3.5 Exercises

① Monotonicity of Sample Complexity

Let $\mathcal{H}$ be a hypothesis class for a binary classification task. Suppose that $\mathcal{H}$ is PAC learnable and its sample complexity is given by $m_{\mathcal{H}}(\cdot,\cdot)$

Show that $m_{\mathcal{H}}$ is monotonically nonincreasing in each of its parameters.

In other words, show that:

Given $\begin{cases} \delta \in (0,1) \\ 0 < \varepsilon_1 \leq \varepsilon_2 < 1 \end{cases}$

We have $m_{\mathcal{H}}(\varepsilon_1, \delta) \geq m_{\mathcal{H}}(\varepsilon_2, \delta)$

Similarly, show that:

Given $\begin{cases} \varepsilon \in (0,1) \\ 0 < \delta_1 \leq \delta_2 < 1 \end{cases}$

We have $m_{\mathcal{H}}(\varepsilon, \delta_1) \geq m_{\mathcal{H}}(\varepsilon, \delta_2)$

Solution:

- The proof follows from the definition, we restated the PAC learnability definition with realizability assumption (for simplicity), with this adjustments: $\begin{cases} 0 < \varepsilon_1 \leq \varepsilon_2 < 1 \\ \text{leads to } m_1 \overset{\text{def}}{=} m_{\mathcal{H}}(\varepsilon_1, \delta) \geq m_{\mathcal{H}}(\varepsilon_2, \delta) \overset{\text{def}}{=} m_2 \end{cases}$

- For every $\delta \in (0,1)$, $0 < \varepsilon_1 \leq \varepsilon_2 < 1$ and distribution $\mathcal{D}$ over $\mathcal{X}$, when running the learning algorithm on $m \geq m_1 \overset{\text{def}}{=} m_{\mathcal{H}}(\varepsilon_1, \delta) \geq m_{\mathcal{H}}(\varepsilon_2, \delta) \overset{\text{def}}{=} m_2$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns a hypothesis h such that, with probability at least $1-\delta$:

$$L_{\mathcal{D}}(h) \leq \varepsilon_1 \leq \varepsilon_2$$

By the minimality of m_2, we conclude that $m_2 \leq m_1$

human language explanation

Human languague explanation:

- Given sample size $m \geq m_1$, we will have $L_D(h) \leq \varepsilon_1$
- Since $\varepsilon_1 \leq \varepsilon_2$, any hypothesis h that is $L_D(h) \leq \varepsilon_1$, also satisfy $L_D(h) \leq \varepsilon_2$. And since $m \geq m_1$ is required in the first case, it also satisfy the second (looser) condition, $m \geq m_1 \geq m_2$
- "Minimality of m_2": means that m_2 is the smallest possible sample size such that accuracy ε_2 is guaranteed.

2 Show that a hypothesis class is PAC learnable (with realizability)

Let $\begin{cases} X \text{ be a discrete domain} \\ H_{singleton} = \underbrace{\{h_z : z \in X\}} \cup \underbrace{\{h^-\}} \end{cases}$

Each $z \in X$, $h_z(x) = \begin{cases} 1 & \text{if } x = z \\ 0 & \text{otherwise} \end{cases}$ $\quad h^-(x) = 0 \;\; \forall x \in X$

for all x, nothing is special

for each z, there is exactly one h_z that label it "special"

Realizability assumption holds

that means at least 1 h in $H_{singleton}$ satisfy $L_D(h) = 0$.

a) Describe an algorithm that implements the ERM rule for learning $H_{singleton}$ in the realizable setup

- Since realizability holds, we need to come up with an algorithm s.t. $L_S(h) = 0$
- Let A be the algorithm that returns hypothesis h_S with the following property:

$$h_S = \begin{cases} h_x & \text{if } \exists x \in S \text{ s.t. } f(x) = 1 \\ h^- & \text{otherwise} \end{cases}$$

Clearly, $L_S(h_S) = 0$, and so A is ERM

b) Show that $\mathcal{H}_{singleton}$ is PAC learnable. Provide an upper bound on the sample complexity

- Let D be distribution over X and $\varepsilon \in (0,1)$
- Based on the definition of A in previous section:
  - If $f = h^-$, then A return true hypothesis, so $L_{D,f}(h^-) = 0$
  - Suppose $\exists\, x \in X$ s.t. $f(x) = 1$

    Let $S|_x = (x_1, \dots, x_m)$ be instances of training set S

    We try to upper bound $D^m(\{S|_x : L_{D,f}(h_S) > \varepsilon\})$ → How bad training set S represent D
    - If $\exists\, x \in S|_x$, then A returns true hypothesis, so $L_{D,f}(h_S) = 0$
    - So the only scenario left for us to upper bound on is when $\exists\, x \notin S|_x$
- Also, it can be proven that:

  $D(x) = L_{D,f}(h)$ —proof→

Lets analyze possible values of true error on individual instance x:

