Mehryar Mohri
Foundations of Machine Learning 2023
Courant Institute of Mathematical Sciences
Homework assignment 1
February 7, 2023
Due: February 21, 2023

A Imbalanced data

1. Let S be a training sample of size m. Let the bias of the data be $\mathbb{P}[+1]$. Give an (unbiased) estimate $\widehat{p}$ of the bias based on the sample. Show that with probability at least $1-\delta$, $|\widehat{p} - \mathbb{P}[+1]| \leq \sqrt{\frac{\log(2/\delta)}{2m}}$. How would you use your estimate to design an algorithm? For example, how would you use that to change the algorithm for learning axis-aligned rectangles?

2. Suppose we define the bias as $\mathbb{E}_S\Big[\frac{m_+}{m_- + \gamma}\Big]$, where m_+ is the number of positive examples and m_- is the number of negative examples. Can you give an estimate of this quantity and show that it is close to it with high probability? [hint: it might be useful to use $\mathbb{E}[1/(1+\text{Binomial}(m,p))] \leq 1/((m+1)p)$, which you would have to prove first.]

Solution:

1. Let $S_m = \sum_{k=1}^m X_k$. An unbiased estimate is $\widehat{p} = \frac{S_m}{m}$. By Hoeffding's inequality,

$$\mathbb{P}\left[|\widehat{p} - \mathbb{P}[+1]| > \sqrt{\frac{\log(2/\delta)}{2m}}\right] = \mathbb{P}\left[|S_m - \mathbb{E}[S_m]| > \sqrt{\frac{m\log(2/\delta)}{2}}\right] \leq 2\exp\left(-\frac{2\frac{m\log(2/\delta)}{2}}{m}\right) = \delta.$$

Therefore, with probability at least $1-\delta$, $|\widehat{p} - \mathbb{P}[+1]| \leq \sqrt{\frac{\log(2/\delta)}{2m}}$.

Use the estimate to design an algorithm for learning axis-aligned rectangles: if the estimated value $\widehat{p}$ is greater than $\frac{1}{2}$, the algorithm may return the largest axis-aligned rectangle that does not contain the points labeled with -1, in order to minimize the number of false negatives. Conversely, if the estimated value $\widehat{p}$ is less than or equal to $\frac{1}{2}$, the algorithm may return the tightest axis-aligned rectangle that contains the points labeled with $+1$, in order to minimize the number of false positives.

2. To simplify, let us assume that $\gamma = m$. The same proof can be applied to other values of $\gamma > 0$ as well. Let $f(x_1, \ldots, x_m) = \frac{\sum_{k=1}^m x_k}{2m - \sum_{k=1}^m x_k}$ for any points $x_1, \ldots, x_m \in \mathcal{X}$. An unbiased estimate is $f(S) = \frac{\sum_{k=1}^m X_k}{2m - \sum_{k=1}^m X_k}$. Note that for any $i \in [m]$ and any points $x_1, \ldots, x_m, x_i' \in \mathcal{X}$,

$$\begin{aligned}
&|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \\
&= \left|\frac{\sum_{k=1}^m x_k}{2m - \sum_{k=1}^m x_k} - \frac{\sum_{k=1}^m x_k + x_i' - x_i}{2m - \sum_{k=1}^m x_k + x_i - x_i'}\right| \\
&= \left|\frac{(\sum_{k=1}^m x_k)(2m - \sum_{k=1}^m x_k + x_i - x_i') - (2m - \sum_{k=1}^m x_k)(\sum_{k=1}^m x_k + x_i' - x_i)}{(2m - \sum_{k=1}^m x_k)(2m - \sum_{k=1}^m x_k + x_i - x_i')}\right| \\
&= \left|\frac{2m(x_i - x_i')}{(2m - \sum_{k=1}^m x_k)(2m - \sum_{k=1}^m x_k + x_i - x_i')}\right| \\
&\leq \frac{2}{m}.
\end{aligned}$$

By McDiarmid's inequality,

$$\mathbb{P}\left[|f(S) - \mathbb{E}[f(S)]| > \sqrt{\frac{2\log(2/\delta)}{m}}\right] \leq 2\exp\left(-\frac{\frac{4\log(2/\delta)}{m}}{\frac{4}{m}}\right) = \delta.$$

Therefore, with probability at least $1-\delta$, $|f(S) - \mathbb{E}[f(S)]| \leq \sqrt{\frac{2\log(2/\delta)}{m}}$.

