2.4 Exercises

2.4.1. Overfitting of polynomial matching:

- We have shown that the predictor defined as:

$$h_S(x) = \begin{cases} y_i & \text{if } \exists i \in [m] \text{ s.t. } x_i = x \\ 0 & \text{otherwise} \end{cases}$$

will leads to overfit. The goal of this exercise is to show that this predictor can be described as a threshold polynomial.

- Show that given a training set $S = \{(x_i, f(x_i))\}_{i=1}^{m} \subseteq (\mathbb{R}^d \times \{0,1\})^m$ there exists a polynomial p_S such that $h_S(x) = 1$ if and only if $p_S(x) \geq 0$, where h_S is defined as above.

Solution:

- We will prove the statement: $h_S(x) = 1$ i.o.i. $p_S(x) \geq 0$ by consider these 4 cases:

$$\begin{cases} \cdot\ m = 1 \text{ and } S = ((x_1, 0)) \\ \cdot\ m = 1 \text{ and } S = ((x_1, 1)) \\ \cdot\ m = 2 \text{ and } S = ((x_1, 1), (x_2, 0)) \\ \cdot\ m = 2 \text{ and } S = ((x_1, 1), (x_2, 1)) \end{cases}$$

and suggest what p_S looks like along the way.

Suppose that $\exists i \in [m]$ s.t. $y_i = 1$, then:

- Consider the case $m=1$ and $S = ((x_1, 0))$

$h_S(x) = 0$ for any $x \in X$

In this case, define $p_S(x) := -1$ or any other value < 0 will make the statement true

next page

- Consider the case $m = 1$ and $S = ((x_1, 1))$:

$$\underline{h_S(x)} = \begin{cases} 1 & \text{if } x = x_1 \\ 0 & \text{otherwise (meaning } x \in X \setminus \{x_1\}) \end{cases}$$

In human language, $h_S(x)$ equals 1 if instance x is in the training set, equals 0 if instance x is not in training set

Define $p_S(x) := -\|x - x_1\|^2$. We can see that :

$$p_S(x) = \begin{cases} 0 & \text{if } x = x_1 \\ < 0 & \text{otherwise } (x \in X \setminus \{x_1\}) \end{cases}$$

- Consider the case $m = 2$ and $\begin{cases} S = ((x_1, 1), (x_2, 0)) \\ \text{or } S = ((x_1, 1), (x_2, 1)) \end{cases}$

- Case $S = ((x_1, 1), (x_2, 0))$ (1)

$$h_S(x) = \begin{cases} 1 & \text{if } x = x_1 \\ 0 & \text{otherwise } (x \in X \setminus \{x_1\}) \end{cases}$$

So similar to case $m=1$, define $p_S(x) := -\|x - x_i\|^2$ will lead to similar result :

$$p_S(x) = \begin{cases} 0 & \text{if } x = x_1 \\ < 0 & \text{otherwise } (x \in X \setminus \{x_1\}) \end{cases}$$

So the statement is true

- Case $S = ((x_1, 1), (x_2, 1))$ (2)

$$h_S(x) = \begin{cases} 1 & \text{if } x \in \{x_1, x_2\} \\ 0 & \text{otherwise } (x \in X \setminus \{x_1, x_2\}) \end{cases}$$

Define $p_S(x) := -(\|x - x_1\|^2)(\|x - x_2\|^2)$, we can see that:

$$p_S(x) = \begin{cases} 0 & \text{if } x \in \{x_1, x_2\} \\ < 0 & \text{otherwise } (x \in X \setminus \{x_1, x_2\}) \end{cases}$$

So the statement is true

↳ continue

Using induction from (1) and (2), we can generalize p_S as:

$$p_S(x) := -\prod_{i \in [m],\, y_i = 1} \| x - x_i \|^2$$

Then $p_S(x)$ is a polynomial such that

$$p_S(x) = \begin{cases} 0 & \forall i \in [m] \text{ s.t. } y_i = 1 \text{ and } x = x_i \\ < 0 & \text{otherwise } (x \in X \setminus \{ x_i \mid i \in [m] \text{ and } y_i = 1 \}) \end{cases}$$

