Machine learning HW 2

Understanding Machine Learning: From Theory to Algorithms[Exercises for chapters 3 and 5]

Mehrnaz jalili:400422061- Arash Sajjadi:400422096
Email: Mehrnazjalili1991@gmail.com
Email: arash.sajjadi@yahoo.com

Date: November 28, 2021

chapter 3

Exercise 3.2:

$$\text{Preliminary assumptions}\begin{cases} \mathcal{X}\longrightarrow \qquad \text{Discrete} \\ \mathcal{H}=\{h_z : z\in\mathcal{X}\}\cup\{h^-\} \end{cases} \qquad h_z\left(x\right)=\begin{cases}1 & x=z\\ 0 & x\neq z\end{cases}$$

3.2.1:

Our algorithm will be that if z is in our training set, this member will be labeled $+$, and all other members will be labeled $-$. If z is not in our training set, all members must get a label. In both cases, the amount of training set's error is **zero**, so it is evident that our algorithm is an **ERM**.

3.2.2:

If there is no member with label $+$ in set $\mathcal{X}$ (whit respect to distribution $\mathcal{D}$), it is evident that our true label is a fixed function always returns 0 (assign to each member of label $-$) The question becomes a little more complicated when a member in $\mathcal{X}$ does not have a label $+$. In this case, two modes are possible. The first case is that we see this member in the training set, in which case, according to the algorithm presented in the previous section, we can claim to have achieved the true label. The second or rather the most challenging part of our problem is when the member with the label $+$ is inside $\mathcal{X}$ but does not appear in the training set.In this we have:$\mathcal{D}(z)=\epsilon, \mathcal{D}(x\neq z)=1-\epsilon$.[1].

$$\begin{aligned} \mathbb{P}_{S\sim\mathcal{D}^m}\left[S|_x, L_{\mathcal{D},f}>\epsilon\right]<\delta \\ \mathbb{P}\left[x^+\notin S|_m\right]= \\ \left(1-\mathbb{P}(x^+)\right)^m=\left(\mathbb{P}(x^-)\right)^m= \\ \left(1-\epsilon\right)^m\leq e^{-\epsilon m} \end{aligned}$$

At this stage we will easily have: $e^{-\epsilon m}\leq\delta\Rightarrow m\geq\frac{\log\left(\frac{1}{\delta}\right)}{\epsilon}$. Therfore, sample complexity is bounded by

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

Exercise 3.3:Concentric Circles

$$\text{Preliminary assumptions}\begin{cases} \mathcal{X}=\mathbb{R}^2 & y=\{0,1\} \\ \mathcal{H}=\{h_r : r\in\mathbb{R}_+\} & h_r(x)=\mathbb{1}_{[\|x\|\leq r]} \end{cases}$$

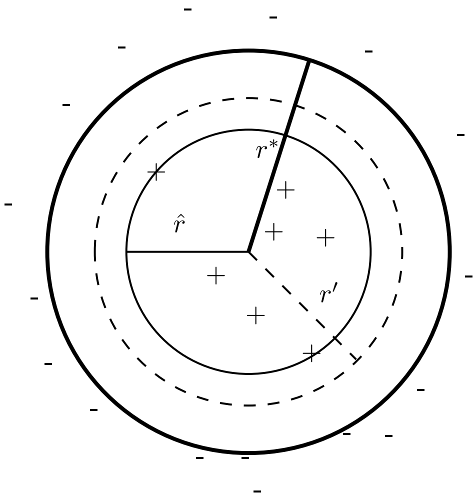
Consider Algorithm A in such a way that it returns the smallest circle that contains all the positive elements in our training set. Similarly to question number 3 in Chapter 2 of this book, this algorithm can be considered as an ERM.

We have to prove that $\mathcal{H}$ is PAC learnable.To solve the problem, we need a number of definitions, which we will introduce in this section.

