CS 260: Machine Learning **Spring 2020**

Homework 1

Student Name: Andrew Choi *UID: 205348339*

- Feel free to talk to other students in the class when doing the homework. You should, however, write down your solution yourself. You also must indicate on each homework with whom you collaborated and cite any other sources you use including Internet sites.
- You will write your solution in LaTeX and submit the **pdf file through Gradescope**. You also need to submit the **zipped LaTeX files to CCLE**. We will grade your homework based on the final version of the pdf file submitted to Gradescope. We will not grade the zipped Latex files on CCLE. However, failure to submitting your LaTax files to CCLE will incur 2 points penalty out of 100 points.
- The homework (both pdf and zipped Latex source files) is due at **1:59 PM before the class**.

1. Exercise **2.2**

   Show that

   $$\mathop{\mathbb{E}}_{S|x\sim\mathcal{D}^m}[L_s(h)] = L_{(\mathcal{D},f)}(h)$$

   Can be proved through linearity of expectation:

   $$\begin{aligned}\mathop{\mathbb{E}}_{S|x\sim\mathcal{D}^m}[L_s(h)] &= \mathop{\mathbb{E}}_{S|x\sim\mathcal{D}^m}\left[\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}[h(x_i)\neq f(x_i)]\right] \\ &= \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{E}}_{x_i\sim\mathcal{D}^m}[\mathbb{1}[h(x_i)\neq f(x_i)]] \\ &= \frac{1}{m}\sum_{i=1}^{m}\mathop{\mathbb{P}}_{x_i\sim\mathcal{D}^m}[h(x_i)\neq f_(x_i)] \\ &= L_{(\mathcal{D},f)}(h)\end{aligned}$$

2. Exercise **2.3**

   a. As A produces the smallest rectangle enclosing all positive training set examples, then $L_S(A(S)) = 0$. Therefore, A is an ERM algorithm by definition.

   b. Fix some distribution $\mathcal{D}$ over $\mathcal{X}$. Let R^* be the rectangle that generates the labels and f be the corresponding hypothesis.

   By definition, algorithm A implies $R(S) \subseteq R^*$. Following this, we define the event in which S does not contain a positive instance in $R_i\ \forall\ i \in \{1,2,3,4\}$ below.

   $$R_i^- = \{S|x : S|x \cap R_i = \emptyset\}$$

   As the probability mass of R_i^- is $1-\epsilon/4\ \forall\ i$, the probability of R_i^- occurring is

   $$\mathop{\mathbb{P}}_{S|x\sim\mathcal{D}^m}[R_i^-] = (1-\epsilon/4)^m\ \forall\ i \in\ \{1,2,3,4\}$$

From here, we can see that we can upper bound the probability of sampling a S that leads to a generalization error higher than ϵ as the union of the sets R_i^-.

$$\mathop{\mathbb{P}}_{S|x\sim\mathcal{D}^m}[\{S|x : L_{(\mathcal{D},f)}(h_s) > \epsilon\}] \leq \mathop{\mathbb{P}}_{S|x\sim\mathcal{D}^m}\left(\bigcup_{i=1}^{4} R_i^-\right)$$

Using the union bound, we know the following.

$$\begin{aligned}
\mathop{\mathbb{P}}_{S|x\sim\mathcal{D}^m}[\{S|x : L_{(\mathcal{D},f)}(h_s > \epsilon\}] &\leq \mathop{\mathbb{P}}_{S|x\sim\mathcal{D}^m}\left(\bigcup_{i=1}^{4} R_i^-\right) \\
&\leq \sum_{i=1}^{4} \mathop{\mathbb{P}}_{S|x\sim\mathcal{D}^m}\left(R_i^-\right) \\
&= 4(1-\epsilon/4)^m \\
&\leq 4e^{-\epsilon m/4}
\end{aligned}$$

Plugging in $m = 4\log(4/\delta)/\epsilon$ results in a value of δ.