2.4.2 Linearity of expectation

Let $\begin{cases} H \text{ be a class of binary classifiers over a domain } X. \\ D \text{ be an unknown distribution over } X \\ f \text{ be the target hypothesis in } H \end{cases}$

Fix some $h \in H$, show that the expected value of $L_S(h)$ over the choice of $S|_x$ equals $L_{D,f}(h)$, namely:

$$\mathbb{E}_{S|_x \sim D^m} [L_S(h)] = L_{D,f}(h)$$

Solution:

$$\mathbb{E}_{S|_x \sim D^m} [L_S(h)] = \mathbb{E}_{S|_x \sim D^m} \left[\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{h(x_i) \neq f(x_i)} \right] \quad \langle \text{definition of } L_S(h) \rangle$$

$$= \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{x_i \sim D} \left[\mathbb{1}_{h(x_i) \neq f(x_i)} \right] \quad \langle \text{linearity of expectation} \rangle$$

$$= \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{x \sim D} \left[\mathbb{1}_{h(x) \neq f(x)} \right] \quad \langle x_1, \dots, x_m \text{ are i.i.d } \mathbb{E}[g(x_1)] = \mathbb{E}[g(x_2)] \text{ for any function } g$$

$$= m \left(\frac{1}{m} \right) \underbrace{\mathbb{P}_{x \sim D} [h(x) \neq f(x)]}_{\text{definition of true error}}$$

$$= L_{D,f}(h)$$

2.4.3 Axis aligned rectangles:

Given an axis aligned rectangle classifier defined as:

$$h_{(a_1, b_1, a_2, b_2)}(x_1, x_2) = \begin{cases} 1 & \text{if } a_1 \leqslant x_1 \leqslant b_1 \text{ and } a_2 \leqslant x_2 \leqslant b_2 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{where } \begin{cases} a_1, b_1, a_2, b_2 \in \mathbb{R} \\ a_1 \leqslant b_1, \quad a_2 \leqslant b_2 \end{cases}$$

The hypothesis class of all axis aligned rectangles is defined as:

$$\mathcal{H}^2_{rec} = \{ h_{(a_1, b_1, a_2, b_2)} : a_1 \leqslant b_1 \text{ and } a_2 \leqslant b_2 \}$$

Note that this is an infinite hypothesis class

Throughout this exercise we use realizability assumption, which state:

$$\exists h^* \in \mathcal{H}^2_{rec} \quad \text{s.t.} \quad L_{D,f}(h^*) = 0,$$

and $L_S(h^*) = 0$ where training set S is sampled according to D and labeled by f

① Let A be the algorithm that returns the smallest rectangle enclosing all positive examples in the training set. Show that A is an ERM

Solution:

- Let $\begin{cases} \cdot\, S = ((x_i, y_i))_{i=1}^m \text{ be training set} \\ \cdot\, R(S) \text{ be the rectangle returned by learner } A \\ \cdot\, A(S): X \to Y \text{ be the corresponding hypothesis} \end{cases}$

We need to show that:

$$A(S) \in \underset{h \in \mathcal{H}^2_{rec}}{\operatorname{argmin}} \; L_S(h)$$

- Since A return all positive examples from training set:

$$A(S)(x_i) = y_i = 1, \quad \forall i \in [m] \qquad (1)$$

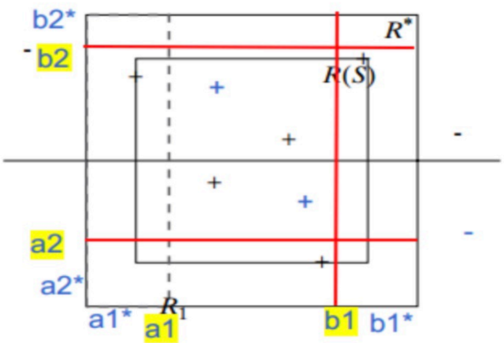
- Also, by realizability assumption, $\exists h^* \in \mathcal{H}^2_{rec}$ s.t. $L_S(h^*) = 0$

which means: $h^*(x_i) = y_i = 1, \quad \forall i \in [m] \qquad (2)$

- From (1) and (2), we can conclude that $L_S(A(S)) = 0$, and so A is an ERM.

