CS 260: Machine Learning	**Spring 2020**
Homework 1	
Hand Out: March.30	*Due: April.15*

1. Exercise 2.2 Answer:
   By definition it holds that

$$L_S(h) = \frac{|i \in [m] : h(x_i) \neq y_i|}{m} = \frac{1}{m}\sum_{i=1}^{m} 1\{h(x_i) \neq y_i\},$$

   where $1\{A\}$ is indicator function of set A, and examples in S are i.i.d. according to distribution $\mathcal{D}$. We directly have:

$$\begin{aligned}
\mathbb{E}_{S|_x \sim \mathcal{D}^m}[L_S(h)] &= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{x_i \sim \mathcal{D}}[1(h(x_i) \neq y_i)] \\
&= \frac{1}{m}\sum_{i=1}^{m} 1 \cdot P_{x_i \sim \mathcal{D}}(h(x_i) \neq y_i) + 0 \cdot P_{x_i \sim \mathcal{D}}(h(x_i) = y_i) \\
&= \frac{1}{m} \cdot m \cdot P_{x \sim \mathcal{D}}[h(x) \neq y] \\
&= L_{(\mathcal{D},f)}(h). \blacksquare
\end{aligned}$$

2. Exercise 2.3 Answer:

   1. Note that A labels all the positive instances in the training set correctly. Furthermore, by the realizability assumption, all the negative instances should be outside the tightest rectangle that contains all the positive instances. Thus A also labels all the negative instances in the training set correctly. So A is an ERM. $\blacksquare$
   2. Note that all the positive instances are in R^* while algorithm A can only have access to positive instances in the training set which is a subset of theses positive instances. Since $R(S)$ is the smallest rectangle enclosing the positive instances in the training set, thus $R(S) \subseteq R^*$.

      Based on the definition of $L_{(\mathcal{D},f)}$, we get

$$L_{(\mathcal{D},f)}(R(S)) = \mathcal{D}(R^* \setminus R(S)).$$

      For any fixed $0 < \epsilon < 1$ and rectangles $R_1, R_2, R_3,\ R_4$ defined as in the hint. Define F_i as the set of instances that do not fall in R_i, i.e.,

$$F_i = \{S|_x : S|_x \cap R_i = \phi\}.$$

      Applying the union bound yields

$$\mathcal{D}^m(\{S : L_{(\mathcal{D},f)}(A(S)) > \epsilon\}) \leq \mathcal{D}^m\big(\cup_{i=1}^{4} F_i\big) \leq \sum_{i=1}^{4} \mathcal{D}^m(F_i). \tag{1}$$

Next, we only need to ensure that $\mathcal{D}^m(F_i) \leq \delta/4$. According to the definition of F_i, we actually have

$$\mathcal{D}^m(F_i) = (1-\epsilon/4)^m \leq \exp(-m\epsilon/4).$$

Plugging the above result into (1), we arrive at the conclusion that A returns a hypothesis with error at most ϵ with probability at least $1-\delta$ if the training set size is greater than $4\log(4/\delta)/\epsilon$. ■

3. We can define a similar hypothesis class as follows,

$$h_{(a_1,b_1,\ldots,a_d,b_d)} = \begin{cases} 1, & \text{if } \forall i, a_i \leq x_i \leq b_i \\ 0, & \text{otherwise.} \end{cases}$$

The only difference is that instead of 4 strips, we have $2d$ strips (2 strips for each dimension). Replacing 4 with $2d$ in previous result yields the conclusion. ■

4. The algorithm A returns the smallest polyhedron enclosing all positive examples, so A only need to traverse all examples in training set and find the minimum and maximum in each dimension, thus the runtime is:

$$O(md) = O\left(\frac{d^2\log(2d/\delta)}{\epsilon}\right)$$

which is polynomial in $d, 1/\epsilon, \log(1/\delta)$. ■

3. Exercise 3.1 Answer:
Use $\mathcal{D}$ to denote the unknown distribution over $\mathcal{X}$, and use $f \in \mathcal{H}$ to denote the target hypothesis. Denote A as the algorithm that learns $\mathcal{H}$ with sample complexity $m_{\mathcal{H}}(\cdot,\cdot)$. Given some fixed $\delta \in (0,1)$, and suppose $0 < \epsilon_1 \leq \epsilon_2 \leq 1$. We define $m_1 = m_{\mathcal{H}}(\epsilon_1, \delta), m_2 = m_{\mathcal{H}}(\epsilon_2, \delta)$. Based on the definition, given an i.i.d. training sequence of size m that satisfies $m = m_1$, we can guarantee that, with probability at least $1-\delta$, the algortihm A will return a hypothesis h that satisfies

$$L_{(\mathcal{D},f)}(h) \leq \epsilon_1 \leq \epsilon_2.$$

According to the definition and minimality of m_2, we can derive that $m_2 \leq m_1$. Similarly, we can prove another argument and omit here.

