Computational and Statistical Learning Theory

Problem set 8

Due: December 1st

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

Problems

1. **Stability :**

   (a) Use McDiamird's inequality to show that if a learning rule $\mathcal{A}$ is $\epsilon(m)$ stable, then with probability greater than $1-\delta$,

   $$L(\mathcal{A}) \leq \hat{L}(\mathcal{A}) + f(\epsilon(m), \log(1/\delta), m)$$

   where f has only a polynomial dependence on $\log(1/\delta)$. Write down f explicitly.

   (b) Use the above guarantee to analyze the RERM learning rule $\mathcal{A}(S) = \operatorname{argmin} \hat{L}(w) + \frac{\lambda}{2}\|w\|_2^2$, for linear prediction with a Lipschitz bounded loss, and obtain a learning guarantee for the norm-bounded linear predictor class $\{w | \|w\|_2 \leq B\}$.

   Write down the resulting sample complexity, as well as the sample complexity obtained from the stability-based analysis we did in class, and the sample complexity of ERM we obtained from concentration-based arguments.

2. **Boosting :**
   For any binary hypothesis class $\mathcal{H}$ over $\mathcal{X}$, and some $\epsilon < \frac{1}{2}$ and $\gamma < 1$, assume there exists a learning rule $\mathcal{A}(S)$ and a sample size m s.t. for any distribution $D(\mathcal{X}, \pm 1)$ where $\inf_{h \in \mathcal{H}} L_D(h) = 0$, we have that w.p. $\geq 1-\gamma$ over $S \sim D^m$, $L(\mathcal{A}(S)) < \epsilon$. Prove an upper bound on the VC-dimension of $\mathcal{H}$ as a function of m, ϵ and δ, with a polynomial dependence on m. We can conclude that from a statistical perspective, weak learning implies strong learning.

3. **Boosting :**

(a) Consider linear prediction with the $0/1$ loss using the class of sparse linear predictors: $\{w \in R^d | \|w\|_0 \leq B\}$, where $\|w\|_0$ is the number of non-zeros in w, over $\mathcal{X} = R^d$. Analyze the VC-dimension of this class.

(b) Let $\mathcal{H}$ be a binary (± 1) hypothesis class over $\mathcal{X}$ with VC-dim($\mathcal{H}$) $\leq d$. Consider the class

$$\mathcal{H}_B = \{x \mapsto \text{sign}(\sum_{i=1\ldots B} \alpha_i h_i) | \alpha_i \in R, h_i \in \mathcal{H}\}$$

Analyze the VC-dimension of this class.

4. **Boosting :**
Derive a length-$(d+1)$ compression scheme for a learning rule which is an ERM over linear separators in R^d (with respect to the $0/1$ error).

5. **Boosting :**
Combine the confidence boosting and accuracy boosting arguments to rigorously show that weak learning implies strong learning. In particular, considering a hypothesis class $\{\mathcal{H}_n\}$, if we have a learning rule $\mathcal{A}$ s.t. for some $\epsilon = 1 - \gamma < 2$ and some $\delta < 1$, for every n and every distribution D over $\mathcal{X}_n$, w.p.$\geq 1 - \delta$, $L(\mathcal{A}(D)) < \epsilon$ and $\mathcal{A}$ requires $m_{\mathcal{A}}(n) \geq \text{poly}(n)$ samples and runtime, specify how AdaBoost can be used to obtain $L(\tilde{\mathcal{A}}(D, \tilde{\epsilon}, \tilde{\delta})) \leq \tilde{\epsilon}$ w.p. $\geq 1 - \tilde{\delta}$ with sample and runtime complexity $\tilde{m}(n) \leq \text{poly}(n, 1/\tilde{\epsilon}, \log(1/\tilde{\delta}))$. Provide an explicit expression for the bound on the resulting sample complexity.