CSE 250: Machine Learning Theory Spring 2017

Homework 1

Instructor: Raef Bassily **Due on:** Tue May 2

Instructions and Notes

- For your proofs, you may use any result covered in class, and its analysis, but please cite the result that you use.
- All problems have equal weights.
- The assignment will be graded on clarity and correctness. If your arguments are not clear and have holes in them, it will be assumed inncorrect.
- **Notation:** For any positive integer k, I will use the notation $[k]$ to denote $\{1, \ldots, k\}$. I will use $\langle \mathbf{x}, \mathbf{y} \rangle$ to denote the inner-product between two vectors $\mathbf{x}, \mathbf{y}$. Also, $\|\mathbf{x}\|_p$ will denote the L_p norm of a vector $\mathbf{x}$.
- **Note:** Whenever PAC-learnability is addressed, **realizability condition** will be assumed **unless stated otherwise.**

Problem 1: Learning Concentric Circles

Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0, 1\}$, and let $\mathcal{H}_{\mathsf{circles}}$ be the class of concentric circles in the plane, that is, $\mathcal{H}_{\mathsf{circles}} = \{h_r : r \in \mathbb{R}_+\}$, where, for any $x \in \mathbb{R}^2$, $h_r(x) \triangleq \mathbf{1}\left(\|x\|_2 \leq r\right)$. Prove that $\mathcal{H}_{\mathsf{circles}}$ is PAC-learnable (assume realizability), and its sample complexity is upper-bounded by

$$n_{\mathcal{H}_{\mathsf{circles}}}(\epsilon, \delta) = \left\lceil \frac{\ln(1/\delta)}{\epsilon} \right\rceil.$$

Problem 2: Boosting the Confidence of a Learner

Let $\mathcal{H}$ be a hypothesis class. Assuming realizability, suppose you are given an algorithm $\mathcal{A}$ that, given a parameter $0 < \epsilon < 1$ and $n(\epsilon)$ i.i.d samples from the underlying distribution D, produces a hypothesis $h \in \mathcal{H}$ such that $\mathbf{err}(h; D) \leq \epsilon$ with probability at least $\frac{1}{10}$.

Given $\mathcal{A}$ and a procedure **EX** which when called outputs a fresh independent labeled example from the distribution D,

1. Describe a PAC-learning algorithm for $\mathcal{H}$, i.e., an algorithm that can learn $\mathcal{H}$ with confidence at least $1 - \delta$ for arbitrarily small δ. (A clear description of your algorithm would suffice for this part.)
2. Prove that your algorithm PAC-learns $\mathcal{H}$ (i.e., derive $n_{\mathcal{H}}(\epsilon, \delta)$).

Problem 3: Learning Point Functions

Let $\mathcal{X} = [d]$ for some fixed $d \in \mathbb{N}$, let $\mathcal{Y} = \{0, 1\}$, and let $\mathcal{H}_{\mathsf{points}}$ be the class of point functions, that is, $\mathcal{H}_{\mathsf{points}} = \{h_z : \ z \in \mathcal{X}\}$ where, for any $x, z \in \mathcal{X}$, $h_z(x) \triangleq \mathbf{1}(x = z)$. Assuming realizability,

1. Describe an algorithm that PAC-learns $\mathcal{H}_{\mathsf{points}}$.
2. Prove that $\mathcal{H}_{\mathsf{points}}$ is PAC-learnable showing that the sample complexity is upper-bounded by

$$n_{\mathcal{H}_{\mathsf{points}}} = \left\lceil \frac{2\ln\left(\frac{2}{\epsilon\delta}\right)}{\epsilon} \right\rceil.$$

3. Hence, show that the general bound for finite Hypothesis classes dicussed in class (Occam's razor) is not tight in general.

Problem 4: Axis-Aligned Rectangles

The goal of this problem is to show that there is no loss in generality in the argument discussed in class for the proof of learnability of axis-aligned rectangles. In particular, this problem shows how the argument can be extended to a corner case where distribution on the domain points has a discrete component concentrated at the boundary of the true labeling rectangle. In fact, if we assume that the distribution on the domain points has this form, then we can show that less number of examples are required to achieve the desired accuracy/confidence.
Let $\mathcal{X} = \mathbb{R}^2$, $\mathcal{Y} = \{0, 1\}$, and let $\mathcal{H}_{\mathsf{rec}}$ be the class of axis-aligned rectangles, that is,

$$\mathcal{H}_{\mathsf{rec}} = \{h_{(a_1,b_1,a_2,b_2)} : \ a_1 \leq b_1, a_2 \leq b_2\},$$

where, for any $\mathbf{x} = (x_1, x_2) \in \mathbb{R}^2$,

$$h_{(a_1,b_1,a_2,b_2)}(\mathbf{x}) = \mathbf{1}\left(a_1 \leq x_1 \leq b_1 \text{ and } a_2 \leq x_2 \leq b_2\right).$$