$$\begin{aligned}
4e^{-\epsilon m/4} &= 4e^{-\log(4/\delta)} \\
&= 4e^{\log(\delta/4)} \\
&= \delta
\end{aligned}$$

c. For axis aligned rectangles in $\mathbb{R}^d$, we define a set of rectangles $R_i \; \forall \; i \in [2d]$ similar to the previous definition with each a probability mass of exactly $\epsilon/2d$ where d is the number of dimensions. We also define the set of all samples that do not contain a positive instance in R_i as well as the probability of this event occurring.

$$\begin{aligned}
R_i^- &= \{S|x : S|x \cap R_i = \emptyset\} \\
\mathop{\mathbb{P}}_{S|x\sim\mathcal{D}^m}[R_i^-] &= (1-\epsilon/2d)^m \; \forall \; i \in [2d]
\end{aligned}$$

From here we can perform the same steps as before.

$$\begin{aligned}
\mathop{\mathbb{P}}_{S|x\sim\mathcal{D}^m}[\{S|x : L_{(\mathcal{D},f)}(h_s > \epsilon\}] &\leq \mathop{\mathbb{P}}_{S|x\sim\mathcal{D}^m}\left(\bigcup_{i=1}^{2d} R_i^-\right) \\
&\leq \sum_{i=1}^{2d} \mathop{\mathbb{P}}_{S|x\sim\mathcal{D}^m}\left(R_i^-\right) \\
&= 2d(1-\epsilon/2d)^m \\
&\leq 2de^{-\epsilon m/2d}
\end{aligned}$$

Therefore, the right hand side will be $\leq \delta$ when $m \geq \lceil 2d\log(2d/\delta)/\epsilon \rceil$.

d. For each dimension, the algorithm must find the minimum and maximum bound that captures all positive instances. Therefore, given a sample of size m, the runtime is $O(md)$. Using the definition of $m = \lceil 2d \log(2d/\delta)/\epsilon \rceil$, by plugging this value into $O(md)$, we can see that algorithm A is polynomial in $d, 1/\epsilon$, and in $\log(1/\delta)$.

3. Exercise **3.1**

This problem can be solved simply by using the definition of PAC learnability. Since $\mathcal{H}$ is PAC learnable, then for every distribution $\mathcal{D}$ and perfect labeling function $f \in \mathcal{H}$, with probability $1 - \delta$ over the choice of i.i.d. samples S of size m, we know that

$$m \geq \frac{log(|\mathcal{H}|/\delta)}{\epsilon}$$

Therefore, for $0 < \epsilon_1 \leq \epsilon_2 < 1$ and $\delta \in (0, 1)$, we can use the minimality of sample complexity to show

$$m_1 = m_{\mathcal{H}}(\epsilon_1, \delta) \geq m_{\mathcal{H}}(\epsilon_2, \delta) = m_2$$

Similarly for $0 < \delta_1 \leq \delta_2 < 1$ and $\epsilon \in (0, 1)$,

$$m_1 = m_{\mathcal{H}}(\epsilon, \delta_1) \geq m_{\mathcal{H}}(\epsilon, \delta_2) = m_2$$

4. Exercise **3.2**

a. The following algorithm is an ERM under the realizability assumption: If training set S contains a positive instance, then return h_z. Otherwise, if S contains all negative instances, return h^-.

b. We can see that if the true hypothesis is h^-, then the proposed algorithm will return a perfect hypothesis. If there does exist a positive instance, then it is clear to see that the true hypothesis will be $h^* \in h_z$, but for training sets that do not contain this positive instance, the algorithm will produce an imperfect hypothesis. With this in mind, we can define $\mathcal{H}_B$ and $\mathcal{M}$.

$$\mathcal{H}_B = \{h^- : L_{(\mathcal{D},f)}(h^-) > \epsilon\}$$
$$\mathcal{M} = \{S|x : h \in \mathcal{H}_B, L_s(h) = 0\}$$