- all negative case h^-: $L_{x\sim D}(h^-) = 0$
- contain "special" instance:
  - Correctly identify "special" instance: $L_{x\sim D}(h_S) = 0$
  - Wrongly identify "special" instance: $L_{x\sim D}(h_S) = 1$

So possible values are 0 and 1:

By LOTE and definition of A:

$$L_{D,f}(h) = 1 \cdot D(x) + 0 \cdot D^c(x) = D(x)$$

$\Rightarrow D(x) \leqslant \varepsilon$ means $L_{D,f}(h) \leqslant \varepsilon$

$\Rightarrow D(x) > \varepsilon$ means $L_{D,f}(h) > \varepsilon$

$\Leftrightarrow D(x') \leqslant 1-\varepsilon$ means $L_{D,f}(h) > \varepsilon$
$\forall x' \in X|_x$

- Combine 2 points above, we have:

$$\{S|_x : L_{D,f}(h) > \varepsilon\} = \{S|_x : \forall x' \in S|_x,\ D(x') \leqslant 1-\varepsilon\}$$

And so:

$$D^m(\{S|_x : L_{D,f}(h) > \varepsilon\}) = D^m(\{S|_x : \forall x' \in S|_x,\ D(x') \leqslant 1-\varepsilon\}) \leqslant (1-\varepsilon)^m \leqslant e^{-\varepsilon m}$$

- Let $\delta \in (0,1)$ s.t. $e^{-\varepsilon m} \leqslant \delta$. We can conclude that:

$$m \geqslant \frac{\log(1/\delta)}{\varepsilon}$$

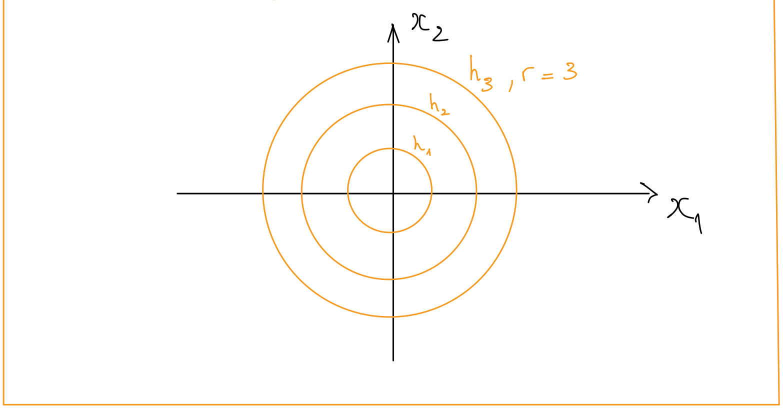
Therefore, $\mathcal{H}_{singleton}$ is PAC learnable with $m_{\mathcal{H}_{singleton}} \leqslant \left\lceil \frac{\log(1/\delta)}{\varepsilon} \right\rceil$

(3) Another prove PAC learnable (with realizability)

Let $\begin{cases} \mathcal{X} = \mathbb{R}^2, \quad \mathcal{Y} = \{0, 1\} \\ \mathcal{H} \text{ be hypothesis class of concentric circles in the plane:} \\ \quad \mathcal{H} = \{h_r : r \in \mathbb{R}_+\}, \text{ where } h_r(x) = \mathbb{1}_{\|x\| \leqslant r} \end{cases}$

Prove that $\begin{cases} \mathcal{H} \text{ is PAC learnable (assume realizability)} \\ \mathcal{H} \text{ sample complexity is bounded by } m_{\mathcal{H}}(\varepsilon, \delta) \leqslant \left| \frac{\log(1/\delta)}{\varepsilon} \right| \end{cases}$

Visualize hypothesis class $\mathcal{H}$



Solution: Proving PAC learnability has 2 steps
- Come up with alg A that is ERM
- Upper bound sample complexity

- Let D be the distribution over $\mathcal{X}$, $\varepsilon \in (0, 1)$ and f be target hypothesis
- Let A be algorithm that returns the smallest circle enclosing all the positive examples from the training set S, where $\begin{cases} C(S) \text{ be the circle returned} \\ r(S) \text{ be the circle's radius} \\ A(S) : \mathcal{X} \to \mathcal{Y} \text{ is the hypothesis} \end{cases}$

We can easily see that $L_S(A(S)) = 0$, so A is ERM

Proof:
- Since A returns positive examples: $A(S)(x) = 1$
- Realizability assumption: $\exists h^* \in \mathcal{H}$ s.t. $h^*(x) = 1$

$\Rightarrow A(S) = h^*$, and so $L_S(A(S)) = 0$

conclude A is ERM

- By realizability assumption, $\exists\, h^* \in \mathcal{H}$ s.t. $L_{D,f}(h^*) = 0$

Let $\begin{cases} C^* & \text{be the circle corresponds to hypothesis } h^* \\ r^* & \text{be the corresponding radius} \end{cases}$

It can be proven that $C(S) \subseteq C^*$, where C^* is the circle enclosing all the positive examples.