B PAC learning

1. Show that the concept class of the union of two intervals in $\mathbb{R}$ is PAC learnable. Give a rigorous description of the algorithm and the proof.

2. The proof of the theorem given in class for a finite hypothesis set in the consistent case is not sufficiently explicit. What we want to prove is: $\mathbb{P}[\widehat{R}_S(h_S) = 0 \Rightarrow R(h_S) \leq \epsilon] \geq 1 - \delta$. Prove that that is equivalent to $\mathbb{P}[\widehat{R}_S(h_S) = 0 \wedge h_S \in \mathcal{H}_\epsilon] \leq \delta$. Explain why we then bound $\mathbb{P}[\exists h \in \mathcal{H}_\epsilon \colon \widehat{R}_S(h) = 0]$.

3. Suppose we have a sequence of distributions $\mathcal{D}_1, \ldots, \mathcal{D}_t, \ldots$. Let S be a sample of m independently drawn points with $x_i \sim \mathcal{D}_i$. We are in a deterministic setting where $y_i = f(x_i)$ for some function f. Let $\mathcal{H}$ be a finite hypothesis set and let ℓ be a loss function taking values in $[0,1]$, $\ell(h(x_i), y_i) \in [0,1]$. The loss function ℓ is definite, that is $\ell(y, y') = 0$ iff $y = y'$. Show that: $\mathbb{P}[\exists h \in \mathcal{H} \colon \mathbb{E}_{i \sim \mathrm{Unif}\{1,\ldots,m\}, x \sim \mathcal{D}_i}[\ell(h(x), y)] > \epsilon \wedge \mathbb{E}_{x \sim S}[\ell(h(x), y)] = 0] \leq |\mathcal{H}| e^{-m\epsilon}$.

Solution:

1. Consider the concept class formed by unions of two closed intervals $[a,b] \bigcup [c,d]$. We can define a simple PAC-learning algorithm as follows. For a training sample S, the algorithm returns the hypothesis h_S:

   - if there are two separate sequences of positively labeled points in the training data (separated by negative points), then return the union of two intervals $\overline{[a,b]} \bigcup \overline{[c,d]}$ with $\overline{[a,b]} \subset [a,b]$ and $\overline{[c,d]} \subset [c,d]$, where $\overline{[a,b]}$ is the smallest interval containing the first sequence of positive points, and $\overline{[c,d]}$ is the smallest interval containing the second sequence of positive points.
   - Otherwise, return the smallest interval $\overline{[a,d]}$ containing all the positive points, which can be written as the union of two closed intervals.

   Let $[a,b] \bigcup [c,d]$ be the target concept. Let $\epsilon > 0$. We can assume that $\mathbb{P}[[a,b]] > \epsilon/3$ and $\mathbb{P}[[c,d]] > \epsilon/3$. Other cases are either trivial or simple to analyze as for what follows. As in the proof for axis-aligned rectangles, consider four regions r_1, r_2, r_3 and r_4 defined as follows. r_1 is an interval of the form $[a, \bar{b}]$, $\bar{b} \leq b$ such that $\mathbb{P}\big[[a, \bar{b}]\big] > \epsilon/6$. Similarly, r_2, r_3 and r_4 are regions bordering the endpoints of the two intervals, each with probability $\epsilon/6$.

   Now, by the algorithm's definition and a geometric argument similar to the case of axis-aligned rectangles, if $R(h_S) > \epsilon$, then either the union of intervals predicted misses at least one of the regions r_i, $i \in [1,4]$, or $\mathbb{P}[(b,c)] > \epsilon/3$ and no training point falls in (b,c) (second case of the hypothesis returned by the algorithm). Thus, using the union bound and considering the probability of each point falling outside the (b,c) when $\mathbb{P}[(b,c)] > \epsilon/3$ is at most $(1 - \epsilon/3)$, we have:

   $$\begin{aligned}
   \underset{S \sim \mathcal{D}^m}{\mathbb{P}}[R(h_S) > \epsilon] &\leq \underset{S \sim \mathcal{D}^m}{\mathbb{P}}\left[\bigcup_{i=1}^{4} \{h_S \bigcap r_i = \varnothing\}\right] + (1 - \epsilon/3)^m \\
   &\leq \sum_{i=1}^{4} \underset{S \sim \mathcal{D}^m}{\mathbb{P}}[\{h_S \bigcap r_i = \varnothing\}] + (1 - \epsilon/3)^m \\
   &\leq 4(1 - \epsilon/6)^m + (1 - \epsilon/3)^m \\
   &\leq 4e^{-m\epsilon/6} + e^{-m\epsilon/3} \\
   &\leq 5e^{-m\epsilon/6}.
   \end{aligned}$$

   Setting $\delta > 0$ to match the upper bound yields that for $m \geq \frac{6}{\epsilon} \log \frac{5}{\delta}$, with probability at least $1 - \delta$, $R(h_S) \leq \epsilon$.

