Instructor: Rocco A. Servedio

Computer Science 4252: Introduction to Computational Learning Theory
Problem Set #5 Fall 2023

Due 11:59pm Wednesday, December 6, 2023

See the course Web page for instructions on how to submit homework. **Important:** To make life easier for the TAs, **please start each problem on a new page.**

Problem 1 Let's consider (a slight generalization of) the simple-three-stage boosting scenario that we analyzed in class. Suppose that $h_1 : X \to \{0,1\}$ is such that under distribution $\mathcal{D}_1 = \mathcal{D}$, the weak hypothesis h_1 achieves error $p := \mathbf{Pr}_{\boldsymbol{x}\sim\mathcal{D}_1}[h(\boldsymbol{x}) \neq c(\boldsymbol{x})] < 1/2$ (in class we fixed $p = 0.4$).

(a) In class we described a way to simulate a draw from $EX(c,\mathcal{D}_2)$ given h_1 and access to $EX(c,\mathcal{D}_1)$. Using that approach, what is the expected number of calls to $EX(c,\mathcal{D}_1)$ that need to be performed in order to simulate a single draw from $EX(c,\mathcal{D}_2)$ in our current general-value-of-p setting?

(b) The approach of part (a) can be carried out without knowing the value of p. Suppose now that the value of p is known to you. Describe another way to simulate a draw from $EX(c,\mathcal{D}_2)$ given h_1 and access to $EX(c,\mathcal{D}_1)$, and analyze the efficiency of your approach (the expected number of calls to $EX(c,\mathcal{D}_1)$ needed to simulate a single draw from $EX(c,\mathcal{D}_2)$). Compare the relative efficiency of the two approaches.

(For both parts of the problem you may use standard facts about the geometric distribution as can be found in many sources, for example on Wikipedia.)

Problem 2 Let us say that an algorithm A is a *FNO weak learner with advantage γ for concept class $\mathcal{C}$* (for "False Negative Only") if it has the following property: for any $c \in \mathcal{C}$ and any distribution $\mathcal{D}$, given access to $EX(c,\mathcal{D})$, algorithm A outputs a hypothesis h such that

- If $c(x) = 0$ then $h(x) = 0$; and
- $\mathbf{Pr}_{x\sim\mathcal{D}_1}[h(x) = 1] \geq \frac{1}{2} + \gamma$. Here $\mathcal{D}_1$ is the distribution $\mathcal{D}$ restricted to $\{x : c(x) = 1\}$.

(For simplicity we do not consider the confidence parameter δ, i.e. we assume that an FNO weak learner always outputs a hypothesis as described above.)

Suppose you are given an FNO weak learner with advantage $\gamma = 1/10$ which in fact *always* (when run on $EX(c,\mathcal{D})$) outputs a hypothesis h such that $\mathbf{Pr}_{x\sim\mathcal{D}_1}[h(x) = 1] = 0.6$. Analogous to what we did in class, describe and analyze a simple *two-stage* booster for such a weak learner; i.e. your boosting algorithm should run the weak learner *twice* to produce its final hypothesis.

Problem 3 Let $\mathcal{C}$ be any concept class over $\{0,1\}^n$. Show that if $\mathcal{C}$ is efficiently PAC learnable, then there is an efficient algorithm that, given $0 < \delta < 1$ and a sample S of m examples labeled according to some concept c in $\mathcal{C}$, outputs with probability at least $1 - \delta$ a hypothesis h such that

(i) h is consistent with S, and

(ii) $\text{size}(h) \leq p(n, \text{size}(c), \log m, 1/\delta)$ for some polynomial p.

(**Hint:** Use the AdaBoost algorithm together with the efficient PAC learning algorithm for $\mathcal{C}$. You may define the size of a hypothesis h as you wish for this problem, provided that your definition is reasonable, and you may assume, for the purposes of this problem, that each real number which occurs in the representation of h contributes 1 to its size.)

(A cultural note unrelated to solving the problem: recall that the Occam's Razor theorem can be viewed as showing that "compression implies learnability." This problem essentially shows a converse in a very strong sense – note that the hypothesis h which encodes all m of the correct labels for the examples in S is of size only *poly-logarithmic* in m (ignoring other parameters).)

Problem 4 Let $\mathcal{C}$ be any concept class. Suppose that algorithm A is an efficient proper PAC learning algorithm for $\mathcal{C}$ in the noise-free setting (i.e. given access to an example oracle $EX(c, \mathcal{D})$, algorithm A outputs a hypothesis h which belongs to $\mathcal{C}$ and satisfies the PAC criteria). Suppose that moreover there is an efficient PAC learning algorithm B for concept class $\mathcal{C}$ in the presence of random classification noise, but B is not a proper PAC learning algorithm (i.e. the hypotheses which B outputs belong not to $\mathcal{C}$, but to some other hypothesis class H). Show that then there must exist an efficient *proper* PAC learning algorithm for $\mathcal{C}$ in the presence of random classification noise.

Problem 5 Two functions $f, g : X \to \{0,1\}$ are said to be *uncorrelated* with respect to distribution $\mathcal{D}$ over X if

$$\mathbf{Pr}_{\boldsymbol{x} \sim \mathcal{D}}[f(\boldsymbol{x}) = g(\boldsymbol{x})] = \mathbf{Pr}_{\boldsymbol{x} \sim \mathcal{D}}[f(\boldsymbol{x}) \neq g(\boldsymbol{x})].$$

(a) Recall from the third problem set that for S a subset of $\{1, \ldots, n\}$, the *parity function corresponding to* S is the Boolean function $\chi_S : \{0,1\}^n \to \{0,1\}$,

$$\chi_S(x_1, \ldots, x_n) = \sum_{i \in S} x_i \mod 2$$

which counts whether the number of input bits in coordinates indexed by S is odd or even. The concept class of *parity functions* $\mathcal{P}$ over $\{0,1\}^n$ includes all 2^n functions that can be described in this way. Formally,

$$\mathcal{P} = \{\chi_S : S \subseteq [n]\}.$$

Show that any two distinct parity functions χ_{S_1}, χ_{S_2} over $\{0,1\}^n$ are uncorrelated with respect to the uniform distribution over $\{0,1\}^n$. (By results that we will discuss in class, this means that

the class $\mathcal{P}$ of all parity functions over $\{0,1\}^n$ takes exponential time to learn in the Statistical Query model. Note that you don't need to know what the Statistical Query model is in order to solve this entire problem.)

(b) Let $\mathcal{C}_{DNF}$ be the class of n-term DNF formulas over $\{0,1\}^n$. Show that there is a distribution $\mathcal{D}$ over $\{0,1\}^n$ such that there are $N = n^{\Omega(\log n)}$ many pairwise uncorrelated concepts in $\mathcal{C}_{DNF}$. By the results we will discuss in class, this means that the class $\mathcal{C}_{DNF}$ does not have an efficient (polynomial in n time) learning algorithm in the Statistical Query model.

(c) Now let $\mathcal{C}_{DT}$ be the class of n-leaf decision trees over $\{0,1\}^n$ (decision trees with at most n leaves; see the Oct 4 lecture for a refresher on decision trees). Show that there is a distribution $\mathcal{D}$ over $\{0,1\}^n$ such that there are $N = n^{\Omega(\log n)}$ many pairwise uncorrelated concepts in $\mathcal{C}_{DT}$. By the results we will discuss in class, this means that the class $\mathcal{C}_{DT}$ does not have an efficient (polynomial in n time) learning algorithm in the Statistical Query model.