CSCI699: Theory of Machine Learning Fall 2023

Problem Set 1

Due: September 20 by 11:59 pm

Discussion is allowed and encouraged but everyone should write solutions on their own. Please also mention any collaborators you had substantial discussions with. You are also allowed to consult general resources on the internet (such as one of the books, or other lecture notes online), but you should not search for any the solutions themselves online.

If you use the LaTeX template, then please only keep your answers and remove the questions before submitting.

Homeworks should be written in Latex and submitted via Gradescope. **When you submit on Gradescope, make sure to mark the page which contains each answer.**

Problem 1: PAC Learnability for Axis-Aligned Rectangles

Most of this question appears as Exercise 3 in Chapter 2 of the Understanding Machine Learning book. You are free to look at additional hints given there, if you like.

This question is about showing a PAC learning guarantee for rectangles, and will hopefully help understand different facets of PAC learning. As we discussed in lecture 1 of class, an axis aligned rectangle classifier is a classifier than 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 $h_{(a_1,b_1,a_2,b_2)}$ on an input with coordinates (x_1, x_2) by

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

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

$$\mathcal{H}^2_{\text{rec}} = \{h_{(a_1,b_1,a_2,b_2)}(x_1, x_2) : a_1 \leq b_1, a_2 \leq b_2\}.$$

Throughout this exercise we rely on the realizability assumption.

(a) (3pts) Let A be the algorithm that returns the smallest rectangle enclosing all positive examples in the training set. Show that A is an ERM.

(b) (5pts) Show that if the algorithm A from part (1) receives a training set of size $\geq \dfrac{4\log(4/\delta)}{\epsilon}$ then with probability at least $1-\delta$ it returns a hypothesis with error at most ϵ.

(c) (5pts) Repeat the previous question for the class of axis-aligned rectangles in $\mathbb{R}^d$.

(d) (5pts) Show that the runtime of the algorithm A is polynomial in $d, 1/\epsilon$ and $\log(1/\delta)$. Therefore, we have shown that the class of axis-aligned rectangles in $\mathbb{R}^2$ is *efficiently* PAC-learnable. Though we are not talking too much about efficient runtimes at this point of the class, we will spend more time discussing it soon.

Problem 2: Uniform Convergence for Parameterized Hypothesis Classes

In class, we saw a uniform convergence result for finite hypothesis classes (which then directly implied agnostic PAC learnability for finite hypothesis classes). In this question, we'll see how we can use the result for finite hypothesis classes to also get uniform convergence for certain infinite classes, using the powerful idea of an *ϵ-net*.

Let $\mathcal{H}$ be a hypothesis class consisting of a family of functions indexed by some parameter θ which lies in the set S, i.e. $\mathcal{H} = \{h_\theta : \theta \in S\}$. We assume for this question that the set S is the ball of radius $B > 0$ in $\mathbb{R}^d$:

$$S = \{\theta \in \mathbb{R}^d : \|\theta\|_2 \leq B\}. \tag{1}$$

Assume that the risk $R(h_\theta)$ of any hypothesis h_θ is a *L-Lipschitz* function of θ, defined as follows.

Definition 1. *Let $L \geq 0$. The function $R(h_\theta)$ is a L-Lipschitz function of θ if for all $\theta, \theta' \in S$ we have,*

$$|R(h_\theta) - R(h_{\theta'})| \leq L \left\|\theta - \theta'\right\|_2 .$$

Similarly, the empirical risk $\hat{R}(h_\theta)$ is a L-Lipschitz function of θ if for all $\theta, \theta' \in S$ we have,

$$|\hat{R}(h_\theta) - \hat{R}(h_{\theta'})| \leq L \left\|\theta - \theta'\right\|_2 .$$

(a) (2pts) Briefly discuss when this L-Lipschitz condition could be satisfied (does not need to be rigorous, you can be high-level). You might find it useful to think about losses other than the 0/1 loss we've usually considered in class.

(b) (5pts) Since S is infinite, $\mathcal{H}$ is an infinite hypothesis class. We now define the notion of an ϵ-net of S which allows us to consider only a finite subset of S to show uniform convergence for $\mathcal{H}$.