2. We can demonstrate the equivalence as follows:

   $$\begin{aligned}
   \mathbb{P}[\widehat{R}_S(h_S) = 0 \Rightarrow R(h_S) \leq \epsilon] \geq 1 - \delta &\iff \mathbb{P}[\widehat{R}_S(h_S) \neq 0 \vee R(h_S) \leq \epsilon] \geq 1 - \delta \\
   &\iff \mathbb{P}[\widehat{R}_S(h_S) = 0 \wedge R(h_S) > \epsilon] \leq \delta \\
   &\iff \mathbb{P}[\widehat{R}_S(h_S) = 0 \wedge h_S \in \mathcal{H}_\epsilon] \leq \delta.
   \end{aligned}$$

However, since we do not know which consistent hypothesis $h_S \in \mathcal{H}_\epsilon$ the algorithm will select, and this choice depends on the training sample S, we need to provide a uniform convergence bound. In other words, we require a bound that holds for the set of all consistent hypotheses in $\mathcal{H}_\epsilon$. Therefore, we will bound $\mathbb{P}[\exists h \in \mathcal{H}_\epsilon\colon \widehat{R}_S(h) = 0]$, which provides an upper bound of $\mathbb{P}[\widehat{R}_S(h_S) = 0 \wedge h_S \in \mathcal{H}_\epsilon]$.

3. For any $\epsilon > 0$, define $\mathcal{H}_\epsilon$ by $\mathcal{H}_\epsilon = \left\{h \in \mathcal{H} : \mathbb{E}_{i\sim \mathrm{Unif}\{1,\ldots,m\}, x\sim \mathcal{D}_i}[\ell(h(x), y)] > \epsilon\right\}$. Since $\ell \in [0,1]$ is definite, for any i, we have $\mathbb{E}_{x\sim\mathcal{D}_i}[\ell(h(x),y)] \le \mathbb{E}_{x\sim\mathcal{D}_i}\left[1_{h(x)\neq y}\right] = \mathbb{P}_{x\sim\mathcal{D}_i}[h(x) \neq y]$. Thus, by the union bound, the following holds:

$$
\begin{aligned}
&\mathbb{P}\left[\exists h \in \mathcal{H}\colon \mathop{\mathbb{E}}_{i\sim \mathrm{Unif}\{1,\ldots,m\}, x\sim \mathcal{D}_i}[\ell(h(x), y)] > \epsilon \wedge \mathop{\mathbb{E}}_{x\sim S}[\ell(h(x),y)] = 0\right] \\
&= \mathbb{P}\left[\bigcup_{h\in\mathcal{H}_\epsilon}\left\{\mathop{\mathbb{E}}_{x\sim S}[\ell(h(x),y)] = 0\right\}\right] \\
&\le \sum_{h\in\mathcal{H}_\epsilon} \mathbb{P}\left[\mathop{\mathbb{E}}_{x\sim S}[\ell(h(x),y)] = 0\right] && \text{(union bound)} \\
&= \sum_{h\in\mathcal{H}_\epsilon} \prod_{i=1}^{m} \mathop{\mathbb{P}}_{x_i\sim\mathcal{D}_i}[h(x_i) = y_i] && (\ell \text{ is definite}) \\
&= \sum_{h\in\mathcal{H}_\epsilon} \prod_{i=1}^{m} \left(1 - \mathop{\mathbb{P}}_{x_i\sim\mathcal{D}_i}[h(x_i) \neq y_i]\right) \\
&\le \sum_{h\in\mathcal{H}_\epsilon} \left(1 - \frac{\sum_{i=1}^m \mathbb{P}_{x_i\sim\mathcal{D}_i}[h(x_i)\neq y_i]}{m}\right)^m && \text{(AM–GM inequality)} \\
&\le \sum_{h\in\mathcal{H}_\epsilon} \left(1 - \frac{\sum_{i=1}^m \mathbb{E}_{x_i\sim\mathcal{D}_i}[\ell(h(x_i), y_i)]}{m}\right)^m && (\mathbb{E}_{x\sim\mathcal{D}_i}[\ell(h(x),y)] \le \mathbb{P}_{x\sim\mathcal{D}_i}[h(x)\neq y]) \\
&= \sum_{h\in\mathcal{H}_\epsilon} \left(1 - \mathop{\mathbb{E}}_{i\sim \mathrm{Unif}\{1,\ldots,m\}, x\sim \mathcal{D}_i}[\ell(h(x), y)]\right)^m \\
&\le |\mathcal{H}|(1-\epsilon)^m \\
&\le |\mathcal{H}| e^{-m\epsilon},
\end{aligned}
$$

