Computational and Statistical Learning Theory

Problem set 2

Due: May 1st

Please send your solutions to `learning-submissions@ttic.edu`

Notation:

- Input space: $\mathcal{X}$
- Label space: $\mathcal{Y} = \{\pm 1\}$
- Sample: $(x_1, y_1), \ldots, (x_m, y_m) \in \mathcal{X}$
- Hypothesis Class: $\mathcal{H}$
- Risk: $L_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbf{1}_{h(x)\neq y}\right]$
- Empirical Risk: $L_S(h) = \frac{1}{m}\sum_{(x,y)\in S}\mathbf{1}_{h(x)\neq y}$

1. Let $\mathcal{X} = \{0,1\}^*$. Write down a precise learning rule A (eg. based on the SRM++ rule from class) s.t. for any computable learning rule B, there exists a number c s.t. for any $\delta < 1, L_0 < 1$, for any sample size m, and any distribution $\mathcal{D}$, if w.p.$\geq 1-\delta$ over $S \sim \mathcal{D}^m$, $L_{\mathcal{D}}(B(S)) \leq L_0$, then w.p. $\geq 1-\delta/2$ over $S \sim \mathcal{D}^{m'(m,c,\epsilon,\delta)}$, $L_{\mathcal{D}}(A(S)) \leq L_0 + \epsilon$. Prove an explicit bound that is as tight as you can (do not worry about constant factors) for $m'(m, c, \epsilon, \delta)$.

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

$$\textsc{Agreement}_{\mathcal{H}} = \{(S,k) \mid S \subseteq \mathcal{X}_n \times \{\pm 1\}, k \in \mathbb{Z}, \exists_{h\in\mathcal{H}_n} |\{(x,y)\in S \,|\, h(x) = y\}| \geq k\}$$

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

3. Let $\mathcal{X}_n = \{0,1\}^n$, for any function $k(n)$ define

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

   where

$$h_{w,\theta(x)} = \begin{cases} 1 & \langle w, x\rangle \geq \theta \\ -1 & \text{otherwise.} \end{cases}$$

For each of the following $k(n)$, establish whether the hypothesis class is efficiently properly PAC learnable, efficiently properly agnostically PAC learnable, efficiently PAC learnable and efficiently agnostically PAC learnable (under reasonable complexity assumptions). That is, for each of these four different settings, if it is learnable, provide an appropriate learning rule, state the sample complexity, and note whether it is greater than the best sample complexity possible with any (even intractable) learning rule. If it is not learnable, then briefly discuss why it is not learnable (you use the results presented in the class).

(a) $k(n) = 3$

(b) $k(n) = \sqrt{n}$ (Efficient proper learning is extra credit)

(c) $k(n) = n$

4. Consider $\mathcal{X} = \mathbb{R}^d$, the class of linear predictors $\mathcal{H} = \{h_w(x) \mapsto \langle w, x \rangle \,|w \in \mathbb{R}^d\}$ and learning by minimizing the hinge loss versus the zero-one error.

   (a) Show that for any $\epsilon > 0$ and $\alpha < 1$, there exists a sample S such that $\inf_w L_S^{01}(h_w) \leq \epsilon$ but for any $w = \arg\min_w L_S^{\text{hinge}}(h_w)$, $L_S^{01}(h_w) > \alpha$.

   (b) We now consider a model that is more restricted that the agnostic model, but still allows for errors. Specifically, we consider the random classification noise model. In the random classification noise model we assume the following on the source distribution $\mathcal{D}(x, y)$: there exists a linear predictor $x \mapsto \langle w_0, x \rangle$ such that $y \perp x \mid \text{sign}(\langle w_0, x \rangle)$ and $\mathbb{P}(y = 1| \langle w_0, x \rangle > 0) = 1 - p$ for some noise probability $p < 1/2$. Prove that for any distribution of this form, $L_{\mathcal{D}}^{01}(\arg\min_w L_{\mathcal{D}}^{\text{hinge}}(h_w)) = \inf L_{\mathcal{D}}^{01}(h_w)$. That is, at least on the distribution, minimizing the hinge loss does minimize the zero-one error.

   (c) (Extra credit) Prove that for any ϵ, δ, there exists $m(\epsilon, \delta)$ s.t. for any source distribution $\mathcal{D}$ of the form above, w.p.$> 1 - \delta$, $L_{\mathcal{D}}^{01}(\arg\min_w L_S(\mathcal{D}^{\text{hinge}}(h_w)) \leq \inf_w L_{\mathcal{D}}^{01}(h_w) + \epsilon$. That is, minimizing the hinge loss is an efficient learning algorithm under the random classification noise assumption.

5. (Optional) Prove that for the class $\mathcal{H}_n$ of half-spaces (linear predictors) over $\{0, 1\}^n$, the problem $\textsc{Agreement}_{\mathcal{H}}$ is NP-hard.

   Hint: Consider the decision problem $\textsc{HittingSet}$:

$$\textsc{HittingSet} = \left\{(C, k) \;\middle|\; C \subseteq 2^{[n]}, \exists_{R,|R|=k} \forall_{A \in C} A \cap R \neq \emptyset\right\}$$

   That is, the input is a collection C of subset of the integers $1..n$, and an integer k, and the problem is to decide whether there exists a set of cardinality at most k that "hits" (has non-empty intersection) with all sets in C. The problem $\textsc{HittingSet}$ is a classic NP-hard problem, and you may base your proof on this fact.

   First, show that a restricted version of $\textsc{HittingSet}$ where all sets in C are required to be the same size is also NP-hard (e.g. show a simple reduction from $\textsc{HittingSet}$). Then, consider the following mapping from inputs (C, k), where all sets in C are of cardinality exactly t, to a labeled sample in $\mathbb{R}^{sn}$ (for convenience, we will index vectors in $\mathbb{R}^{sn}$ as $v_{i,j}$ where $1 \leq i \leq s$ and $1 \leq j \leq n$, and denote $e_{i,j}$ the vector of all-zeros except a single one at (i, j)):

- Positive points at $\sum_{i=1}^{s} e_{i,j}$ for each $j = 1 \ldots n$.
- Negative points at $\sum_{j \in A} e_{i,j}$ for each $i = 1 \ldots s$ and each $A \in C$.

Use the above mapping to construct a reduction from the restricted version of HITTINGSET to AGREEMENT$_{\mathcal{H}}$.