Telecom Paris — Introduction to Machine Learning (APM-0EL05-TP)
Teacher: A. Tchamkerten — Course 1

ASSIGNMENT 1 - SOLUTIONS

Exercise 1 (Best predictor when distribution is known)**.** Suppose $(X, Y) \sim P_{X,Y}$ take finitely many values. A statistician is 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 $h^*(x) = \arg\max_y P(y|x)$.

Solution. We have

$$\begin{aligned} Pr(h(X) \neq Y) &= \sum_x Pr(Y \neq h(X)|X = x)Pr(X = x) \\ &= \sum_x (1 - Pr(Y = h(x)|X = x))Pr(X = x) \\ &\geq \sum_x (1 - Pr(Y = h^*(x)|X = x))Pr(X = x) \qquad (1) \end{aligned}$$

where the inequality follows from the definition of $h^*(x)$. □

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 $L_{(P,f)}(h)$, namely,

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

Solution. By the linearity of expectation,

$$\begin{aligned} \mathop{\mathbb{E}}_{S \sim P^m} [L_S(h)] &= \mathop{\mathbb{E}}_{S \sim P^m} \left[\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{h(X_i) \neq f(X_i)\} \right] \\ &= \frac{1}{m} \sum_{i=1}^{m} \mathop{\mathbb{E}}_{X_i \sim P} [\mathbb{1}\{h(X_i) \neq f(X_i)\}] \\ &= \frac{1}{m} \sum_{i=1}^{m} \mathop{\mathbb{P}}_{X_i \sim P} [h(X_i) \neq f(X_i)] \\ &= \frac{1}{m} m \, L_{(P,f)}(h) \\ &= L_{(P,f)}(h). \end{aligned}$$

□

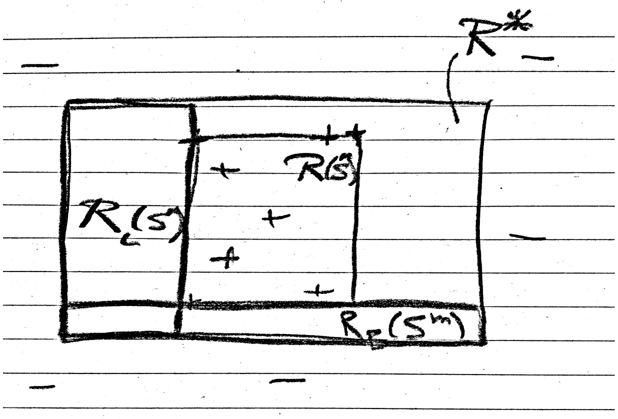

Figure 1: The outside rectangle R^* corresponds to f. The rectangle in the middle corresponds to $R(S^m)$. R_L and R_B correspond to the left and right stripes. R_R and R_T are not represented. The difference $R^* \backslash R(S^m)$ is included in the union of the four stripes.

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

$$h_{(a_1,b_1,a_2,b_2)}(x_1, x_2) = \begin{cases} 1 & \text{if } a_1 \leq x_1 \leq b_1 \text{ and } a_2 \leq x_2 \leq b_2 \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

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

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

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 ϵ.

   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 $R(S^m) \subseteq R^*$.
   - 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 ϵ, 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. 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, $1/\epsilon$, and in $\log(1/\delta)$.

Solution.

1. Observe that by definition A achieves zero on all instances in the training set. Since the loss function is nonnegative, we deduce that A is an ERM.

2. Fix some distribution P over $\mathcal{X}$, and define R^* as in the hint. Let f be the hypothesis associated with R^*. We have

$$L_{(P,f)}(A(s^m)) = P(R^* \setminus R(s^m)) = P(\cup_{i\in\{L,T,R,B\}} R_i(s^m)).$$

Therefore, if s^m induces a "large error" under distribution P, i.e., is such that

$$L_{(P,f)}(A(s^m)) > \varepsilon,$$

it necessarily satisfies

$$P(R_i(s^m)) > \varepsilon/4 \tag{3}$$

for some $i \in \{L,T,R,B\}$. So let us assume that s^m satisfy (3) for some $i \in \{L,T,R,B\}$ —for otherwise there is nothing to prove. Denote by $I(s^m)$ the set of indices i in $\{L,T,R,B\}$ such that $P(R_i(s^m)) > \varepsilon/4$. Observe that if $i \in I(s^m)$ then necessarily the m data points of s^m all belong to a region whose probability is at most $(1-\varepsilon/4)^m$, that is

$$P^m(i \in I(s^m)) \leq (1-\varepsilon/4)^m.$$

Therefore,

$$\begin{aligned}
P^m(L_{(P,f)}(A(S^m)) > \varepsilon) &\leq P^m(I(S^m) \neq \emptyset) \\
&= P^m\left(\bigcup_{i\in\{L,T,R,B\}} \{i \in I(S^m)\}\right) \\
&\leq \sum_{i\in\{L,T,R,B\}} P^m\left(i \in I(S^m)\right) \\
&\leq \sum_{i\in\{L,T,R,B\}} (1-\varepsilon/4)^m \\
&= 4(1-\varepsilon/4)^m \\
&\leq 4e^{-m\varepsilon/4}.
\end{aligned}$$

We deduce that if

$$m > (4/\varepsilon)\ln(4/\delta)$$

then with probability $\geq 1-\delta$ the error will be $\leq \varepsilon$, irrespectively of P.

3. The hypothesis class of axis aligned rectangles in $\mathbb{R}^d$ is defined as follows. Given real numbers $a_1 \leq b_1$, $a_2 \leq b_2$, ...,$a_d \leq b_d$, define the classifier $h_{(a_1,b_1,\ldots,a_d,b_d)}$ by

$$h_{(a_1,b_1,\ldots,a_d,b_d)}(x_1,\ldots,x_d) = \begin{cases} 1 & \text{if } \forall i \in [d], a_i \leq x_i \leq b_i \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

The class of all axis-aligned rectangles in $\mathbb{R}^d$ is defined as

$$\mathcal{H}^d_{\text{rec}} = \{h_{(a_1,b_1,\ldots,a_d,b_d)} : \forall i \in [d], a_i \leq b_i\}.$$

It can be seen that the same algorithm proposed above is an ERM for this case as well. The sample complexity is analyzed similarly. The only difference is that instead of 4 strips, we have $2d$ strips (2 strips for each dimension). Thus, it suffices to draw a training set of size $\left\lceil \frac{2d\log(2d/\delta)}{\epsilon} \right\rceil$.

4. For each dimension, the algorithm has to find the minimal and the maximal values among the positive instances in the training sequence. Therefore, its runtime is $\mathcal{O}(md)$. Since we have shown that the required value of m is at most $\left\lceil \frac{2d\log(2d/\delta)}{\epsilon} \right\rceil$, it follows that the runtime of the algorithm is indeed polynomial in d, $1/\epsilon$, and $\log(1/\delta)$.

□