Instructor: Rocco A. Servedio

Computer Science 4252: Introduction to Computational Learning Theory
Problem Set #4 Fall 2023

Due 11:59pm Wednesday, November 15, 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 The "Chernoff bounds" we presented in class were bounds on the tail probability for a random variable $\boldsymbol{X} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_m$ where the $\boldsymbol{X}_i$'s are independent $\{0,1\}$-valued random variables, each of which takes value 1 with probability p. This bound is just the tip of a large iceberg — similar bounds are known under a much broader range of conditions. The point of the current problem is to explore (a little bit of) this.

Let's have some fun, and prove a related tail bound for sums of independent random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ that are *not* restricted to have values all in the set $\{0,1\}$.

(a) Show that if $\boldsymbol{Z}$ is any real-valued random variable, then for every $r > 0$ we have $\mathbf{Pr}[\boldsymbol{Z} > z] \leq e^{-rz}\,\mathbf{E}[e^{r\boldsymbol{Z}}]$.

(b) Let $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ be independent random variables that have $\mathbf{Pr}[\boldsymbol{Y}_i = 1] = \mathbf{Pr}[\boldsymbol{Y}_i = -1] = 1/2$, let $\boldsymbol{X}_i = w_i \boldsymbol{Y}_i$ for some real values $w_1, \ldots, w_n$, and let $\boldsymbol{X} = \boldsymbol{X}_1 + \cdots + \boldsymbol{X}_n$.

Show that for any real value r, we have $\mathbf{E}[e^{r\boldsymbol{X}}] \leq e^{r^2\|w\|^2/2}$, where $\|w\|^2 = w_1^2 + \cdots + w_n^2$. (**Hint:** Use (and justify if you can) the fact that $\cosh(x)$, which we recall is $\frac{1}{2}(e^x + e^{-x})$, is at most $e^{x^2/2}$.)

(c) Combine parts (a) and (b) to infer that for any $t > 0$, we have $\mathbf{Pr}[\boldsymbol{X} \geq t] \leq e^{-rt + r^2\|w\|^2/2}$.

(d) Use (c) to obtain the following tail bound on $\boldsymbol{X}$: $\mathbf{Pr}[\boldsymbol{X} \geq t] \leq e^{-t^2/(2\|w\|^2)}$.

Problem 2 In class we saw an algorithm for PAC learning monotone disjunctions which had the following property: if the algorithm is run on a target concept that is a monotone disjunction of length at most k, it outputs a hypothesis which is a monotone disjunction of length at most only only slightly longer than k. In this problem you'll show that that it is a computationally hard problem to PAC learn using a hypothesis whose length is at most *exactly* k.

More precisely, suppose that there is a PAC learning algorithm $\mathcal{A}$ for monotone disjunctions that runs in time $\text{poly}(n, 1/\varepsilon, 1/\delta)$ and has the following property: for all k, if $\mathcal{A}$ is run on a monotone disjunction of length k, it outputs a hypothesis that is a monotone disjunction of length at most k. Show that then there is a randomized $\text{poly}(n)$-time algorithm which optimally solves any instance of SET COVER with high probability. (Since SET COVER is NP-complete, this would mean that NP is contained in RP, which is viewed as being very unlikely.)

Problem 3 Let C be a concept class whose VC dimension is d, and for $s \geq 1$ denote by C_s the class $C_s = \{c = c_1 \cup \ldots \cup c_s \mid c_i \in C\}$. Show that for all $s \geq 1$, the VC dimension of C_s is at most $2ds \log(3s)$. (Hint: Think about the growth function $\Pi_C(m)$.)

Problem 4
(a) Let $X = \mathbb{N} = \{0, 1, 2, 3, \ldots\}$ and let $\mathcal{C}_+$ be the concept class over X defined as follows:

$$\mathcal{C}_+ = \{c_{+0}, c_{+1}, c_{+2}, c_{+3}, \ldots\} \quad \text{where} \quad c_{+i} = \{i + j : j \in \mathbb{N}\}$$

(so, for example, the concept c_{+4} is the subset of X defined as $c_{+4} = \{4, 5, 6, 7, \ldots\}$). Give an efficient PAC learning algorithm for $\mathcal{C}_+$ and explain why your algorithm is correct (analyze its sample complexity and running time).
(b) As in part (a), let $X = \mathbb{N} = \{0, 1, 2, 3, \ldots\}$ but now let $\mathcal{C}_\times$ be the concept class over X defined as follows:

$$\mathcal{C}_\times = \{c_{\times 0}, c_{\times 1}, c_{\times 2}, c_{\times 3}, \ldots\} \quad \text{where} \quad c_{\times i} = \{i \times j : j \in \mathbb{N}\}$$

(so, for example, the concept $c_{\times 4}$ is the subset of X defined as $c_{\times 4} = \{0, 4, 8, 12, \ldots\}$). Argue that there is no PAC learning algorithm for $\mathcal{C}_\times$ (if you are not sure exactly what this means, see the next problem for clarification).

Problem 5 Part (b) of the previous problem implies that there is no *a priori* fixed sample size which suffices for PAC learning the concept class $\mathcal{C}_\times$ for all distributions. To be more precise, it tells us that there is no function $m(1/\varepsilon, 1/\delta)$ such that the following holds: There is an algorithm which, given ε, δ and access to $EX(c, \mathcal{D})$ where $\mathcal{D}$ is any distribution over $\mathbb{N}$ and c is an unknown target concept $c_{\times i}$ in $\mathcal{C}_\times$, draws $m(1/\varepsilon, 1/\delta)$ samples from $EX(c, \mathcal{D})$ and with probability $1 - \delta$ outputs an ε-accurate hypothesis for c.

However, while there is no fixed sample size $m(1/\varepsilon, 1/\delta)$ that suffices for every distribution, in fact for every distribution there is some finite sample size that suffices for it. Establishing this, for the concept class $\mathcal{C}_\times$, is the point of the current problem.

Show that for every distribution $\mathcal{D}$ over $\mathbb{N}$, there is a function $m_{\mathcal{D}}(1/\varepsilon, 1/\delta)$ (which may depend on $\mathcal{D}$) and an algorithm $A_{\mathcal{D}}$ (which also may depend on $\mathcal{D}$) such that the following holds: If $A_{\mathcal{D}}$ is given ε, δ and access to $EX(c_{\times i}, \mathcal{D})$ where $c_{\times i}$ is an unknown element of $\mathcal{C}_{\times i}$, it draws $m_{\mathcal{D}}(1/\varepsilon, 1/\delta)$ samples from $EX(c, \mathcal{D})$ and with probability $1 - \delta$ outputs an ε-accurate hypothesis for c.