We can now show that $\mathcal{H}_{Singleton}$ is PAC learnable and upper bound its sample complexity by the following. Note that $\mathcal{H}_B$ only consists of one hypothesis, therefore

$$\begin{aligned}\mathcal{D}^m(\{S|x : L_{(\mathcal{D},f)}(h_s) > \epsilon\}) &\leq \mathcal{D}^m(\{S|x : h \in \mathcal{H}_B, L_s(h) = 0\}) \\ &\leq (1 - \epsilon)^m \\ &\leq e^{-\epsilon m}\end{aligned}$$

Therefore, the sample complexity is upper bounded by the following.

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

5. Exercise **3.3**

   Under the realizability assumption, we can assume that the true hypothesis is the circle with radius r^* that contains all positive instances and excludes all negative instances. Therefore, a learning algorithm A that generates the tightest circle which contains all positive instances from training set S is an ERM.

   For this problem, it can be seen that the misleading set includes all training sets that lead to a hypothesis which outputs a circle with a radius that is too small resulting in $L_{\mathcal{D}}(h) > \epsilon$. This radius can be considered the "threshold" radius and be denoted by r^-. We can then define the probability of the region between r^- and r^* as ϵ as shown below.

$$\mathop{\mathbb{P}}_{x\sim\mathcal{D}}\left[x \in \mathbb{R}^2 : r^- \leq ||x|| \leq r^*\right] = \epsilon$$

   From here, we can calculate the probability of sampling from the misleading set $\mathcal{M}$ which would be the complement region and compute the upper bound of the sample complexity.

$$\mathop{\mathbb{P}}_{S|x\sim\mathcal{D}^m}[S|x : S|x \in \mathcal{M}] = (1-\epsilon)^m \leq e^{-\epsilon m}$$

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

6. Exercise **3.4**

   We can see that the hypothesis class $\mathcal{H}$ is a finite class. The number of total hypotheses include the total number of possible Boolean conjunctions provided d variables plus the all-negative hypothesis. Therefore, $|\mathcal{H}| = 3^d + 1$. With this we have proven that $\mathcal{H}$ is PAC learnable and can calculate the sample complexity as the following.

$$m_{\mathcal{H}}(\epsilon,\delta) \leq \left\lceil \frac{\log((3^d+1)/\delta)}{\epsilon} \right\rceil$$

   We can use an algorithm A that scans through each instance x_i of length d and continuously removes hypotheses that contradict the label y_i. For example, if $y_i = 1$, and the first element of of the first instance $x_{i,1} = 0$, then remove all hypotheses whose conjunction contains $x_{i,1}$. It is clear to see that such an algorithm is ERM. As this algorithm is linear with the instance length d and there are m instances, the algorithm has a runtime of $O(md)$.

7. Exercise **3.5**

   Note that $L_{\bar{\mathcal{D}}_{m,f}} > \epsilon$ implies

$$\frac{\sum_{i=1}^{m} \mathbb{P}_{x\sim\mathcal{D}_i}[h(x) = f(x)]}{m} < 1 - \epsilon$$

Therefore, we can derive the following.

$$\begin{aligned}
\mathbb{P}[L_{\bar{\mathcal{D}}_m,f}(h) > \epsilon, L_s(h) = 0] &= \prod_{i=1}^{m} \mathop{\mathbb{P}}_{x\sim\mathcal{D}_i}[h(x) = f(x)] \\
&= \left(\left(\prod_{i=1}^{m} \mathop{\mathbb{P}}_{x\sim\mathcal{D}_i}[h(x) = f(x)] \right)^{\frac{1}{m}} \right)^m \\
&\leq \left(\frac{\sum_{i=1}^{m} \mathbb{P}_{x\sim\mathcal{D}_i}[h(x) = f(x)]}{m} \right)^m \\
&< (1-\epsilon)^m \\
&\leq e^{-\epsilon m} \\
&\leq |\mathcal{H}| e^{-\epsilon m}
\end{aligned}$$

