Mehryar Mohri
Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 1
September 17, 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 $1 - \delta$, $X \leq f^{-1}(\delta)$.

   Solution: For any $\delta > 0$, let $t = f^{-1}(\delta)$. Plugging this in $\Pr[X > t] \leq f(t)$ yields $\Pr[X > f^{-1}(\delta)] \leq \delta$, that is $\Pr[X \leq f^{-1}(\delta)] \geq 1 - \delta$. □

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 $\Pr[X \geq n] - \Pr[X \geq n + 1]$).

   Solution: We assume that X is a bounded random variable to avoid any convergence issues (although the statement is still true in the general case).

   By definition of expectation and using the hint, we can write

   $$\mathrm{E}[X] = \sum_{n \geq 0} n \Pr[X = n] = \sum_{n \geq 1} n(\Pr[X \geq n] - \Pr[X \geq n + 1]).$$

   Note that in this sum, for $n \geq 1$, $\Pr[X \geq n]$ is added n times and subtracted $n - 1$ times, thus $\mathrm{E}[X] = \sum_{n \geq 1} \Pr[X \geq n]$.

   More generally, by definition of the Lebesgue integral, for any non-negative random variable X, the following identity holds:

   $$\mathrm{E}[X] = \int_0^{+\infty} \Pr[X \geq t] \, dt.$$

   □

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).

Solution: Let $\widehat{p}_+$ be the fraction of positively labeled points in $S = (x_1, \ldots, x_m)$:

$$\widehat{p}_+ = \frac{1}{m} \sum_{i=1}^{m} 1_{f(x_i)=+1}$$

Since the points are drawn i.i.d.,

$$\mathrm{E}[\widehat{p}_+] = \frac{1}{m} \sum_{i=1}^{m} \mathop{\mathrm{E}}_{S \sim D^m}[1_{f(x_i)=+1}] = \mathop{\mathrm{E}}_{S \sim D^m}[1_{f(x_1)=+1}] = \mathop{\mathrm{E}}_{x \sim D}[1_{f(x)=+1}] = p_+.$$

Thus, by Hoeffding's inequality, for any $\epsilon > 0$,

$$\Pr[|p_+ - \widehat{p}_+| > \epsilon] \leq 2e^{-2m\epsilon^2}.$$

Setting δ to match the right-hand side yields the result. □

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'?

Solution: The probability that R' misses region r_j is the product of the probability p for each point x_i of the training sample to either not fall in r_j or be positive and fall in r_j with the label flipped to negative due to noise.

$$\begin{aligned}
p &= \Pr[x \notin r_j \vee (x \in r_j \wedge x \text{ positive} \wedge \text{ label of } x \text{ flipped})] \\
&= \Pr[x \notin r_j \vee (x \in r_j \wedge \text{ label of } x \text{ flipped})] \\
&= \Pr[x \notin r_j] + \Pr[(x \in r_j \wedge \text{ label of } x \text{ flipped})] \\
&= (1 - \Pr[x \in r_j]) + \eta \Pr[x \in r_j] \\
&= (1 - \eta)(1 - \Pr[x \notin r_j]) + \eta \\
&\leq (1 - \eta)(1 - \epsilon/4) + \eta \\
&= (1 - \epsilon/4) + \eta\epsilon/4 \leq 1 - \epsilon(1 - \eta')/4.
\end{aligned}$$

□

(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.

Solution: The probability that $\Pr[error(R') > \epsilon]$ is upper bounded by the probability that R' misses at least one region r_j. Thus, by the union bound,

$$\Pr[error(R') > \epsilon] \leq 4\Big(1 - \epsilon(1 - \eta')/4\Big)^m \leq 4e^{-m\epsilon(1-\eta')/4}.$$

Setting δ to match the upper bound leads to the following: with probability at least $1 - \delta$, for $m \geq \frac{4}{(1-\eta')\epsilon} \log \frac{4}{\delta}$, $error(R') \leq \epsilon$. □

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 $d(h^*) = \eta$.

   Solution: The probability that the label of a point be incorrect is η. A label is incorrect iff it differs from the label given by the target h^*. □

(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.

Solution: The label of a point disagrees with the one given by h either because its label is correct (probability $1 - \eta$) and h misclassifies that point (probability $error(h)$), or because its label is incorrect (probability η) and h classifies it correctly (probability $1 - error(h)$). Since these two events are disjoint, the probability of their union is the sum of the probability and

$$\begin{aligned} d(h) &= (1-\eta) error(h) + \eta(1 - error(h)) \\ &= \eta + (1-2\eta)\, error(h). \end{aligned}$$

□

(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 $\epsilon' = \epsilon(1 - 2\eta')$.

Solution: In view of the previous question, if $error(h) > \epsilon$,

$$\begin{aligned} d(h) &= \eta + (1-2\eta)\, error(h) \\ &\geq \eta + (1-2\eta)\epsilon \\ &\geq \eta + (1-2\eta')\epsilon \\ &= d(h^*) + (1-2\eta')\epsilon, \end{aligned}$$

where we used $d(h^*) = \eta$. □

(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:

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

Solution: By Hoeffding's inequality $\Pr[\widehat{d}(h^*) - d(h^*) > \epsilon'/2] \leq e^{-m\epsilon'^2/2}$. Setting $\delta/2$ to match the right-hand side yields the result. □

(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 $h \in H$:

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

Solution: By the union bound and Hoeffding's inequality $\Pr[\exists h\colon d(h) - \widehat{d}(h) > \epsilon'/2] \leq |H| e^{-m\epsilon'^2/2}$. Setting $\delta/2$ to match the right-hand side yields the result. □

(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).

Solution: By the union bound, for any $\delta > 0$, with probability at least $1 - \delta$, for $m \geq \frac{2}{\epsilon'^2}(\log |H| + \log \frac{2}{\delta})$, both inequalities of the previous two questions hold, the previous one for all $h \in H$. Thus, using the equality of the hint, with probability at least $1 - \delta$, for $m \geq \frac{2}{\epsilon'^2}(\log |H| + \log \frac{2}{\delta})$, the following holds for all $h \in H$ with $error(h) > \epsilon$:

$$\begin{aligned}\widehat{d}(h) - \widehat{d}(h^*) &= [\widehat{d}(h) - d(h)] + [d(h) - d(h^*)] + [d(h^*) - \widehat{d}(h^*)] \\ &\geq -\epsilon'/2 + \epsilon' - \epsilon'/2 = 0,\end{aligned}$$

and thus such hypotheses h are not selected by L since they do not admit a minimal $\widehat{d}(h)$.

This shows that algorithm L can be used for PAC-learning in the presence of the noise described and in the consistent case where the target concept is in H. Nevertheless, the computational complexity of L is in general not polynomial. In general, the problem of finding the hypothesis with minimal $\widehat{d}(h)$ is NP-complete. □