

# Seminar class 1

## Exercise 1 (exercise 2.1 in the book)

**Overfitting of polynomial matching:** We have shown that the predictor defined in Equation (2.3) leads to overfitting. While this predictor seems to be very unnatural, the goal of this exercise is to show that it can be described as a thresholded polynomial. That is, show that given a training set  $S = \{(\mathbf{x}_i, f(\mathbf{x}_i))\}_{i=1}^m \subseteq (\mathbb{R}^d \times \{0, 1\})^m$ , there exists a polynomial  $p_S$  such that  $h_S(\mathbf{x}) = 1$  if and only if  $p_S(\mathbf{x}) \geq 0$ , where  $h_S$  is as defined in Equation (2.3). It follows that learning the class of all thresholded polynomials using the ERM rule may lead to overfitting.

$$h_S(x) = \begin{cases} y_i & \text{if } \exists i \in [m] \text{ s.t. } x_i = x \\ 0 & \text{otherwise.} \end{cases} \quad (2.3)$$

## Exercise 2 (exercise 2.3 in the book)

**Axis aligned rectangles:** An axis aligned rectangle classifier in the plane is a classifier that assigns the value 1 to a point if and only if it is inside a certain rectangle.

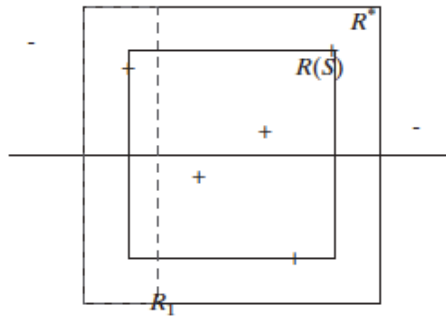


Figure 2.2. Axis aligned rectangles.

Formally, given real numbers  $a_1 \leq b_1, a_2 \leq b_2$ , define the classifier  $h_{(a_1, b_1, a_2, b_2)}$  by

$$h_{(a_1, b_1, a_2, b_2)}(x_1, x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq b_1 \text{ and } a_2 \leq x_2 \leq b_2 \\ 0 & \text{otherwise} \end{cases}. \quad (2.10)$$

The class of all axis aligned rectangles in the plane is defined as

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

Note that this is an infinite size hypothesis class. Throughout this exercise we rely on the realizability assumption.

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

*Hint:* Fix some distribution  $\mathcal{D}$  over  $\mathcal{X}$ , let  $R^* = R(a_1^*, b_1^*, a_2^*, b_2^*)$  be the rectangle that generates the labels, and let  $f$  be the corresponding hypothesis. Let  $a_1 \geq a_1^*$  be a number such that the probability mass (with respect to  $\mathcal{D}$ ) of the rectangle  $R_1 = R(a_1^*, a_1, a_2^*, b_2^*)$  is exactly  $\epsilon/4$ . Similarly, let  $b_1, a_2, b_2$  be numbers such that the probability masses of the rectangles  $R_2 = R(b_1, b_1^*, a_2^*, b_2^*)$ ,  $R_3 = R(a_1^*, b_1^*, a_2^*, a_2)$ ,  $R_4 = R(a_1^*, b_1^*, b_2, b_2^*)$  are all exactly  $\epsilon/4$ . Let  $R(S)$  be the rectangle returned by  $A$ . See illustration in Figure 2.2.

- Show that  $R(S) \subseteq R^*$ .
  - 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  $\epsilon$ .
  - For each  $i \in \{1, \dots, 4\}$ , upper bound the probability that  $S$  does not contain an example from  $R_i$ .
  - Use the union bound to conclude the argument.
3. Repeat the previous question for the class of axis aligned rectangles in  $\mathbb{R}^d$ .
  4. Show that the runtime of applying the algorithm  $A$  mentioned earlier is polynomial in  $d, 1/\epsilon$ , and in  $\log(1/\delta)$ .

### Exercise 3 (exercise 3.1 in the book)

**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. That is, show that given  $\delta \in (0, 1)$ , and given  $0 < \epsilon_1 \leq \epsilon_2 < 1$ , we have that  $m_{\mathcal{H}}(\epsilon_1, \delta) \geq m_{\mathcal{H}}(\epsilon_2, \delta)$ . Similarly, show that given  $\epsilon \in (0, 1)$ , and given  $0 < \delta_1 \leq \delta_2 < 1$ , we have that  $m_{\mathcal{H}}(\epsilon, \delta_1) \geq m_{\mathcal{H}}(\epsilon, \delta_2)$ .

### Exercise 4 (exercise 3.2 in the book)

Let  $\mathcal{X}$  be a discrete domain, and let  $\mathcal{H}_{\text{Singleton}} = \{h_z : z \in \mathcal{X}\} \cup \{h^-\}$ , where for each  $z \in \mathcal{X}$ ,  $h_z$  is the function defined by  $h_z(x) = 1$  if  $x = z$  and  $h_z(x) = 0$  if  $x \neq z$ .  $h^-$  is simply the all-negative hypothesis, namely,  $\forall x \in \mathcal{X}, h^-(x) = 0$ . The realizability assumption here implies that the true hypothesis  $f$  labels negatively all examples in the domain, perhaps except one.

1. Describe an algorithm that implements the ERM rule for learning  $\mathcal{H}_{\text{Singleton}}$  in the realizable setup.
2. Show that  $\mathcal{H}_{\text{Singleton}}$  is PAC learnable. Provide an upper bound on the sample complexity.

### Exercise 5 (exercise 3.7 in the book)

(\*) **The Bayes optimal predictor:** Show that for every probability distribution  $\mathcal{D}$ , the Bayes optimal predictor  $f_{\mathcal{D}}$  is optimal, in the sense that for every classifier  $g$  from  $\mathcal{X}$  to  $\{0, 1\}$ ,  $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$ .