Statistical Learning Theory (S&DS 669)
Problem Set 2

Instructor Name: Omar Montasser
TA: Herlock Rahimi
Due Date: Tuesday, October 14, 2025 (11:59 PM ET).

Instructions

- You may collaborate with your classmates, but each student must write solutions on their own. You also need to write down who you worked with.
- If you use a language model, then you need to provide a transcript of the interaction as a supplement. Refer to the policy on the Canvas page for more guidelines.
- If you use any sources other than the class material and references available on Canvas, then mention that. It's fine to look up a complicated sum or ineqality, but don't look up an entire solution.
- Submit your solutions in LaTeX by uploading a PDF to Canvas.

Problems

Problem 1. For any family of hypothesis classes $\mathcal{H}_n \subseteq \{\pm 1\}^{\mathcal{X}_n}$, where $\mathcal{X}_n = \{0,1\}^n$, consider the following decision problem:

$$\text{AGREEMENT}_{\mathcal{H}_n} = \{(S,k) \mid S \subseteq \mathcal{X}_n \times \{\pm 1\}, k \in \mathbb{Z}, \exists h \in \mathcal{H}_n \text{ s.t. } |\{(x,y) \in S \mid h(x) = y\}| \geq k\}.$$

Prove that if $\mathcal{H}_n$ is efficiently agnostically properly PAC learnable then $\text{AGREEMENT}_{\mathcal{H}_n} \in$ RP.

Problem 2 (Learning Sparse Linear Predictors)**.** Let $\mathcal{X}_n = \{0,1\}^n$, for any $1 \leq k \leq n$ define

$$\mathcal{H}_n^k = \{h_{w,\theta} \mid \|w\|_0 \leq k, \theta \in \mathbb{R}\}$$

where

$$h_{w,\theta}(x) = \begin{cases} 1 & \langle w, x \rangle \geq \theta \\ -1 & \text{otherwise,} \end{cases} \text{ and } \|w\|_0 := \left|\{j \in [n] : w_j \neq 0\}\right|$$

(a) Show that $\text{vc}(\mathcal{H}_n^k) \leq O\left(k \log(\frac{n}{k})\right)$. [*Hint: use Sauer's lemma.*]

(b) Provide an explicit learning rule A and show that for any distribution D over $\mathcal{X}_n \times \{\pm 1\}$, the following holds with probability at least $1-\delta$ over the draw of $S \sim D^m$:

$$L_D(A(S)) \leq \inf_{w,\theta} \left\{ L_D(h_{w,\theta}) + O\left(\sqrt{\frac{\|w\|_0 \log n + \log \frac{1}{\delta}}{m}} \right) \right\}.$$

(c) Prove the above bound using the following different learning rule (called, validation rule):

- Split S into two equal subsets S_1 and S_2.
- For each $1 \leq k \leq n$, let $w_k := \arg\min_w L_{S_1}(w)$ s.t. $\|w\|_0 \leq k$.
- Let $\hat{k} := \arg\min_k L_{S_2}(w_k)$.
- Return $w_{\hat{k}}$.

Problem 3 (Boosting Confidence δ)**.** Recall the definition of (ϵ, δ)-realizable-PAC-learning,

Definition 1. We say that a learner $\mathcal{A}$ (ϵ, δ)-PAC-learns a hypothesis class $\mathcal{H}$ (in the realizable setting) if $\exists m(\epsilon, \delta) \in \mathbb{N}$ such that for any distribution D where $\inf_{h \in \mathcal{H}} L_D(h) = 0$, with probability at least $1-\delta$ over $S \sim D^{m(\epsilon,\delta)}$, $L_D(\mathcal{A}(S)) \leq \epsilon$.

In this problem, we will explore boosting the confidence parameter δ. Specifically, suppose that we are given an $(\epsilon, 1/2)$-PAC-learner $\mathcal{A}$ for a hypothesis class $\mathcal{H}$. That is, learner $\mathcal{A}$ succeeds in outputting a low-error hypothesis only with probability $1/2$. We will use learner $\mathcal{A}$ to construct another learner $\mathcal{B}$ that (ϵ, δ)-PAC-learns $\mathcal{H}$ (for any δ).

Learner $\mathcal{B}$. Run learner $\mathcal{A}$ on N different iid sets $S_1, \ldots, S_N \sim D^{m_{\mathcal{A}}(\epsilon/2, 1/2)}$ where each set S_i is of size $m_{\mathcal{A}}(\epsilon/2, 1/2)$ (i.e., this is the sample complexity of learner $\mathcal{A}$). This produces N hypotheses $h_1 = \mathcal{A}(S_1), \ldots, h_N = \mathcal{A}(S_N)$. Draw an additional test set $S' \sim D^m$ of size m, and output $h_{i^\star}$ where $i^\star = \mathrm{argmin}_{1 \leq i \leq N} L_{S'}(h_i)$. That is, we output the hypothesis that achieves the smallest error on test set S' among the N hypotheses.