8. Exercise **3.6**

   Suppose $\mathcal{H}$ is agnostic PAC learnable and let A be a learning algorithm that learns $\mathcal{H}$ with sample complexity $m_{\mathcal{H}}(\epsilon, \delta)$. Then if $\mathcal{H}$ is agnostic PAC learnable, $\mathcal{H}$ is also PAC learnable under the realizability assumption. Recall the definition of agnostic PAC learnable where with probability of at least $1 - \delta$,

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

   We can assume that there exists a labeling function f that produces y. Furthermore, under the realizability assumption, we have that there exists an $h^* \in \mathcal{H}$ such that $L_{\mathcal{D}}(h^*) = 0$. Then, for every $m > m_{\mathcal{H}}(\epsilon, \delta)$, if A is provided training set S of m i.i.d. instances labelled by f, then with probability of at least $1 - \delta$,

   $$\begin{aligned}
   L_{\mathcal{D}}(A(S)) &\leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \\
   &= \epsilon
   \end{aligned}$$

9. Exercise **4.1**

   For any learning algorithm A, distribution $\mathcal{D}$, and loss function whose range is $[0, 1]$, show that for every $\epsilon, \delta > 0$, there exists $m(\epsilon, \delta)$ such that $\forall m \geq m(\epsilon, \delta)$

   $$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) > \epsilon] < \delta$$

   Show that this is equivalent to

   $$\lim_{m\to\infty} \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] = 0$$

   We can solve this by using Markov's inequality.

   $$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) > \epsilon] \leq \frac{\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))]}{\epsilon}$$

Assuming the above limit statement is true, it can be be assumed that there exists a m^* such that

$$\lim_{m \to m^*} \mathop{\mathbb{E}}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(A(S))] = \delta\epsilon$$

We can then conclude the proof by the following where $\forall m \geq m^*$

$$\begin{aligned} \mathop{\mathbb{P}}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(A(S)) > \epsilon] &\leq \frac{\mathbb{E}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(A(S))]}{\epsilon} \\ &= \frac{\epsilon\delta}{\epsilon} \\ &= \delta \end{aligned}$$

10. Exercise **4.2**

Prove that if a loss function has range $[a, b]$, then the sample complexity satisfies

$$m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta) \leq \left\lceil \frac{2\log(2|\mathcal{H}|/\delta)(b-a)^2}{\epsilon^2} \right\rceil$$

We can prove this by using Hoeffding's Inequality. Assume a loss function with range $[a, b]$. We can then obtain the following.

$$\mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) = \mathbb{P}\left[|\frac{1}{m}\sum_{i=1}^{m} \theta_i - \mu| > \epsilon \right] \leq 2\exp(-2m\epsilon^2/(b-a)^2)$$

Using the union bound, we then obtain the following.

$$\begin{aligned} \mathcal{D}^m(\{S : \exists h \in \mathcal{H} | L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) &\leq \sum_{h \in \mathcal{H}} 2\exp(-2m\epsilon^2/(b-a)^2) \\ &= 2|\mathcal{H}|\exp(-2m\epsilon^2/(b-a)^2) \end{aligned}$$

We can then conclude the proof by setting the right side equal to δ and solving for $m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$.

$$\begin{aligned} 2|\mathcal{H}|\exp(-2m_{\mathcal{H}}^{UC}(\epsilon/2)^2/(b-a)^2) &= \delta \\ -2m_{\mathcal{H}}^{UC}(\epsilon/2)^2/(b-a)^2 &= \log\left(\frac{\delta}{2|\mathcal{H}|}\right) \\ m_{\mathcal{H}}^{UC} &= \frac{2\log\left(2|\mathcal{H}|/\delta\right)(b-a)^2}{\epsilon^2} \end{aligned}$$