4. Exercise 3.2 Answer:

1. Realizability assumption ensures that ERM should output a hypothesis h^* such that $L_S(h^*) = 0$. The proposed algorithm selects hypothesis h_z if some positive instance $z \in \mathcal{X}$ appears in the training set, otherwise it selects all-negative hypothesis h^-. Clearly, this is an ERM algorithm.
2. We want to prove:
$$\mathcal{D}^m(\{S|_x : L_{\mathcal{D},f}(A(S)) > \epsilon\}) \leq \delta$$
Let $\mathcal{H}_B$ be the set of bad hypothesis:
$$\mathcal{H}_B = \{h \in \mathcal{H}_B : L_{\mathcal{D},f}(h) > \epsilon\}$$

Let M be the set of misleading samples:

$$M = \{S|_x : \exists h \in \mathcal{H}_B, L_S(h) = 0\}$$

We observe that:

$$\{S|_x : L_{\mathcal{D},f}(A(S)) > \epsilon\} \subseteq M = \bigcup_{h \in \mathcal{H}_B} \{S|_x : L_S(h) = 0\}$$

If h^- is the true hypothesis, then no positive examples exists in training set. Thus no misleading samples, $M = \emptyset$, and $\mathcal{D}(\{S|_x : L_{\mathcal{D},f}(A(S)) > \epsilon\}) \leq \mathcal{D}(M) = 0$. If h_z is the true hypothesis, and z does not appear in training set (misleading samples), then algorithm A might return the hypothesis h^-, which would be the only bad hypothesis if $L_{\mathcal{D},f}(h^-) = \mathcal{D}(\{z\}) > \epsilon$:

$$\begin{aligned}\mathcal{D}^m(\{S|_x : L_{\mathcal{D},f}(A(S)) > \epsilon\}) &\leq \mathcal{D}^m(M) \\ &= \mathcal{D}^m(\{S|_x : L_S(h^-) = 0\}) \\ &= (1-\epsilon)^m \leq e^{-\epsilon m} \leq \delta\end{aligned}$$

Thus $\mathcal{H}_{singleton}$ is PAC learnable with existing $m_{\mathcal{H}}(\epsilon, \delta) = \lceil \frac{\log(1/\delta)}{\epsilon} \rceil$. ■

5. Exercise 3.3 Answer:
Realizability assumes that there is a true hypothesis h_{r*}. Suppose algorithm A returns the smallest circle enclosing all positive training points, and denote the hypothesis as $A(S) = h_r$. Training samples would be misleading if none of them appears in the area $\{x \in \mathbb{R}^2 : r \leq \|x\| \leq r*\}$, and h_r would be the bad hypothesis. The probability of an example appearing in that area is: $\mathcal{D}(\{x : r \leq \|x\| \leq r*\}) = L_{\mathcal{D},f}(h_r) > \epsilon$, then:

$$\begin{aligned}\mathcal{D}^m(\{S|_x : L_{\mathcal{D},f}(A(S)) > \epsilon\}) &\leq \mathcal{D}^m(\{S|_x : L_S(h_r) = 0\}) \\ &= (1 - L_{\mathcal{D},f}(h_r))^m \\ &= (1-\epsilon)^m \leq e^{-\epsilon m} \leq \delta\end{aligned}$$

Thus $\mathcal{H}$ is PAC learnable with sample complexity $m_{\mathcal{H}}(\epsilon, \delta) = \lceil \frac{\log(1/\delta)}{\epsilon} \rceil$. ■

6. Exercise 3.4 Answer:
The hypothesis class contains all conjunctions of literals over the d variables. We can prove that it is finite, because each literal i has 3 choices: $x_i, \bar{x}_i$ or neither of these appear in the corresponding conjunction. Therefore, we can derive the upper bound of the whole hypothesis $|\mathcal{H}| = 3^d + 1$, which means that $|\mathcal{H}|$ is finite. According to Corollary 3.2, we can get that $\mathcal{H}$ is PAC learnable and the related sample complexity is bounded by

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

The algorithm is implemented as follows: denote the training sequence as $S = \big((\mathbf{x}^1, y^1), ..., (\mathbf{x}^m, y^m)\big)$. We start with the hypothesis h_0 that

$$h_0 = x_1 \wedge \bar{x}_1 \wedge \ldots \wedge x_d \wedge \bar{x}_d.$$

For each $(\mathbf{x}^j, y^j)$, if $y^j = 1$, then we remove $\bar{x}_i$ if $\mathbf{x}_i^j = 1$, then we remove x_i if $\mathbf{x}_i^j = 0$. Then algorithm A is able to return the target hypothesis for S. It is easy to get that the runtime is linear in $m \cdot d$. ■

