CSCI-GA.2566-001
Foundations of Machine Learning

Problem Set 1 Aditya Ramesh

Exercise 1.1. Let $X = \mathbb{R}^2$. Give a PAC-learning algorithm for C, the set of concepts of the form $c = \{(x_1, x_2) \in \mathbb{R}^2 : r_1^2 \leq x_1^2 + x_2^2 \leq r_2^2\}$, for real numbers $r_1, r_2 > 0$.

Solution. Let D be the distribution over the input space X, and let $S \sim D^m$ be a sample of points discriminated by the target annulus $A \in C$. Our learning algorithm returns the annulus $A_S \in H$ with inner radius s_1 and outer radius s_2. The radii s_1 and s_2 describe the smallest and largest circles, respectively, that each contain positively-labeled points in S. That is,

$$s_1 = \inf\{r : \bar{B}(0, r) \cap S \neq \varnothing\} \quad \text{and} \quad s_2 = \sup\{r : \bar{B}(0, r) \cap S \neq \varnothing\},$$

where $\bar{B}(0, r)$ is the closed ball of center 0 and radius r, and S is the set of positively-labeled points in S. If we assume that S is nonempty, then the set $\{r : \bar{B}(0, r) \cap S \neq \varnothing\}$ is also nonempty and uniformly bounded. It follows that s_1 and s_2 exist.

Now let $\epsilon, \delta > 0$ be given. It is clear that our learning algorithm runs in polynomial time with respect to $1/\epsilon$ and $1/\delta$. Note also that the annulus A_S returned by our algorithm is contained in the target annulus A. If $\Pr[A] \leq \epsilon$, then because $A_S \subseteq A$, the only kind of errors that A_S can make are false negatives. Further, the region of error is precisely $A \setminus A_S$. Therefore $\Pr[R(A_S) \leq \epsilon] = 1$.

Now assume that $\Pr[A] > \epsilon$. We construct two annuli A_1 and A_2, each contained in A and having probability mass $\epsilon/2$. The inner radius of A_1 is r_1, and the outer radius of A_1 is given by solving

$$\pi(a_1^2 - r_1^2) = \frac{\epsilon}{2}$$

for a_1. Similarly, the outer radius of A_2 is r_2, and the inner radius a_2 is obtained by solving

$$\pi(r_2^2 - a_2^2) = \frac{\epsilon}{2}$$

for a_2. Observe that if A_S intersects A_1 and A_2, then the region of error is contained in $A_1 \cup A_2$, and therefore bounded by the probability mass of this union. So if A_S intersects A_1 and A_2, then the region of error has probability mass at most ϵ. By contraposition, if the region of error has probability mass more than ϵ, then A_S does not intersect either A_1 or A_2. Formally,

$$\begin{aligned}
\Pr[R(A_S) > \epsilon] &\leq \Pr[\cup_i \{A_S \cap A_i = \varnothing] \\
&\leq \sum_i \Pr[\{A_S \cap A_i = \varnothing\}] && \text{(union bound)} \\
&= \sum_i \Pr_{\substack{S \sim D^m \\ x \in S}}[x \notin A_S \cap A_i] && \text{(expanding definition of event)} \\
&= \sum_i (\Pr_{x \sim D}[x \notin A_S \cap A_i])^m. && \text{(i.i.d. sampling)}
\end{aligned}$$

Now consider the quantity $\Pr_{x \sim D}[x \notin A_S \cap A_i]$. By the chain rule, we can bound this quantity as follows. (Note that we have dropped the subscripts $x \sim D$ for legibility.)

$$\begin{aligned}
\Pr[x \notin A_S \cap A_i] &= \Pr[x \notin A_S \cap A_i \,|\, x \in A] \Pr[x \in A] + \\
&\qquad \Pr[x \notin A_S \cap A_i \,|\, x \notin A] \Pr[x \notin A] \\
&\leq \left(1 - \frac{\epsilon/2}{\Pr[x \in A]}\right) \Pr[x \in A] + 1 \cdot (1 - \Pr[x \in A]) \\
&= 1 - \frac{\epsilon}{2}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\Pr[R(A_S) > \epsilon] &\leq \sum_i (\Pr_{x \sim D}[x \notin A_S \cap A_i])^m \\
&\leq \sum_i \left(1 - \frac{\epsilon}{2}\right)^m \\
&\leq 2 \exp\left(\frac{-m\epsilon}{2}\right). \qquad (1)
\end{aligned}$$

To find the necessary sample size m to achieve the desired δ probability bound, we set δ equal to the RHS of (1) and solve for m, giving

$$m = \frac{2}{\epsilon} \log\left(\frac{2}{\delta}\right).$$

Then for all $\epsilon, \delta > 0$, there exists $m = \textit{poly}(1/\epsilon, 1/\delta)$ such that

$$\Pr[R(A_S) \leq \epsilon] \geq 1 - \delta.$$

Hence our algorithm is a PAC-learning algorithm for C, and we conclude that C is PAC-learnable.