Note that in the description of learner $\mathcal{B}$ we have left N (the number of times of running $\mathcal{A}$) and m (the size of the test set) unspecified. In order to guarantee that learner $\mathcal{B}$ (ϵ, δ)-PAC-learns $\mathcal{H}$,

(a) What should N be so that at least one hypothesis among $h_1, \ldots, h_N$ has error at most $\epsilon/2$ on D, with probability at least $1 - \delta/2$? (The answer should be a function of δ).

(b) Using Chernoff bounds, determine an explicit bound on the size of the test set m such that the hypothesis that performs best on the test set S' has error at most ϵ on D with probability at least $1 - \delta/2$ over $S' \sim D^m$? **Hint:** determine a size m and a threshold τ such that with high probability, the promised good hypothesis has error at most τ on S', and all hypotheses with error more than ϵ on D have error more than τ on S'.

Note that (a) and (b) combined together imply that $\mathcal{B}$ (ϵ, δ)-PAC-learns $\mathcal{H}$.

(c) Write down the sample complexity of learner $\mathcal{B}$, $m_{\mathcal{B}}(\epsilon, \delta)$, as a function of $m_{\mathcal{A}}(\epsilon/2, 1/2)$, N, and m.

Problem 4 (Boosting and Hardness of Efficient Learning)**.** Recall the class of halfspaces over $\mathbb{R}^n$

$$\mathcal{H}_n = \{h_w : \mathbb{R}^n \to \{\pm 1\} \mid h_w(x) = \text{sign}(\langle w, x \rangle), w \in \mathbb{R}^n\},$$

and the class of intersection of k halfspaces ($k > 1$)

$$\mathcal{H}_n^k = \{h(x) = h_{w_1}(x) \wedge h_{w_2}(x) \wedge \cdots \wedge h_{w_k}(x) \mid h_{w_1}, \ldots, h_{w_k} \in \mathcal{H}_n\},$$

where

$$h(x) = \begin{cases} +1 & \text{if } h_{w_1}(x) = \cdots = h_{w_k}(x) = +1 \\ -1 & \text{otherwise.} \end{cases}$$

In this problem, we will prove the following claim:

Theorem 1. If $\mathcal{H}_n$ is efficiently-agnostically-PAC-learnable, then for any polynomial $k(n) = \text{poly}(n)$, $\mathcal{H}_n^{k(n)}$ is efficiently-PAC-learnable (in the realizable setting).

(a) Prove that for any distribution D over $\mathcal{X} \times \mathcal{Y}$, if there exists $h \in \mathcal{H}_n^{k(n)}$ such that $L_D(h) = 0$, then there exists a halfspace $h_w \in \mathcal{H}_n$ such that $L_D(h_w) \leq 1/2 - 1/2k^2$. [*Hint: use the probabilistic method.*]

(b) By relying on Part (a) and the assumption in Theorem 1 that $\mathcal{H}_n$ is efficiently-agnostically-PAC-learnable, suggest a weak-PAC-learner for $\mathcal{H}_n^{k(n)}$ that runs in time polynomial in n. Explicitly state the error parameter ϵ and confidence parameter δ of the weak-PAC-learner, and the number of samples required as a function of n and $k(n)$.

(c) Use AdaBoost to establish the claim in Theorem 1. Write down the sample complexity and runtime of AdaBoost, as a function of the sample complexity and runtime of the weak-PAC-learner from Part (b). This should be expressed as a function of $n, k(n), \epsilon, \delta$.

A Remark. Several results in the literature (see e.g., [Tie24] and references therein) show hardness of efficiently weakly-PAC-learning $\mathcal{H}_n^{k(n)}$ (i.e., interesection of $k(n)$ halfspaces) under standard cryptographic assumptions. Combined with Theorem 1, this implies hardness of efficiently agnostically-PAC-learning $\mathcal{H}_n$ (i.e., halfspaces). This is an example of how boosting can be used to prove a computational hardness result.

References

[Tie24] Stefan Tiegel. Improved hardness results for learning intersections of halfspaces. In Shipra Agrawal and Aaron Roth, editors, *The Thirty Seventh Annual Conference on Learning Theory, June 30 - July 3, 2023, Edmonton, Canada*, volume 247 of *Proceedings of Machine Learning Research*, pages 4764–4786. PMLR, 2024.