7. Exercise 3.5 Answer:
To start with, if $L_{(\bar{\mathcal{D}}_m,f)}(h) > \epsilon$, based on the definition of $\bar{\mathcal{D}}_m$, we can get that

$$\frac{\mathbb{P}_{X\sim\mathcal{D}_1}[h(X) = f(X)] + ... + \mathbb{P}_{X\sim\mathcal{D}_m}[h(X) = f(X)]}{m} \leq 1 - \epsilon, \tag{2}$$

where the inequality follows from the fact that $L_{(\bar{\mathcal{D}}_m,f)}(h) > \epsilon$. On the other hand, we have

$$\begin{aligned}
\mathbb{P}_{S\sim\Pi_{i=1}^m\mathcal{D}_i}[L_S(h) = 0] &= \Pi_{i=1}^m \mathbb{P}_{X\sim\mathcal{D}_i}[h(X) = f(X)] \\
&= \left(\left(\Pi_{i=1}^m \mathbb{P}_{X\sim\mathcal{D}_i}[h(X) = f(X)]\right)^{1/m}\right)^m \\
&\leq \left(\frac{\sum_{i=1}^m \mathbb{P}_{X\sim\mathcal{D}_i}[h(X) = f(X)]}{m}\right)^m. \qquad (3)
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{P}_{S\sim\Pi_{i=1}^m\mathcal{D}_i}[L_S(h) = 0 \text{ and } L_{\bar{\mathcal{D}}_m,f}(h) > \epsilon] &= \mathbb{P}_{S\sim\Pi_{i=1}^m\mathcal{D}_i}[L_S(h) = 0 | L_{\bar{\mathcal{D}},f}(h) > \epsilon]\mathbb{P}[L_{\bar{\mathcal{D}},f}(h) > \epsilon] \\
&\leq \mathbb{P}_{S\sim\Pi_{i=1}^m\mathcal{D}_i}[L_S(h) = 0 | L_{\bar{\mathcal{D}},f}(h) > \epsilon] \\
&\leq (1-\epsilon)^m \\
&\leq e^{-m\epsilon}
\end{aligned}$$

the first inequality is based on (3), and the second inequality follows from (2). This completes the proof. ■

8. Exercise 3.6 Answer:
First we assume that $\mathcal{H}$ is agnostic PAC learnable, and denote A as the learning algorithm that learns $\mathcal{H}$ with sample complexity $m_{\mathcal{H}}(\cdot,\cdot)$. Next, we try to show that $\mathcal{H}$ is PAC learnable using the algorithm A.

Here we use $\mathcal{D}$ be the distribution over $\mathcal{X}$, and f be the target function. We assume that $\mathcal{D}$ is a joint distribution over $\mathcal{X} \times \{0,1\}$. Based on the realizability assumption, we have $\inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) = 0$. Let $\epsilon, \delta \in (0,1)$, for the sample size $m \geq m_{\mathcal{H}}(\epsilon,\delta)$, then with probability at least $1-\delta$, it returns a hypothesis h^* that satisfies

$$\begin{aligned}
L_{\mathcal{D}}(h^*) &\leq \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \\
&= 0 + \epsilon \\
&= \epsilon,
\end{aligned}$$

which concludes our proof. ■

9. Exercise 4.1 Answer:
Proof:

- We only need to prove that for any $\eta > 0$, we have $\mathbb{E}_{S\sim\mathcal{D}^m} L_{\mathcal{D}}(A(S)) \leq \eta$ for sufficient large m. Note that by the conclusion in (1), for any fixed $\eta \in (0,1)$,

and every distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, there exists $m(\eta/2, \eta/2)$ such that for every $m \geq m(\eta/2, \eta/2)$, we have

$$\mathbb{P}_{S\sim\mathcal{D}_m}\left[L_{\mathcal{D}}(A(S)) > \frac{\eta}{2}\right] < \frac{\eta}{2}.$$

Note that $L_{\mathcal{D}}(A(S)) < 1$, we have

$$\begin{aligned}\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] &< \mathbb{P}(L_{\mathcal{D}}(A(S)) > \eta/2) \cdot 1 + \mathbb{P}(L_{\mathcal{D}}(A(S)) \leq \eta/2) \cdot \eta/2 \\ &\leq \frac{\eta}{2} + \frac{\eta}{2} \\ &= \eta.\end{aligned}$$

- By Markov's inequality, we have

$$\mathbb{P}(L_{\mathcal{D}}(A(S)) > \epsilon) \leq \mathbb{E}[L_{\mathcal{D}}(A(S))]/\epsilon.$$

We only need to choose big enough m, let $\mathbb{E}[L_{\mathcal{D}}(A(S))] < \epsilon\delta$, then we have our conclusion.

■

10. Exercise 4.2 Answer:
Proof: The first inequality holds trivialy. By Hoeffding's inequality, we have

$$\mathbb{P}(|L_{\mathcal{D}}(h) - L_S(h)| > \epsilon/2) \leq 2\exp(-2m\epsilon^2/(b-a)^2).$$

Then we make RHS to be at most $\delta/|\mathcal{H}|$, applying union bound, we get the answer. ■