Exercise 1.2. Let $S = (x_1, \ldots, x_m)$ be a sample of size m, and fix $h : X \to R$.

(a) Denote by $u = (h(x_1), \ldots, h(x_m))^t$ the vector of predictions of h for S. Give an upper bound on the empirical Rademacher complexity $\hat{\mathfrak{R}}_S(H)$ of $H = \{h, -h\}$ in terms of $\|u\|_2$. Suppose that $h(x_i) \in \{0, -1, +1\}$ for all $i \in [1, m]$. Express the bound on the Rademacher complexity in terms of the sparsity measure $n = |\{i : h(x_i) \neq 0\}|$. What is the upper bound for the extreme values of this sparsity measure?

(b) Let F be the family of functions from X to $\mathbb{R}$. Give an upper bound on the empirical Rademacher complexity of $F + h$ and $F \pm h$ in terms of $\hat{\mathfrak{R}}_X(F)$ and $\|u\|_2$.

Solution.

(a) Let $\sigma = (\sigma_1, \ldots, \sigma_m)^t$ be a vector of Rademacher variables. By $\sigma \cdot u$, we denote the inner product between the vectors σ and u. We have

$$\begin{aligned}
\hat{\mathfrak{R}}_S(H) &= \mathrm{E}_\sigma\left[\sup_{h \in H} \frac{\sigma \cdot u}{m}\right] && \text{(by definition)} \\
&= \mathrm{E}_\sigma\left[\max\left\{\frac{\sigma \cdot u}{m}, -\frac{\sigma \cdot u}{m}\right\}\right] \\
&= \mathrm{E}_\sigma\left[\left|\frac{\sigma \cdot u}{m}\right|\right] \\
&= \mathrm{E}_\sigma\left[\frac{\|u\|_1}{m}\right] \\
&= \frac{1}{m}(\mathrm{E}_\sigma[|\sigma \cdot u|]^2)^{1/2} \\
&\leq \frac{1}{m}(\mathrm{E}_\sigma[(\sigma \cdot u)^2])^{1/2} && \text{(Jensen's inequality)} \\
&= \frac{1}{m}\left(\mathrm{E}_\sigma\left[\sum_{i,j \in [1,m]} \sigma_i \sigma_j u_i u_j\right]\right)^{1/2} \\
&\leq \frac{1}{m}\left(\sum_{i,j \in [1,m]} \mathrm{E}_\sigma[\sigma_i \sigma_j u_i u_j]\right)^{1/2}.
\end{aligned}$$

Note that if $i \neq j$, then

$$\mathrm{E}_\sigma[\sigma_i \sigma_j u_i u_j] = \mathrm{E}_\sigma[\sigma_i u_i]\mathrm{E}_\sigma[\sigma_j u_j] = 0.$$

Therefore

$$\begin{aligned}\hat{\mathfrak{R}}_S(H) &\leq \frac{1}{m}\left(\sum_{i,j\in[1,m]} E_\sigma[\sigma_i\sigma_j u_i u_j]\right)^{1/2} \\ &= \frac{1}{m}\left(\sum_{i=1}^{m} E_{\sigma_i}[\sigma_i^2 u_i^2]\right)^{1/2} \\ &= \frac{1}{m}\left(\sum_{i=1}^{m} u_i^2\right)^{1/2} \\ &= \frac{\|u\|_2}{m}.\end{aligned}$$

Unlike the upper bound derived in the original submission for this assignment, this one still holds under the assumption that $h(x_i) \in \{0, -1, +1\}$. Therefore, we can immediately conclude that

$$\hat{\mathfrak{R}}_S(H) \leq \frac{\sqrt{n}}{m}.$$

If $n = 1$, then $\hat{\mathfrak{R}}_S(H) \leq 1/m$. If $n = m$, then $\hat{\mathfrak{R}}_S(H) \leq 1/\sqrt{m}$.

(b) We have

$$\begin{aligned}
\hat{\mathfrak{R}}_S(F+h) &= \mathrm{E}_\sigma\left[\sup_{g\in F+h}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(x_i)\right] \\
&= \mathrm{E}_\sigma\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^{m}\sigma_i (f+h)(x_i)\right] \\
&= \mathrm{E}_\sigma\left[\sup_{f\in F}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(x_i)\right)+\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i)\right] \\
&= \mathrm{E}_\sigma\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(x_i)\right]+\mathrm{E}_\sigma\left[\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i)\right] \\
&= \hat{\mathfrak{R}}_S(F)+0=\hat{\mathfrak{R}}_S(F).
\end{aligned}$$

Letting $H=\{-h,h\}$ as in part (a), we have

$$\begin{aligned}
\hat{\mathfrak{R}}_S(F\pm h) &= \mathrm{E}_\sigma\left[\sup_{g\in F\pm h}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g(x_i)\right] \\
&= \mathrm{E}_\sigma\left[\sup_{\substack{f\in F\\ h\in H}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i (f+h)(x_i)\right] \\
&= \mathrm{E}_\sigma\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(x_i)+\sup_{h\in H}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i)\right] \\
&= \mathrm{E}_\sigma\left[\sup_{f\in F}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(x_i)\right]+\mathrm{E}_\sigma\left[\sup_{h\in H}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(x_i)\right] \\
&\le \hat{\mathfrak{R}}_S(F)+\frac{\|u\|_2}{\sqrt{m}}.
\end{aligned}$$

