Mehryar Mohri
Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 1
September 16, 2016
Due: October 04, 2016

A. Probability tools

1. Let $f\colon (0, +\infty) \to \mathbb{R}$ be a function admitting an inverse f^{-1} and let X be a random variable. Show that if for any $t > 0$, $\Pr[X > t] \leq f(t)$, then, for any $\delta > 0$, with probability at least $1 - \delta$, $X \leq f^{-1}(\delta)$.

2. Let X be a discrete random variable taking non-negative integer values. Show that $\mathrm{E}[X] = \sum_{n \geq 1} \Pr[X \geq n]$ (*hint*: rewrite $\Pr[X = n]$ as $\Pr[X \geq n] - \Pr[X \geq n + 1]$).

B. Label bias

1. Let D be a distribution over $\mathcal{X}$ and let $f\colon \mathcal{X} \to \{-1, +1\}$ be a labeling function. Suppose we wish to find a good approximation of the label bias of the distribution D, that is of p_+ defined by:

$$p_+ = \Pr_{x \sim D}[f(x) = +1]. \tag{1}$$

Let S be a finite labeled sample of size m drawn i.i.d. according to D. Use S to derive an estimate $\widehat{p}_+$ of p_+. Show that for any $\delta > 0$, with probability at least $1 - \delta$, $|p_+ - \widehat{p}_+| \leq \sqrt{\frac{\log(2/\delta)}{2m}}$ (carefully justify all steps).

C. Learning in the presence of noise

1. In Lecture 2, we showed that the concept class of axis-aligned rectangles is PAC-learnable. Consider now the case where the training points received by the learner are subject to the following noise: points negatively labeled are unaffected by noise but the label of a positive training point is randomly flipped to negative with probability $\eta \in (0, \frac{1}{2})$. The exact value of the noise rate η is not known to the learner but an upper bound η' is supplied to him with $\eta \leq \eta' < 1/2$. Show that the algorithm described in class returning the tightest rectangle containing positive points can still PAC-learn axis-aligned

rectangles in the presence of this noise. To do so, you can proceed using the following steps:

(a) Using the notation of the lecture slides, assume that $\Pr[R] > \epsilon$. Suppose that $error(R') > \epsilon$. Give an upper bound on the probability that R' misses a region r_j, $j \in [1, 4]$ in terms of ϵ and η'?

(b) Use that to give an upper bound on $\Pr[error(R') > \epsilon]$ in terms of ϵ and η' and conclude by giving a sample complexity bound.

2. [Bonus question] In this section, we will seek a more general result. We consider a finite hypothesis set H, assume that the target concept is in H, and adopt the following noise model: the label of a training point received by the learner is randomly changed with probability $\eta \in (0, \frac{1}{2})$. The exact value of the noise rate η is not known to the learner but an upper bound η' is supplied to him with $\eta \leq \eta' < 1/2$.

   (a) For any $h \in H$, let $d(h)$ denote the probability that the label of a training point received by the learner disagrees with the one given by h. Let h^* be the target hypothesis, show that $d(h^*) = \eta$.

   (b) More generally, show that for any $h \in H$, $d(h) = \eta + (1 - 2\eta)\, error(h)$, where $error(h)$ denotes the generalization error of h.

   (c) Fix $\epsilon > 0$ for this and all the following questions. Use the previous questions to show that if $error(h) > \epsilon$, then $d(h) - d(h^*) \geq \epsilon'$, where $\epsilon' = \epsilon(1 - 2\eta')$.

   (d) For any hypothesis $h \in H$ and sample S of size m, let $\widehat{d}(h)$ denote the fraction of the points in S whose labels disagree with those given by h. We will consider the algorithm L which, after receiving S, returns the hypothesis h_S with the smallest number of disagreements (thus $\widehat{d}(h_S)$ is minimal). To show PAC-learning for L, we will show that for any h, if $error(h) > \epsilon$, then with high probability $\widehat{d}(h) \geq \widehat{d}(h^*)$. First, show that for any $\delta > 0$, with probability at least $1 - \delta/2$, for $m \geq \frac{2}{\epsilon'^2} \log \frac{2}{\delta}$, the following holds:

$$\widehat{d}(h^*) - d(h^*) \leq \epsilon'/2$$

   (e) Second, show that for any $\delta > 0$, with probability at least $1 - \delta/2$, for $m \geq \frac{2}{\epsilon'^2}(\log |H| + \log \frac{2}{\delta})$, the following holds for all $h \in H$:

$$d(h) - \widehat{d}(h) \leq \epsilon'/2$$

(f) Finally, show that for any $\delta > 0$, with probability at least $1 - \delta$, for $m \geq \frac{2}{\epsilon^2(1-2\eta')^2}(\log|H| + \log\frac{2}{\delta})$, the following holds for all $h \in H$ with $error(h) > \epsilon$:

$$\widehat{d}(h) - \widehat{d}(h^*) \geq 0.$$

(*hint*: use $\widehat{d}(h) - \widehat{d}(h^*) = [\widehat{d}(h) - d(h)] + [d(h) - d(h^*)] + [d(h^*) - \widehat{d}(h^*)]$ and use previous questions to lower bound each of these three terms).