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 $\mathbb{E}_S[m_+/m_-]$, 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.]

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: $\mathbb{P}[\exists h \in \mathcal{H} \colon \mathbb{E}_{i \sim \text{Unif}\{1,\ldots,m\}, x \sim \mathcal{D}_i}[\ell(h(x), y)] > \epsilon \wedge \mathbb{E}_{x \sim S}[\ell(h(x), y)] = 0] \leq |\mathcal{H}| e^{-m\epsilon}$.

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 $\ell_{0-1}(h,x,y) = 1_{\rho_h(x,y) \leq 0}$. 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 $\ell_{\log}(h,x,y) = -\log\Big(\frac{h(x,y)}{\sum_{y' \in \mathcal{Y}} h(x,y')}\Big)$. Derive the Bayes classifier and Bayes error for $\ell_{\log}$.