[1] We can also show it by $\mathcal{D}(x^+)=\epsilon, \mathcal{D}(x^-)=1-\epsilon$

- R^* is a circle that represents our true label. But obviously, we do not have access to this circle because we do not have access to all the data in the universe of our data. So, and finding this circle is ideal for us. We also represent the radius of this circle with r^*. Also, the corresponding member of this circle in the set of hypotheses could be considered to be h^*.
- Consider $\hat{R}$ as the output of Algorithm A, which we talked about earlier. Similarly, $\hat{r}$ is the radius of the circle $\hat{R}$. $\hat{h}$ is also a member of our set of hypotheses that represents $\hat{R}$ for us. [2]
- Perhaps the key to solving this problem depends on the definition of R'. We define R' such that $r' \leq r^*$ and also for every $\epsilon, \delta \in (0,1)$, we have $\mathcal{D}_{\mathcal{X}}\left(\{x : r' \leq \|x\| \leq r^*\}\right) = \epsilon$. r' and h' are defined similarly to the previous items.

It is clear that $\mathcal{D}(L_{\mathcal{D},f}(h_S) \geq \epsilon)$ is bounded above by the probability that no point *in* S belongs to $(R^* - R')$. This probability of this event is bounded above by $(1-\epsilon)^m$. Having Taylor series, $(1-\epsilon)^m \leq e^{-\epsilon m}$



Our personal perception of this problem

$$\mathcal{D}(L_{\mathcal{D},f}(h_S) \geq \epsilon) = \mathcal{D}(S \cap \hat{R} = \varnothing) \overset{\text{i.i.d}}{=} (1-\epsilon)^m \leq e^{-\epsilon m} \leq \delta$$

Therfore, $\mathcal{H}$ is PAC learnable and the sample complexity is:

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

Exercise 3.4:Boolean conjunctions

Let's value $\mathcal{H}$ With the help of each other. There are three modes for each variable.

1. x_i exists in the multiplication function of variables.
2. $\overline{x}_i$ exists in the multiplication function of variables.
3. There is no combination of x_i in our function.

$$|\mathcal{H}| = 3^d \Rightarrow |\mathcal{H}| : \text{finite} \Rightarrow \mathcal{H} \text{ is PAC learnable} \Rightarrow m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log\left(\frac{|\mathcal{H}|}{\delta}\right)}{\epsilon} \right\rceil$$

[2] $L_S(\hat{h}(x_{i\in[m]})) = 0$

$$m_{\mathcal{H}}(\epsilon,\delta) \leq \left\lceil \frac{\log\left(\frac{3^d}{\delta}\right)}{\epsilon} \right\rceil = \left\lceil \frac{d.\log(3)-\log(\delta)}{\epsilon} \right\rceil$$

We couldn't answer this question entirely, but almost half of the questions have been answered.

Exercise 3.5:I.N.I.D

$$\text{Preliminary assumpti} \begin{cases} \mathcal{D}_m = \frac{\sum_{i=1}^m \mathcal{D}_i}{m} \\ \mathcal{X} \rightarrow \text{independent} \\ S = \{x_1,...,x_m\} \Rightarrow |S| = m \\ \text{realizability}\checkmark \qquad f \in \mathcal{H}, \epsilon \in (0,1) \end{cases}$$

We have to show that $\mathbb{P}\left[\exists h \in \mathcal{H} : L_{(\mathcal{D},f)}(h) > \epsilon \text{ and } L_{(S,f)(h)=0)}\right] \leq |\mathcal{H}|\, e^{-\epsilon m}$

$$\mathbb{P}_{S\sim\prod_{i=1}^m \mathcal{D}_i}[L_{S,f}(h)=0] \underset{\text{expand}}{=} \prod_{i=1}^m \underbrace{\mathbb{P}_{x\sim\mathcal{D}_i}[f(x)=h(x)]}_{A_i} = \prod_{i=1}^m A_i$$

According to the question form's hint, we can convert this expression into a summation as follows.

