

## ASSIGNMENT 1

**Exercise 1** (Best predictor when distribution is known). Suppose  $(X, Y) \sim P_{X,Y}$  take finitely many values. A statistician who observes  $X$  and knows  $P_{X,Y}$  is asked to find a prediction rule  $h(X) \in \{0, 1\}$  that minimizes the error probability  $Pr(h(X) \neq Y)$ . Show that the best predictor is  $h^*(x) = \arg \max_y P(y|x)$ .

**Exercise 2.** Let  $\mathcal{H}$  be a class of binary classifiers over a domain  $\mathcal{X}$ . Let  $P$  be an unknown distribution over  $\mathcal{X}$ , and let  $f$  be true hypothesis in  $\mathcal{H}$ . Fix some  $h \in \mathcal{H}$ . Show that the expected value of the empirical loss  $L_S(h)$  equals  $L_{(D,f)}(h)$ , namely,

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

**Exercise 3** (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. 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

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

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:* Let  $R^*$  be the rectangle that generates the labels, and let  $f$  be the corresponding hypothesis. Let  $R(S^m)$  be the rectangle returned by  $A$ . See illustration in Figure 1.

- Show that  $R(S^m) \subseteq R^*$ .

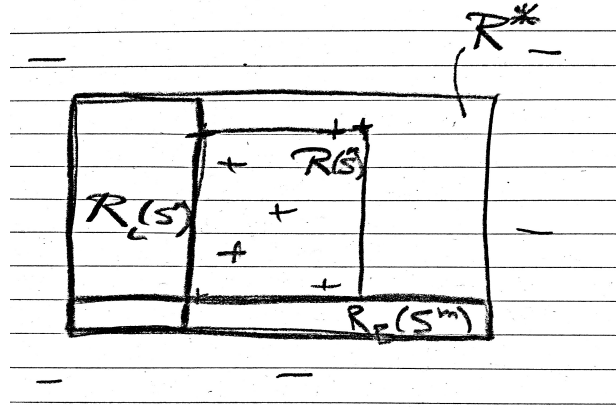


Figure 1: The outer rectangle  $R^*$  corresponds to  $f$ . The rectangle in the middle corresponds to  $R(S^m)$ .  $R_L$  and  $R_B$  correspond to the left and right stripes.  $R_R$  and  $R_T$  are not represented. The difference  $R^* \setminus R(S^m)$  is included in the union of the four stripes.

- Consider the 4 stripes that surround  $R(S^m)$  as shown on Fig. 1—some of those stripes might be the emptyset. Let us denote them by  $R_L(S^m)$ ,  $R_T(S^m)$ ,  $R_R(S^m)$ ,  $R_B(S^m)$  (the left, top, right, and bottom stripes). Show that if the probability under  $P$  of each of these stripes is at most  $\varepsilon/4$ , then the hypothesis returned by  $A(S^m)$  has error of at most  $\varepsilon$ , that is  $L_{P,f}(A(S^m)) \leq \varepsilon$ . Therefore, if  $L_{P,f}(A(S^m)) > \varepsilon$  then  $P(R_i(S^m)) > \varepsilon/4$  for at least some  $i \in \{L, T, R, B\}$ . Define  $I(S^m)$  as the set of stripe indices  $i$  such that  $P(R_i(S^m)) > \varepsilon/4$ . Show that  $P^m(i \in I(S^m)) \leq (1 - \varepsilon/4)^m$ . Conclude.
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  $d$ ,  $1/\varepsilon$ , and in  $\log(1/\delta)$ .