Homework #1

CSE 493S/599S: Advanced Machine Learning
Prof. Ludwig Schmidt
Due: Tuesday, May 16$^{\text{th}}$ 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:

- **This homework contains only the first 4 problems. We will add a 5$^{\text{th}}$ problem by Sunday April 9$^{\text{th}}$.**
- Please submit this homework to Gradescope and link each page of your work to the corresponding problem.
- 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 past offerings of this course at UW, or similar courses at other schools. Using any pre-existing solutions from these sources, from the Web or other textbooks constitues a violation of the academic integrity expected of you and is strictly prohibited.

1 ERM and 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

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

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 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: Fix some distribution 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 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 Fig. 1.

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

(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 $\mathbb{R}^d$.

(4) Show that the runtime of applying the algorithm A mentioned earlier is polynomial in d, $1/\epsilon$, and in $\log(1/\delta)$.

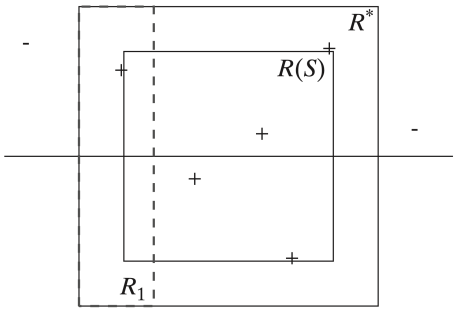

Figure 1: Axis aligned rectangles

[30 points]

---

2 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 $\{0,1\}$, $L_{\mathcal{D}}(f_{\mathcal{D}}) \leq L_{\mathcal{D}}(g)$.

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]

---

3 VC-dimension of axis aligned rectangles

Let $\mathcal{H}^d_{\text{rec}}$ be the class of axis aligned rectangles in $\mathbb{R}^d$. Prove that $\text{VCdim}(\mathcal{H}^d_{\text{rec}}) = 2d$.

[20 points]

---

4 Infinite VC-dimension with one parameter

It is often the case that the VC-dimension of a hypothesis class equals (or can be bounded above by) the number of parameters one needs to set in order to define each hypothesis in the class. For instance, if $\mathcal{H}$ is the class of axis aligned rectangles in $\mathbb{R}^d$, then $\text{VCdim}(\mathcal{H}) = 2d$, which is equal to the number of parameters used to define a rectangle in $\mathbb{R}^d$. Here is an example that shows that this is not always the case. We will see that a hypothesis class might be very complex and even not learnable, although it has a small number of parameters.

Consider the domain $\mathcal{X} = \mathbb{R}$, and the hypothesis class

$$\mathcal{H} = \{x \mapsto \lceil \sin(\theta x) \rceil : \theta \in \mathbb{R}\}$$

(here, we take $\lceil -1 \rceil = 0$). Prove that $\text{VCdim}(\mathcal{H}) = \infty$.

Hint: There is more than one way to prove the required result. One option is by applying the following lemma: If $0.x_1x_2x_3\ldots$, is the binary expansion of $x \in (0, 1)$, then for any natural number m, $\lceil \sin(2^m \pi x) \rceil = (1 - x_m)$, provided that $\exists k \geq m$ s.t. $x_k = 1$.

[30 points]
