Computational and Statistical Learning Theory

Problem set 2

Due: Monday, October 24th

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. 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 $\textsc{Agreement}_{\mathcal{H}} \in \mathbf{RP}$.

2. 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 $k(n)$, answer the following question:

- What is the VC dimension of the hypothesis class ?
- Is the problem efficiently properly PAC learnable? If yes, with what sample complexity and what runtime?
- Is the problem efficiently PAC learnable? If yes, with what sample complexity and what runtime? (Extra credit) Is there a gap in sample complexity vs the best possible with an intractable rule?
- Is the problem efficiently agnostically properly PAC learnable? If yes, with what sample complexity and what runtime?
- Is the problem efficiently agnostically PAC learnable? If yes, with what sample complexity and what runtime? Is there a gap in sample complexity vs the best possible with an intractable rule?

(a) $k(n) = 2$

(b) $k(n) = 3$

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

(d) $k(n) = n$ (Efficient agnostic proper learning is extra credit)

Hint: Consider the decision problem HITTINGSET:

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

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 HITTINGSET is a classic NP-hard problem, and you may base your proof on this fact.

First, show that a restricted version of HITTINGSET where all sets in C are required to be the same size is also NP-hard (e.g. show a simple reduction from 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$_{\mathcal{H}}$.

3. 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 is independent of x given $\text{sign}(\langle w_0, x\rangle)$ 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.

Extra-credit Challenge Problems :

1. Prove that for any polynomial $p(n)$, there exists a family $\mathcal{H}_n$ of hypothesis, such that $\mathcal{H}_n$ is (not necessarily efficiently) PAC learnable with poly$(\log n, 1/\epsilon, \log 1/\delta)$ examples, but that any polynomial-time learning algorithm for $\mathcal{H}_n$ needs at least $p(n)$ examples in order to get error less than 0.1 with probability at least $1/2$. You can follow the steps as:

   I) Use counting argument to prove existence, where we do not require that $h \in \mathcal{H}_n$ are computable. II) Use some acceptable cryptographic assumption and prove the existence of such a class, with all $h \in \mathcal{H}_n$ computable in time poly(n).

Research Problems :

1. Show how a VC-based learning guarantee can be obtained from the PAC-Bayes bound. That is, for any class with VC dimension d, describe a prior p and a learning rule that returns a distribution (randomized hypothesis) q, for which the PAC-Bayes bound leads to a VC-based learning guarantee.