(2) Show that if A receives a training set of size $\geq \frac{4\log(4/\delta)}{\varepsilon}$, then with probability of at least $(1-\delta)$, it returns a hypothesis with error of at most ε.

Hint:
- Fix some distribution D over X
- Let $R^* = R(a_1^*, b_1^*, a_2^*, b_2^*)$ be rectangle that generate labels, and f be corresponding hypothesis.
- Draw
$$\begin{cases} R_1 = R(a_1^*, a_1, a_2^*, b_2^*), \text{ where } a_1 \geq a_1^* \\ R_2 = R(b_1, b_1^*, a_2^*, b_2^*), \text{ where } b_1 \leq b_1^* \\ R_3 = R(a_1^*, b_1^*, a_2^*, a_2), \text{ where } a_2 \geq a_2^* \\ R_4 = R(a_1^*, b_1^*, b_2, b_2^*), \text{ where } b_2 \leq b_2^* \end{cases}$$
such that the probability mass are all $\varepsilon/4$



Solution:

Show that $R(S) \subseteq R^*$

- By definition of R(S).

$$R(S) \subseteq R^*$$

smallest rectangle enclosing all positive examples in training set

rectangle enclosing all positive examples in training set

Show that if S contains (positive) examples in all of the rectangles R_1, R_2, R_3, R_4, then the hypothesis returned by A has error of at most ε.

- Since $R(S) \subseteq R^*$:

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

Why?

. $L_{D,f}(A(S))$ has 2 components: . negative examples inside $R(S)$
. positive examples outside $R(S)$

. $R(S) \subseteq R^*$ eliminate the first component, hence the result.

$$= D(R^* \setminus R(S))$$

- Since $\begin{cases} \text{the probability mass for } R_1, R_2, R_3, R_4 \text{ are all } \varepsilon/4 \\ S \text{ contains all (positive) examples in } R_1, R_2, R_3, R_4 \end{cases}$

then: $D(R^* \setminus R(S)) \leqslant 4 \cdot \frac{\varepsilon}{4} = \varepsilon$

So: $L_{D,f}(A(S)) \leqslant \varepsilon$

For each $i \in \{1, \dots, 4\}$, upper bound the probability that S does not contain an example from R_i (meaning $R(S) = R^*$, or $h(x) = f(x)$)

- We would like to upper bound $D^m(\{S|_x : L_{D,f}(h_S) > \varepsilon\})$
- With the discussion above, if S contains (positive) examples in all of R_1, R_2, R_3, R_4 then $L_{D,f}(A(S)) \leqslant \varepsilon$. Therefore:

$$\{S|_x : L_{D,f}(h_S) > \varepsilon\} = \bigcup_{i=1}^{4} \{S|_x : S|_x \cap R_i = \emptyset\} \quad (1)$$

- It is easy to see that:

$$D^m(\{S|_x : S|_x \cap R_i = \emptyset\}) = \left(1 - \frac{\varepsilon}{4}\right)^m \leqslant e^{-\frac{\varepsilon}{4}m}$$

, for all $i \in \{1, 2, 3, 4\}$

Use the union bound to conclude the argument

- From discussion above:

$$D^m(\{S|_x : S|_x \cap R_i = \emptyset\}) \leq e^{-\frac{\varepsilon}{4}m} \quad , \forall i \in \{1,2,3,4\}$$

- Using union bound, we can find the upper bound:

$$\begin{aligned} D^m(\{S|_x : L_{D,f}(h_S) > \varepsilon\}) &= D^m\left(\bigcup_{i=1}^{4} \{S|_x : S|_x \cap R_i = \emptyset\}\right) \\ &\leq \sum_{i=1}^{4} D^m(\{S|_x : S|_x \cap R_i = \emptyset\}) \\ &= \sum_{i=1}^{4} e^{-\frac{\varepsilon}{4}m} \\ &= 4e^{-\frac{\varepsilon}{4}m} \end{aligned}$$

So the argument holds.