CSE 250: Machine Learning Theory Spring 2017

Mid-term Exam

Instructor: Raef Bassily **Due on:** Thu May 25

Instructions

- Prove your claims. You may use any result covered in class, but you must **cite** the result that you use.

Problem 1 (10 points)

Let $\mathcal{H}$ be a hypothesis class defined over a domain $\mathcal{X}$ where $VC(\mathcal{H}) = K$. Let $\gamma = 1/4$.

(a) Suppose we are given a function $N_{\mathcal{H}} : (0, 1) \to \mathbb{N}$ and a γ-weak learner $\mathcal{A}$ for $\mathcal{H}$ that for every $\delta_0 \in (0, 1)$ and every distribution D over $\mathcal{X} \times \{0, 1\}$, given $N_{\mathcal{H}}(\delta_0)$ i.i.d. training examples from D, outputs a hypothesis $h \in \mathcal{H}$ that has true error $\mathbf{err}(h; D) \leq 1/2 - \gamma$ with probability at least $1 - \delta_0$. Suppose we want to use Adaboost algorithm to construct a strong learner based on $\mathcal{A}$.

Fix any $\epsilon \in (0, 1/4)$ and $\delta \in (0, 1)$. Let D be any distribution over $\mathcal{X} \times \{0, 1\}$. Suppose you can draw i.i.d. examples from D that can be used as an input traninig set S for Adaboost. Describe your instantiation of the Adaboost algorithm **specifying the number of rounds** T **and the size of the training set** S such that with probability at least $1 - \delta$, the output hypothesis $\tilde{h}$ has true error $\mathbf{err}(\tilde{h}; D) \leq \epsilon$. (Express the required quantities in terms of ϵ, δ, K. You can use the asymptotic $O(\cdot)$ notation in your expressions, i.e., you may ignore constant factors.) **(7 points)**

(b) Suppose that for each $\delta_0 \in (0, 1)$, the running time of the weak-learner $\mathcal{A}$ (when given $N_{\mathcal{H}}(\delta_0)$ examples) is at most $\tau(\delta_0)$. Derive an upper bound on the running time of your Adaboost. (State your bound in terms of ϵ, δ, K, and $\tau(\cdot)$. You can use the asymptotic $O(\cdot)$ notation in your expressions, i.e., you may ignore constant factors.) **(3 points)**

Problem 2 (10 points)

(a) For the domain $\mathcal{X} = \mathbb{R}^d$, $d \geq 2$, consider the class of decision stumps defined as

$$\mathcal{H} = \cup_{i=1}^{d} \mathcal{H}_i$$

where

$$\mathcal{H}_i = \{ \; \mathbf{x} \mapsto \mathsf{sign}\,(a x_i + b) : a \in \{-1, +1\}\,, \; b \in \mathbb{R} \; \}$$

where x_i is the i-th component of $\mathbf{x}$, $i \in [d]$. That is, $\mathcal{H}_i$ is the class of bi-directional thresholds based on the i-th component of an input data point $\mathbf{x} \in \mathbb{R}^d$.

Show that the VC dimension of $\mathcal{H}$ is upper-bounded by $5 + 2\log_2(d)$. **(5 points)**

Hint: You may use the following fact: **Fact:** Suppose $\beta, u > 0$. Then, $u \leq 2\log_2(u) + \beta \;\; \Rightarrow \;\; u \leq 5 + 2\beta$.

(b) Consider the class $\mathcal{G}$ of **conjunctions constructed on the basis of decision stumps**, i.e., on the basis of the classes $\{\mathcal{H}_1, \ldots, \mathcal{H}_d\}$ defined in Part 1. More formally, $\mathcal{G}$ is the class that contains all hypothesis $g : \mathbb{R}^d \to \{0, 1\}$ of the form

$$g(\mathbf{x}) = f\left(h_1\left(\mathbf{x}\right), \ldots, h_d\left(\mathbf{x}\right)\right), \; \mathbf{x} \in \mathbb{R}^d$$

where $h_i \in \mathcal{H}_i, i \in [d]$ is a decision stump, and f is a conjunction over d boolean variables. (Here, we will regard the output label of each decision stump as a bit in $\{0, 1\}$).

Give an upper bound (that is as tight as possible) on the VC dimension of the class $\mathcal{G}$. Hence, show that $\mathcal{G}$ is agnostic PAC learnable and give an upper bound on the sample complexity for learning $\mathcal{G}$. **(5 points)**

Problem 3 (10 points)

Consider the hypothesis class $\mathcal{H}$ containing axis-aligned rectangles where the lower-left corner is fixed at the origin. More specifically, for the domain $\mathcal{X} = \mathbb{R}^2$, define

$$\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a \geq 0, b \geq 0\},$$

where

$$h_{a,b}(x_1, x_2) = \begin{cases} 1, & \text{if } x_1 \in [0, a] \text{ and } x_2 \in [0, b], \\ 0, & \text{otherwise.} \end{cases}$$

(Note: this is a subset of the hypothesis class of axis-aligned rectangles, which we studied in the lectures.)

(a) *Without using VC dimension*, show that $\mathcal{H}$ is PAC-learnable. In particular, you should describe an algorithm that PAC-learns $\mathcal{H}$, and derive a corresponding bound on the sample complexity, expressed as a function $n_{\mathcal{H}}(\epsilon, \delta)$. Please compute a concrete bound, avoiding big-O notation. **(6 points)**

(b) Calculate the VC dimension of $\mathcal{H}$ and use this to provide a bound on the sample complexity, expressed as a function $n_{\mathcal{H}}(\epsilon, \delta)$. For this part of the problem, an asymptotic bound expressed using big-O notation is sufficient. **(4 points)**