truth value (0 for false and 1 for true).

For instance, let $d = 3$ and suppose that the true conjunction is $x_1 \wedge \overline{x_2}$. Then, the training set S might contain the following instances:

$$((1,1,1),0),((1,0,1),1),((0,1,0),0)((1,0,0),1).$$

Prove that the hypothesis class of all conjunctions over d variables is PAC learnable and bound its sample complexity. Propose an algorithm that implements the ERM rule, whose runtime is polynomial in $d \cdot m$.

Answer: Unfortunately, I couldn't comprehend this question.

Question 3.5

Let X be a domain and let $D_1, D_2, \ldots, D_m$ be a sequence of distributions over X . Let H be a finite class of binary classifiers over X and let $f \in H$. Suppose we are getting a sample S of m examples, such that the instances are independent but are not identically distributed; the ith instance is sampled from D_i and then y_i is set to be $f(x_i)$. Let $\overline{D_m}$ denote the average, that is, $\overline{D_m} = \frac{(D_1 + \cdots + D_m)}{m}$.
Fix an accuracy parameter $\epsilon \in (0,1)$. Show that

$$\mathbb{P}[\exists h \in H \ s.t. \ L_{(\overline{D_m} \ ,f)}(h) > \epsilon \ and \ L_{(S,f)}(h) = 0] \leq |H| e^{-\epsilon m}$$

Hint: Use the geometric-arithmetic mean inequality.

Answer: We can always assume that we have a bad hypothesis like h that $L_{(\overline{D}_m, f)} \geq \epsilon$. $\overline{D}_m$ means x_1 was sampled from D_1, x_2 was sampled from D_2, and consequently, x_m was sampled from D_m. We can write $L_{(\overline{D_m} \ ,f)}(h) \geq \epsilon$ as :

$$\frac{\mathbb{P}_{X \sim D_1}[h(X) \neq f(X)] + \cdots + \mathbb{P}_{X \sim D_m}[h(X) \neq f(X)]}{m} > \epsilon$$

$$\Rightarrow \frac{\mathbb{P}_{X \sim D_1}[h(X) = f(X)] + \cdots + \mathbb{P}_{X \sim D_m}[h(X) = f(X)]}{m} < 1 - \epsilon$$

Since $f \in H$, there are hypothesises in which have $L_s(h) = 0$. In the class of misleading samples with a bad hypothesis like h in $H_B \subseteq H$, the $L_s(h)$ is always zero. Let's calculate its probability.

$$\mathbb{P}_{\substack{S\sim \prod_{i=1}^{m} D_i \\ h\in H_B}}[L_s(h)=0]=\prod_{i=1}^{m}\mathbb{P}_{S\sim D_i}[h(X)=f(X)]$$ because samples were taken independently

$$=\left(\left(\prod_{\substack{i=1 \\ h\in H_B}}^{m}\mathbb{P}_{S\sim D_i}[h(X)=f(X)]\right)^{\frac{1}{m}}\right)^{m}$$ due to geometric-arithmetic mean inequality:

$$\leq\left(\frac{\sum_{\substack{i=1 \\ h\in H_B}}^{m}\mathbb{P}_{S\sim D_i}[h(X)=f(X)]}{m}\right)^{m}$$

now that we know there is a bad h which is:

$$\mathbb{P}_{S\sim\prod_{i=1}^{m}D_i}[L_s(h)=0]\leq\left(\frac{\sum_{\substack{i=1 \\ h\in H_B}}^{m}\mathbb{P}_{S\sim D_i}[h(X)=f(X)]}{m}\right)^{m}$$

Since we proved for a bad hypothesis we have:

$$\frac{\mathbb{P}_{X\sim D_1}[h(X)=f(X)]+\cdots+\mathbb{P}_{X\sim D_m}[h(X)=f(X)]}{m}<1-\epsilon$$

We can merge 2 above inequalities, and infer that:

$$\mathbb{P}[L_{(\overline{D_m},f)}(h)>\epsilon \textit{ and } L_{(S,f)}(h)=0]\leq\sum_{h\in H_B}(1-\epsilon)^{m}$$ due to Taylor series

$$\leq\sum_{h\in H_B}e^{-m\epsilon}$$

$$=|H_B|e^{-m\epsilon}$$

$$\leq|H|e^{-m\epsilon}$$

Question 3.6

Let H be a hypothesis class of binary classifiers. Show that if H is agnostic PAC learnable, then H is PAC learnable as well. Furthermore, if A is a successful agnostic PAC learner for H, then A is also a successful PAC learner for H.

Answer: We are assuming that H is agnostic PAC learnable, so for any $A(s) = h \in H$:

$$\mathbb{P}\left(L_D(h) \leq \min L_D(h') + \epsilon\right) \geq 1 - \delta$$

Because H is the infinite class of binary classifiers, there exists a flexible h in which can shatter the domain into positive and negative. Whilst that h exists in H, the term $\min L_D(h')$ is zero. Therefore:

$$\mathbb{P}\left(L_D(h) \leq \epsilon\right) \geq 1 - \delta$$

According to above, we can conclude that H is PAC learnable as well.