where for the last step we used the general inequality $1 - x \le e^{-x}$ valid for all $x \in \mathbb{R}$.

C Bayes classifier

In this problem, we consider the multi-class classification setting where $\mathcal{Y} = \{1, \ldots, k\}$. Given a hypothesis set $\mathcal{H}$ of functions mapping from $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, we define the margin as $\rho_h(x,y) = h(x,y) - \max_{y'\neq y} h(x,y')$. Given a distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$, the Bayes error for a loss function $\ell(h,x,y)$ is defined as the infimum of the errors achieved by measurable functions $h\colon \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$:

$$R_\ell^* = \inf_{h\colon\mathcal{X}\times\mathcal{Y}\to\mathbb{R} \text{ measurable}} R_\ell(h),$$

where $R_\ell(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h,x,y)]$. A hypothesis h^* with $R_\ell(h^*) = R_\ell^*$ is called a Bayes classifier. Denote by $p(x,y) = \mathcal{D}(Y = y \mid X = x)$ the conditional probability of $Y = y$ given $X = x$.

1. For a labeled example (x,y), the multi-class zero-one loss is defined by $\ell_{0-1}(h,x,y) = 1_{\rho_h(x,y)\le 0}$. Derive the Bayes classifier and Bayes error for ℓ_{0-1}.

2. For a labeled example (x,y), the multinomial logistic loss is defined by $\ell_{\log}(h,x,y) = -\log\left(\frac{e^{h(x,y)}}{\sum_{y'\in\mathcal{Y}} e^{h(x,y')}}\right)$. Derive the Bayes classifier and Bayes error for $\ell_{\log}$.

Solution:

1. By definition, for any hypothesis h,

$$R_{\ell_{0-1}}(h) = \mathbb{E}_X\left[\sum_{y \in \mathcal{Y}} p(x,y) 1_{\rho_h(x,y) \le 0}\right] \ge \mathbb{E}_X\left[1 - \max_{y \in \mathcal{Y}} p(x,y)\right]$$

where the equality holds if and only if h satisfies

$$\forall x \in \mathcal{X}, \quad \operatorname*{argmax}_{y \in \mathcal{Y}} h(x,y) \subset \operatorname*{argmax}_{y \in \mathcal{Y}} p(x,y), \quad \max_{y \in \mathcal{Y}} \rho_h(x,y) > 0. \tag{1}$$

Therefore, the Bayes classifier for ℓ_{0-1} is defined by (1) and the Bayes error is $\mathbb{E}_X[1 - \max_{y \in \mathcal{Y}} p(x,y)]$.

2. Let $s(x,y) = \frac{e^{h(x,y)}}{\sum_{y' \in \mathcal{Y}} e^{h(x,y')}}$. Using the fact that $\sum_{y \in \mathcal{Y}} s(x,y) = 1$ and the method of Lagrange multipliers, we have that for any hypothesis h,

$$R_{\ell_{\log}}(h) = \mathbb{E}_X\left[-\sum_{y \in \mathcal{Y}} p(x,y) \log(s(x,y))\right] \ge \mathbb{E}_X\left[-\sum_{y \in \mathcal{Y}} p(x,y) \log(p(x,y))\right].$$

where the equality holds if and only if h satisfies

$$\forall (x,y) \in \mathcal{X} \times \mathcal{Y}, \quad s(x,y) = \frac{e^{h(x,y)}}{\sum_{y' \in \mathcal{Y}} e^{h(x,y')}} = p(x,y). \tag{2}$$

Therefore, the Bayes classifier is defined by (2) and the Bayes error is $\mathbb{E}_X\left[-\sum_{y \in \mathcal{Y}} p(x,y) \log(p(x,y))\right]$.