Exercise 1.3. Let $H_1, \ldots, H_p$ be families of hypotheses from X to $\{0,1\}$, and let $f : \{0,1\}^p \to \{0,1\}$. Let F be the family of functions

$$F = \{x \mapsto f(h_1(x), \ldots, h_p(x)) : h_i \in H_i\}.$$

(a) Show that the following inequality holds for the growth function F: for all $m \geq 1$, we have

$$\Pi_F(m) \leq \prod_{k=1}^{p} \Pi_{H_k}(m).$$

(b) Use appropriate choices of f to bound the growth function of the following sets:

$$H_\cup = \{\cup_k \text{supp}(h_k) : h_k \in H_k\}$$
$$H_\cap = \{\cap_k \text{supp}(h_k) : h_k \in H_k\}$$
$$H_\Delta = \{\text{supp}(h_1) \,\Delta\, \text{supp}(h_2) : h_1 \in H_1, h_2 \in H_2\}.$$

Solution.

(a) Let $S = (x_1, \ldots, x_m)$. Then for all $m \geq 1$, we have

$$\begin{aligned}\Pi_F(m) &= \max_{S \subseteq X} |\{(f(x_1), \ldots, f(x_m)) : f \in F\}| \\ &= \max_{S \subseteq X} |\{(f(h_1(x_1), \ldots, h_p(x_1)), \ldots, \\ &\qquad f(h_1(x_m), \ldots, h_p(x_m))) : f \in F, h_i \in H_i\}|\end{aligned}$$

Consider the maximum number of distinct ways in which the points in S can be classified by hypotheses in F due to changes in the first coordinate $h_1(x_i)$ of each element $f(x_i)$. This quantity is bounded above by the maximum number of distinct ways in which H_1 can classify S. That is,

$$\begin{aligned}\Pi_F(m) &= \max_{S \subseteq X} |\{(f(h_1(x_1), \ldots, h_p(x_1)), \ldots, \\ &\qquad f(h_1(x_m), \ldots, h_p(x_m))) : f \in F, h_i \in H_i\}| \\ &\leq \max_{S \subseteq X} |\{(h_1(x_1), \ldots, h_1(x_m)) : h_1 \in H_1\}| \cdot \\ &\qquad \max_{\substack{S \subseteq X \\ c \in \{0,1\}}} |\{(f(c, \ldots, h_p(x_1)), \ldots, f(c, \ldots, h_p(x_m))) : \\ &\qquad f \in F, h_i \in H_i\}| \\ &\leq \max_{S \subseteq X} |\{(h_1(x_1), \ldots, h_1(x_m)) : h_1 \in H_1\}| \cdot \cdots \cdot \\ &\qquad \max_{S \subseteq X} |\{(h_p(x_1), \ldots, h_p(x_m)) : h_p \in H_p\}| \\ &= \prod_{k=1}^{p} \Pi_{H_k}(m),\end{aligned}$$

as desired.

(b) First consider the hypothesis set $H_\cup$, and fix $x \in X$. Note that $h \in H_\cup$ with $h(x) = 1$ if and only if $h_k(x) = 1$ for some $k \in [1, p]$. Therefore $h(x) = \max_k h_k(x)$. Setting $f : \{0,1\}^p \to \{0,1\}$ to the max function and applying part (a) shows that $H_\cup$ is bounded above. Similarly, $h \in H_\cap$ with $h(x) = 1$ if and only if $h_k(x) = 1$ for all $k \in [1, p]$. Therefore $h(x) = \min_k h(k)$. Setting f to the min function and applying part (a) shows that $H_\cap$, too, is bounded above.

Now consider the hypothesis set H_Δ. Observe that for any two sets A and B, we have

$$\begin{aligned} A \,\Delta\, B &= (A \setminus B) \cup (B \setminus A) \\ &= (A \cap B^c) \cup (B \cap A^c) \\ &= ((A \cap B^c) \cup B) \cup ((A \cap B^c) \cup A^c) \\ &= ((A \cup B) \cap (B^c \cup B)) \cup ((A \cup A^c) \cap (B^c \cup A^c)) \\ &= (A \cup B) \cup (B^c \cup A^c) \\ &= (A \cup B) \cap (A \cap B)^c \\ &= (A \cup B) \setminus (A \cap B). \end{aligned}$$

It follows that for any $x \in X$, we have $h \in H_\Delta$ with $h(x) = 1$ if and only if $x \in \text{supp}(H_1)$ and $x \in \text{supp}(H_2)$, but not both. This behavior is captured precisely by the XOR function. So setting $f : \{0,1\}^2 \to \{0,1\}$ to the XOR function and applying part (a) shows that H_Δ is bounded above.