$$\prod_{i=1}^m A_i = \left(\left(\prod_{i=1}^m A_i\right)^{\frac{1}{m}}\right)^m \leq \left(\frac{\sum_{i=1}^m A_i}{m}\right)^m \leq (1-\epsilon)^m \leq e^{-\epsilon m}$$

It is clear that the members of our set of hypotheses that $L_{(\mathcal{D},f)}(h) > \epsilon$ are less than $|H|$. Therefore, up to $|H|$ we have to check the probability of $L_{S,f)}(h) = 0$ on the members of H that$L_{(\mathcal{D},f)}(h) > \epsilon$. Therefore

$$\mathbb{P}\left[\exists h \in \mathcal{H} : L_{(\mathcal{D},f)}(h) > \epsilon \text{ and } L_{(S,f)(h)=0)}\right] \leq |\mathcal{H}|\, e^{-\epsilon m}$$

Exercise 3.6:

$$\mathcal{D}^m\left(S|_x : L_D(h_S) > \min_{h\in\mathcal{H}} L_D(h) > \epsilon\right) < \delta \quad \text{and} \qquad \exists A(S) \in \mathcal{H} \text{ s.t } L_{\mathcal{D},f}(A(S)) = 0 \text{ (assume realizability)}$$

$$\Rightarrow \forall m > m_{\mathcal{H}}(\epsilon,\delta) : \min_{h\in\mathcal{H}} L_D(h) = 0 \qquad \Rightarrow \mathcal{D}_{\mathcal{X}}^m\left(\{S|_x : L_{\mathcal{D},f}(h) < \epsilon\}\right) < \delta$$

$\Rightarrow A$ is a successful PAC learner of $\mathcal{H}$ and $\mathcal{H}$ is PAC learnable.

Exercise 3.7:The Bayes optimal predictor

Let, $x \in \mathcal{X}, c$: conditional probability of a positive label given x.

$$\begin{aligned} \mathbb{P}[f_{\mathcal{D}}X \neq y|X=x)] &= \mathbb{1}_{c_x \geq 0.5}.\mathbb{P}[Y=0|X=x] + \mathbb{1}_{c_x<0.5}.\mathbb{P}[Y=1|X=x] \\ &= \mathbb{1}_{c_x\geq 0.5}.(1-c_x) + \mathbb{1}_{c_x<0.5}.c_x = \min\{c_x, 1-c_x\} \end{aligned}$$

Therefore, it can be calculated that

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

It can be calculated that

$$\begin{aligned} L_{\mathcal{D}}(f_{\mathcal{D}}) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{1}_{[f_{\mathcal{D}}(x)\neq y]}] \\ &= \mathbb{E}_{x\sim\mathcal{D}_X}\left[\mathbb{E}_{y\sim\mathcal{D}_Y[\mathbb{1}_{f_{\mathcal{D}}(x)\neq y}|X=x]}\right] \\ &= \mathbb{E}_{x\sim\mathcal{D}_x}[c_x] \\ &\leq \mathbb{E}_{x\sim\mathcal{D}_X}\left[\mathbb{E}_{y\sim\mathcal{D}_Y|x}[\mathbb{1}_{[g(x)\neq y]|X=x}] = L_{\mathcal{D}}(g)\right] \end{aligned}$$

Exercise 5.2:

1. In general, the fewer features we have, the easier it is to solve the problem. The answer to the problem is probably more linear, and as a result, we will not have the problem of computation and also computational speed. On the other hand, the more data there is, the decrease our error. The problem becomes more difficult. Also, our answer to the functional problem is more complex.

2. $$\mathcal{H}_5 \supseteq \mathcal{H}_2 \qquad e_{\text{AP}}\left(\mathcal{H}_5\right) \leq e_{\text{AP}}\left(\mathcal{H}_2\right) \qquad \text{Complexity of } \mathcal{H}_5 \geq \text{Complexity of } \mathcal{H}_2$$

   Therefore, if the size of our training data set is small, work on two-dimensional data is recommended.