Mehryar Mohri
Foundations of Machine Learning 2024
Courant Institute of Mathematical Sciences
Homework assignment 1
Sep 17, 2024
Due: October 01, 2024

A Probability tools

1. Let $f\colon(0,+\infty)\to\mathbb{R}_+$ be a function that admits an inverse f^{-1}, and let X be a random variable. Suppose that for all $t>0$, the probability that X exceeds t is bounded by $f(t)$, i.e., $\mathbb{P}[X>t]\le f(t)$. Prove that for any $\delta>0$, with probability at least $1-\delta$, the random variable X satisfies $X\le f^{-1}(\delta)$.

2. Let Z be a discrete random variable that takes non-negative integer values. Prove that $\mathbb{E}[Z]=\sum_{n\ge1}\mathbb{P}[Z\ge n]$. *Hint: express* $\mathbb{P}[Z=n]$ *as* $\mathbb{P}[Z\ge n]-\mathbb{P}[Z\ge n+1]$.

Solution:

1. For any $\delta>0$, let $t=f^{-1}(\delta)$. Plugging this in $\mathbb{P}[X>t]\le f(t)$ yields $\mathbb{P}[X>f^{-1}(\delta)]\le\delta$, that is $\mathbb{P}[X\le f^{-1}(\delta)]\ge1-\delta$.

2. We assume that Z 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
$$\mathbb{E}[Z]=\sum_{n\ge0}n\,\mathbb{P}[Z=n]=\sum_{n\ge1}n(\mathbb{P}[Z\ge n]-\mathbb{P}[Z\ge n+1]).$$
   Note that in this sum, for $n\ge1$, $\mathbb{P}[Z\ge n]$ is added n times and subtracted $n-1$ times, thus $\mathbb{E}[Z]=\sum_{n\ge1}\mathbb{P}[Z\ge n]$.

   More generally, by definition of the Lebesgue integral, for any non-negative random variable Z, the following identity holds:
$$\mathbb{E}[Z]=\int_0^{+\infty}\mathbb{P}[Z\ge t]\,dt.$$

B Label bias

1. Let $\mathcal{D}$ be a distribution over $\mathcal{X}$, and let $f\colon\mathcal{X}\to\{-1,+1\}$ be a labeling function. Our goal is to approximate the label bias of the distribution $\mathcal{D}$, denoted by p_+, which is defined as:
$$p_+=\mathbb{P}_{x\sim\mathcal{D}}[f(x)=+1].$$
   Let S be a labeled sample of size m, drawn i.i.d. from $\mathcal{D}$. Using S, derive an estimate $\hat{p}_+$ of p_+. Show that for any $\delta>0$, with probability at least $1-\delta$, the following inequality holds:
$$|p_+-\hat{p}_+|\le\sqrt{\frac{\log(2/\delta)}{2m}}.$$
   Justify each step of the proof carefully.

Solution:

1. 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.,

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

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

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

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 $\mathbb{P}[\mathrm{R}] > \epsilon$. Suppose that $R(\mathrm{R}') > \epsilon$. 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 $\mathbb{P}[R(\mathrm{R}') > \epsilon]$ in terms of ϵ and η' and conclude by giving a sample complexity bound.

Solution:

1. (a) 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.

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

   (b) The probability that $\mathbb{P}[R(\mathrm{R}') > \epsilon]$ is upper bounded by the probability that R' misses at least one region r_j. Thus, by the union bound,

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

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