Definition 2. *Let $\epsilon > 0$. An ϵ-net of S (with respect to the $\|\cdot\|_2$ norm) is a subset $C \subseteq S$ such that $\forall\, \theta \in S, \exists\, \theta' \in C$ such that $\left\|\theta - \theta'\right\|_2 \leq \epsilon$.*

Show that the set C defined as follows is an ϵ-net of S,

$$C = \left\{\theta \in S : \theta_i = \frac{j\epsilon}{\sqrt{d}}, j \in \mathbb{Z}, |j| \leq \frac{B\sqrt{d}}{\epsilon}\right\}.$$

Here θ_i denotes the i-th coordinate of the vector θ, $\mathbb{Z}$ is the set of all integers. (Hint: it might be helpful to draw a picture. We're essentially trying to 'cover' the entire ball S using this grid of points C.)

(c) (3pts) Assume $\epsilon < 3B\sqrt{d}$. Show that $|C| \leq \left(\frac{3B\sqrt{d}}{\epsilon}\right)^d$.

(d) (2pts) Assume that the loss function $\ell(h_\theta(x), y) \in [-M, M]$ for every datapoint (x, y). Using Hoeffding's inequality, show that with probability $\geq 1 - 2|C| \exp\left(-\frac{n\alpha^2}{2M^2}\right)$,

$$|\hat{R}(h_\theta) - R(h_\theta)| \leq \alpha, \ \forall\, \theta \in C.$$

We refer to the above event as E for the rest of the problem.

(e) (5pts) Using the L-Lipschitz property of $R(h_\theta)$ and $\hat{R}(h_\theta)$, show that conditioned on the event E,

$$|\hat{R}(h_\theta) - R(h_\theta)| \leq 2L\epsilon + \alpha, \forall\, \theta \in S. \tag{2}$$

(Hint: by the property of the ϵ-net, for any $\theta \in S$, there is some $\theta' \in C$ such that $\left\|\theta - \theta'\right\|_2 \leq \epsilon$. In part (d) we have shown that the empirical and true risks are close for all $\theta' \in C$. Try to decompose $\hat{R}(h_\theta) - R(h_\theta)$ in a suitable way such that you can then use part (d), the L-Lipschitz property of the functions $R(h_\theta)$ and $\hat{R}(h_\theta)$, and the triangle inequality to bound $|\hat{R}(h_\theta) - R(h_\theta)|$.)

(f) (3pts) The final step is a bit algebraic, so we take all constants $L = B = M = 1$. Show that if we set

$$\alpha = 10\sqrt{\frac{d\log(n)}{n}}, \text{ and } \epsilon = \frac{\alpha}{2},$$

then conditioned on the event E,

$$|\hat{R}(h_\theta) - R(h_\theta)| \leq 20\sqrt{\frac{d\log(n)}{n}}, \forall\, \theta \in S.$$

Using this, bound the failure probability: show that $\Pr(E) \geq 1 - e^{-d}$. This implies that with failure probability $\mathcal{O}(e^{-d})$,

$$|\hat{R}(h_\theta) - R(h_\theta)| \leq 20\sqrt{\frac{d\log(n)}{n}}, \forall\, \theta \in S.$$

Let's look back at what we have shown: our result implies that for some constant $c' > 0$ and $n \geq \frac{c' d\log(d/\gamma)}{\gamma^2}$, with failure probability $\mathcal{O}(e^{-d})$,

$$|\hat{R}(h_\theta) - R(h_\theta)| \leq \gamma, \forall\, \theta \in S.$$

Using the reduction from agnostic PAC learning to uniform convergence, we have therefore shown that the sample complexity of agnostically learning $\mathcal{H}$ to error γ with failure probability $\mathcal{O}(e^{-d})$ is at most $\mathcal{O}\left(\frac{d\log(d/\gamma)}{\gamma^2}\right)$. This technique of using an ϵ-net to show some property of every member of an infinite set comes in handy in many places.

Problem 3: VC dimension

In this problem we will explore VC dimension bounds for various hypothesis classes.