Assuming realizability, let $\mathcal{A}$ denote the algorithm discussed in class that, given a sample of i.i.d. examples from the underlying distribution, outputs the "tightest" rectangle enclosing the 1-labeled examples. Let $h_{(a_1^*,b_1^*,a_2^*,b_2^*)}$ denote the true labeling rectangle. Let $\mathsf{B}(a_1^*, b_1^*, a_2^*, b_2^*)$ denote the boundary of the rectangle $h_{(a_1^*,b_1^*,a_2^*,b_2^*)}$, that is,

$$\mathsf{B}(a_1^*, b_1^*, a_2^*, b_2^*) = \left\{(x_1, x_2) \in \mathbb{R}^2 : \ \left\{x_1 \in \{a_1^*, b_1^*\} \textbf{ and } x_2 \in [a_2^*, b_2^*]\right\} \textbf{ or } \left\{x_1 \in [a_1^*, b_1^*] \textbf{ and } x_2 \in \{a_2^*, b_2^*\}\right\}\right\}$$

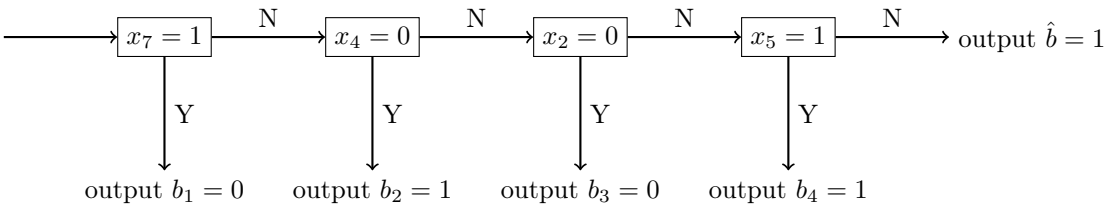
Suppose that the distribution $\mathcal{D}$ over $\mathcal{X}$ is such that $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x} \in \mathsf{B}(a_1^*, b_1^*, a_2^*, b_2^*)] = 1/3$.
For all $0 < \epsilon \leq 1/3$ and all $0 < \delta < 1$, show that the number of examples from a distribution of the above form required to ensure an error not exceeding ϵ with probability at least $1 - \delta$ is **strictly less** than the worst-case quantity $n_{\mathcal{H}_{\mathsf{rec}}}(\epsilon, \delta)$ that we derived in class.

Problem 5: Learning Decision Lists

Suppose $\mathcal{X} = \{0, 1\}^d$, $\mathcal{Y} = \{0, 1\}$, and $\mathcal{H}_{\mathsf{List}}$ is the class of *uni-literal* decision lists.
A decision list $h \in \mathcal{H}_{\mathsf{List}}$ is an ordered sequence of ℓ if-then-else statements for some $\ell \in \mathbb{N}$. That is, it is a sequence $(C_1, b_1), \ldots (C_\ell, b_\ell)$, together with a default bit $\hat{b}$, where each C_i is a boolean formula and b_i is the output bit when C_i evaluates to true. For example:

$$\begin{aligned}
&\text{if } C_1 \text{ then output } b_1 \in \{0, 1\}\\
&\text{else if } C_2 \text{ then output } b_2 \in \{0, 1\}\\
&\qquad\qquad\vdots\\
&\text{else if } C_\ell \text{ then output } b_\ell \in \{0, 1\}\\
&\qquad\text{else output } \hat{b} \in \{0, 1\} \text{ (the default bit)}
\end{aligned}$$

In 1-decision lists (or uni-literal decision lists), each C_i is a conjunction of just **one** literal from the set of boolean literals $\{x_1, \ldots, x_d, \overline{x}_1, \ldots, \overline{x}_d\}$. Here is an illustrated example of a uni-literal decision list of length 4:



Given an unlabeled example $\mathbf{a} = (a_1, \ldots, a_d) \in \{0,1\}^d$ and a decision list h, we assign the variable $x_j = a_j$ for $j \in [d]$ and then run these assignments through the if-then-else statements. Then $h(\mathbf{a})$ is simply the output bit of this procedure.

1. Is $\mathcal{H}_{\mathsf{List}}$ PAC-learnable? Find an upper bound $n_{\mathcal{H}_{\mathsf{List}}}(\epsilon, \delta)$ on its sample complexity.
2. Describe an algorithm that PAC-learns $\mathcal{H}_{\mathsf{List}}$.

Hints

- **Problem 2:**
  1. You may need to run $\mathcal{A}$ multiple times on separate data sets (whose samples are drawn using multiple calls to **EX**). As a result, you will possibly get multiple hypotheses.
  2. Now, you need to find a way to generate a final hypothesis from these hypotheses such that it is accurate with high confidence (arbitrarily small δ rather than just $1/2$). You also need to make sure that whatever hypothesis you output at the end belongs to $\mathcal{H}$ (this is how PAC learnability was defined in class).
- **Problem 3:** Let $\mathcal{D}$ denote the distribution over $[d]$. Let's call a point $j \in [d]$ *critical* if $\mathbb{P}_{x \sim \mathcal{D}}[x = j] > \epsilon/2$. How many critical points in $[d]$ can there be at the most? Let h_S denote the ouptut hypothesis of your algorithm on a sample S. What does the event $\mathbf{err}(h_S; D) > \epsilon$ implies in terms of the relationship between S and critical points?