Theory Homework 1

CSE 493S/599S: Advanced Machine Learning
Instructor: Sewoong Oh
Due: Thursday, May 14$^{\text{th}}$ at 11:59pm

The goal of this homework is to help you better understand the ideas from theoretical machine learning we have covered in class.

Notes:

- You will be assigned a subset of the problems for each homework. Please submit the homework to Gradescope and link each page of your work to the corresponding problem.
- Please typeset your work using LaTeX.
- List every person with whom you discussed any problem in any depth, and every reference (outside of our course slides, lectures, and textbook) that you used.
- You may spend an arbitrary amount of time discussing and working out a solution with your listed collaborators, but **do not take notes, photos, or other artifacts of your collaboration**. Erase the board you were working on, and once you're alone, write up your answers yourself.
- The homework problems have been carefully chosen for their pedagogical value and hence might be similar or identical to those given out in similar courses at UW or other schools. Using any pre-existing solutions from these sources, from the Web or other textbooks constitutes a violation of the academic integrity expected of you and is strictly prohibited.

Version history:

V1 Initial version.

1 ERM and axis aligned rectangles (taught in the 1st theory lecture)

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

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

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 sized hypothesis class. Throughout this exercise we rely on the realizability assumption.

(1) Let A be the algorithm that returns the smallest rectangle $\hat{R}(S)$ enclosing all positive examples in the training set S. Show that $\hat{R}(S)$ minimizes the empirical risk. **[5 points]**

(2) Show that if A receives a training set S 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. **[10 points]**

Hint: Fix some distribution D over $\mathcal{X}$, let

$$R^* = R(a_1^*, b_1^*, a_2^*, b_2^*) \triangleq \{(x,y) \mid a_1^* \leq x_1 \leq b_1^*, a_2^* \leq x_2 \leq b_2^*\}$$

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 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 $\hat{R}(S)$ be the rectangle returned by A. See illustration in Fig. 1.

(a) Show that $\hat{R}(S) \subseteq R^*$.

(b) 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.

(c) For each $i \in \{1, \ldots, 4\}$, upper bound the probability that S does not contain an example from R_i.

(d) Use the union bound to conclude the argument.

(3) Repeat the previous question for the class of axis aligned rectangles in $\mathbb{R}^d$. **[10 points]**

(4) Show that the runtime of applying the algorithm A mentioned in part (3) is polynomial in d, $1/\epsilon$, and in $\log(1/\delta)$. **[5 points]**

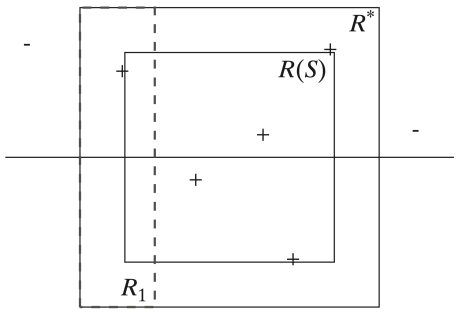

Figure 1: Axis aligned rectangles

[30 points]

---

2 The Bayes optimal predictor **(taught in the 2nd theory lecture)**

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 $\{0,1\}$, $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$.

Hint: For $x \in \mathcal{X}$, let α_x denote the conditional probability of a positive label given x. Show that $\mathbb{P}[f_{\mathcal{D}}(X) \neq y \,|\, X = x] = \min\{\alpha_x, 1-\alpha_x\}$ and that for any classifier $g : \mathcal{X} \to \{0,1\}$, we have $\mathbb{P}[g(X) \neq y \,|\, X = x] \geq \min\{\alpha_x, 1-\alpha_x\}$. Finally, conclude that $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$.

[20 points]