(a) Let $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \{-1, +1\}$ and $\mathcal{H} = \{f_{\theta,b}(x) = \text{sign}\left(\langle x, \theta\rangle + b\right) \ : \ \theta \in \mathbb{R}^d, b \in \mathbb{R}\}$ be the set of d-dimensional linear classifiers. Prove VCdim$(\mathcal{H}) = d + 1$ following the two steps below.

(i) (3pts) Construct $d+1$ points $x_1, \ldots, x_{d+1} \in \mathbb{R}^d$ and argue that for any labeling $y_1, \ldots, y_{d+1} \in \{-1, +1\}$, there exists $h \in \mathcal{H}$ such that $h(x_t) = y_t$ for all $t = 1, \ldots, d + 1$.

(ii) (4pts) Prove that for any $d+2$ points $x_1, \ldots, x_{d+2} \in \mathbb{R}^d$, there exists a labeling $y_1, \ldots, y_{d+2} \in \{-1, +1\}$ such that no $h \in \mathcal{H}$ satisfies $h(x_t) = y_t$ for all $t = 1, \ldots, d + 2$. (Hint: use the fact that $m + 1$ points in an m-dimensional space must be linearly dependent.)

(b) We will now get a VC dimension bound for neural networks. For $k = \{1, \ldots, M\}$, let r_k be some positive integer and $\mathcal{H}_k : \{-1, +1\}^{r_k} \to \{-1, +1\}$ be some function class with growth function $\Pi_{\mathcal{H}_k}$ and VC-dimension d_k. Further define a vector-valued function class mapping from $\{-1, +1\}^{r_k}$ to $\{-1, +1\}^{r_{k+1}}$ as

$$\mathcal{F}_k = \left\{h(x) = (f_1(x), \ldots, f_{r_{k+1}}(x)) \ : \ f_1, \ldots, f_{r_{k+1}} \in \mathcal{H}_k\right\}.$$

Define $r_{M+1} = 1$. Then the following class represents an M-layer feedforward neural net

$$\mathcal{H} = \{h_M \circ \cdots \circ h_1 : \{-1, +1\}^{r_1} \to \{-1, +1\} \ : \ h_1 \in \mathcal{F}_1, \ldots, h_M \in \mathcal{F}_M\}$$

where $\circ$ represents function composition (try to draw a picture to help understand the notation).

(i) (4pts) Prove that the growth function of $\mathcal{H}$ is bounded as

$$\Pi_{\mathcal{H}}(n) \le \prod_{k=1}^{M} \left(\Pi_{\mathcal{H}_k}(n)\right)^{r_{k+1}}.$$

(ii) (2pts) Let $d = \sum_{k=1}^{M} r_{k+1} d_k$. Prove $\Pi_{\mathcal{H}}(n) \le (en)^d$ for $n > d + 1$.

(iii) (3pts) Further show VCdim$(\mathcal{H}) = \mathcal{O}(d \ln d)$. (Hint: take logarithms, and you might find the inequality $1 + \ln x \le 2\sqrt{x}$ useful.)

This shows a VC-dimension bound which holds for deep neural networks, and is essentially the same as the number of parameters of the network. This bound can be loose for modern neural networks where the number of parameters d can run in billions, but we can still can good test accuracy with much fewer samples. Understanding the relationship between the number of parameters of neural networks and how much data we need to learn has been a topic of much study recently [1, 2].

(c) (4pts) For the examples of hypothesis classes we've seen so far, the number of parameters gave a good estimate of the VC dimension of the hypothesis class. This is often, but not always the case. Let $\mathcal{X} = \mathbb{R}, \mathcal{Y} = \{-1, +1\}$ and $\mathcal{H} = \{f_\theta(x) = \text{sign}(\sin(\theta x)) \; : \; \theta \in \mathbb{R}\}$. Prove that for any n, $\mathcal{H}$ shatters the set $\{x_1, \ldots, x_n \; : \; x_i = 2^{-2i}, i \in [n]\}$. This implies that $\text{VCdim}(\mathcal{H}) = \infty$. Therefore there is a hypothesis classes with just a single parameter which is not PAC-learnable. (Hint: for any labeling $\{y_1, \ldots, y_n\}$, consider $\theta = \pi(1 + \sum_{i=1}^{n} (1 - y_i) 2^{2i-1})$.)

References

[1] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[2] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.