<u>Proof:</u> $\left.\begin{array}{l} \circ\ C(S) \text{ is the smallest circle enclosing positive examples} \\ \circ\ C^* \text{ is the circle enclosing positive examples} \end{array}\right\}$

$\Rightarrow \quad C(S) \subseteq C^*$

- Since $C(S) \subseteq C^*$, we can prove that $L_{D,f}(A(S)) = D(C^* \setminus C(S))$

<u>Proof:</u> $L_{D,f}(A(S)) = D(\{x \in X : A(S)(x) \neq f(x)\})$ <definition>

$= D(\{x \in X : x \notin S|_x \text{ and } f(x) = 1\})$

$= D(\{C^* \setminus C(S)\})$

Why?

$L_{D,f}(A(S))$ consists of:

- $x \in S|_x$ s.t. $f(x) = 0$
- $x \notin S|_x$ s.t. $f(x) = 1$

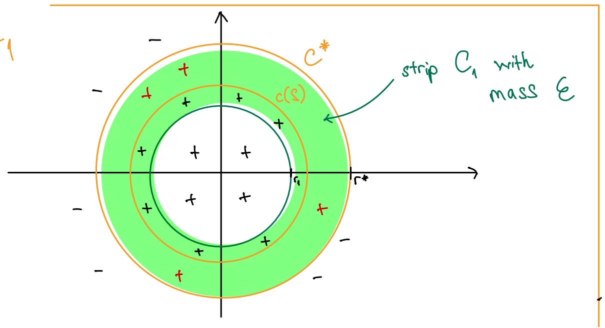
However, $C(S) \subseteq C^*$ means fist case cannot happens

- Let $r_1 \leqslant r^*$ be a number such that the probability mass of C_1 is ε.

where corresponding strip $C_1 = \{x \in \mathbb{R}^2 : r_1 \leqslant \|x\| \leqslant r^*\}$

Visualize C_1

+ : contributes to error



- From the visualization, we can prove that:

$$L_{D,f}(A(S)) \leq \varepsilon$$

<u>Proof:</u>

$$D(C^*) - D(C(S)) \leq D(C_1)$$

$$\Leftrightarrow \underbrace{D(C^* \setminus C(S))}_{\text{positives outside of } C(S)} \leq \underbrace{D(C_1)}_{\text{all positives in strip } C_1}$$

$$\Leftrightarrow L_{D,f}(A(S)) \leq \varepsilon$$ ⟨previous discussion, subs $D(C_1)$⟩

- Now, we would like to upper bound $D^m(\{S|_x : L_{D,f}(h_S) > \varepsilon\})$.

With the discussion above, we can prove that:

$$\{S|_x : L_{D,f}(h_S) > \varepsilon\} = \{S|_x : S|_x \cap C_1 = \emptyset\}$$

↳ there are positives in the strip

<u>Proof:</u>

Since $\{S|_x : L_{D,f}(h_S) \leq \varepsilon\} = \{S|_x : S|_x \cap C_1 \neq \emptyset\}$

↳ there are no positives in the strip

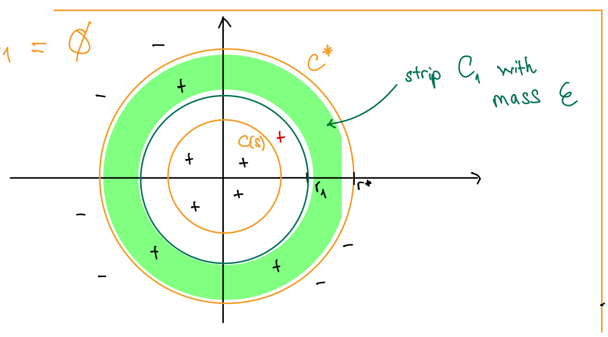
So the opposite holds:

$$\{S|_x : L_{D,f}(h_S) > \varepsilon\} = \{S|_x : S|_x \cap C_1 = \emptyset\}$$

↳ bad event

Visualize $S|_x \cap C_1 = \emptyset$

$+$: contributes to error



- Therefore, we can conclude that:

$$D^m(\{S|_x : L_{D,f}(h_S) > \varepsilon\}) = D^m(\{S|_x : S|_x \cap C_1 = \emptyset\})$$

$$\leq (1-\varepsilon)^m$$

$$\leq e^{-\varepsilon m}$$

Let $\delta \in (0,1)$ such that $e^{-\varepsilon m} \leq \delta$, then $m \geq \frac{\log(1/\delta)}{\varepsilon}$, and so

$\mathcal{H}$ is PAC learnable with $m_{\mathcal{H}} \leq \left\lceil \frac{\log(1/\delta)}{\varepsilon} \right\rceil$