Mehryar Mohri
Foundations of Machine Learning 2014
Courant Institute of Mathematical Sciences
Homework assignment 1
September 17, 2014
Due: September 30, 2014

A. PAC learning of n-dimensional rectangles

Give a PAC-learning algorithm for C, the set of axis-aligned n-dimensional rectangles in $\mathbb{R}^n$, that is $C = \{[a_1, b_1] \times \cdots \times [a_n, b_n] \colon a_i, b_i \in \mathbb{R}\}$. You should give a careful proof similar to what was given in class for axis-aligned rectangles (case $n = 2$). How does the sample complexity vary as a function of n?

Solution: We let R' be the smallest n-dimensional rectangle consistent with the given sample. The proof is similar to the one given in class for rectangles in the plane except that here we need to consider $2n$ regions r_i, $i \in [1, 2n]$, along each face of the target n-dimensional rectangle with $\Pr[r_i] \geq \frac{\epsilon}{2n}$ and with $\Pr[r_i - f_i] < \frac{\epsilon}{2n}$ where f_i is the internal face of r_i. Arguing as in the proof given in class, assuming that $\Pr[\mathsf{R}] > \epsilon$, if $\Pr[R(\mathsf{R}') > \epsilon]$ then R' must miss at least one region r_i. The probability that none of the m sample points falls into region r_i is $(1 - \epsilon/2n)^m$. By the union bound, this shows that

$$\Pr[R(\mathsf{R}') > \epsilon] \leq 2n(1 - \epsilon/2n)^m \leq 2n e^{-\frac{\epsilon m}{2n}}. \tag{1}$$

Setting δ to the right-hand side shows that for

$$m \geq \frac{2n}{\epsilon} \log \frac{2n}{\delta}, \tag{2}$$

with probability at least $1 - \delta$, $R(\mathsf{R}') \leq \epsilon$.

B. Rademacher complexity of regularized neural networks

Let the input space be $X = \mathbb{R}^{n_1}$. In this problem, we consider the family of regularized neural networks defined by the following set of functions mapping X to $\mathbb{R}$:

$$\mathcal{H} = \left\{ \mathbf{x} \mapsto \sum_{j=1}^{n_2} w_j \sigma(\mathbf{u}_j \cdot \mathbf{x}) \colon \|\mathbf{w}\|_1 \leq \Lambda', \|\mathbf{u}_j\|_2 \leq \Lambda, \forall j \in [1, n_2] \right\},$$

where σ is an L-Lipschitz function. As an example, σ could be the sigmoid function which is 1-Lipschitz.

1. Show that $\widehat{\mathfrak{R}}_S(\mathcal{H}) = \frac{\Lambda'}{m} \operatorname{E}_{\boldsymbol{\sigma}} \Big[\sup_{\|\mathbf{u}\|_2 \leq \Lambda} \big|\sum_{i=1}^m \sigma_i \sigma(\mathbf{u} \cdot \mathbf{x}_i)\big|\Big]$.

   Solution:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{H}) &= \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_1 \leq \Lambda', \|\mathbf{u}_j\|_2 \leq \Lambda} \sum_{i=1}^m \sigma_i \sum_{j=1}^{n_2} w_j \sigma(\mathbf{u}_j \cdot \mathbf{x}_i)\right] \\
&= \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_1 \leq \Lambda', \|\mathbf{u}_j\|_2 \leq \Lambda} \sum_{j=1}^{n_2} w_j \sum_{i=1}^m \sigma_i \sigma(\mathbf{u}_j \cdot \mathbf{x}_i)\right] \\
&= \frac{\Lambda'}{m} \operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{u}_j\|_2 \leq \Lambda} \max_{j \in [1, n_2]} \left|\sum_{i=1}^m \sigma_i \sigma(\mathbf{u}_j \cdot \mathbf{x}_i)\right|\right] && \text{(all the weight put on largest term)} \\
&= \frac{\Lambda'}{m} \operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{u}_j\|_2 \leq \Lambda, j \in [1, n_2]} \left|\sum_{i=1}^m \sigma_i \sigma(\mathbf{u}_j \cdot \mathbf{x}_i)\right|\right] \\
&= \frac{\Lambda'}{m} \operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{u}\|_2 \leq \Lambda} \left|\sum_{i=1}^m \sigma_i \sigma(\mathbf{u} \cdot \mathbf{x}_i)\right|\right].
\end{aligned}$$

2. Use the following form of Talagrand's lemma valid for all hypothesis sets H and L-Lipschitz function Φ:

$$\frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{h \in H}\left|\sum_{i=1}^m \sigma_i (\Phi \circ h)(x_i)\right|\right] \leq \frac{L}{m} \operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{h \in H}\left|\sum_{i=1}^m \sigma_i h(x_i)\right|\right],$$

   to upper bound $\widehat{\mathfrak{R}}_S(\mathcal{H})$ in terms of the empirical Rademacher complexity of $\mathcal{H}'$, where $\mathcal{H}'$ is defined by

$$\mathcal{H}' = \{\mathbf{x} \mapsto s(\mathbf{u} \cdot \mathbf{x}) \colon \|\mathbf{u}\|_2 \leq \Lambda, s \in \{-1, +1\}\}.$$

   Solution: By Talagrand's lemma, since σ is L-Lipschitz, the following inequality holds:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{H}) &\leq \frac{\Lambda' L}{m} \operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in H}\left|\sum_{i=1}^m \sigma_i \mathbf{u}\cdot\mathbf{x}_i\right|\right]\\
&= \frac{\Lambda' L}{m} \operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in H}\sup_{s\in\{-1,+1\}} s\sum_{i=1}^m \sigma_i \mathbf{u}\cdot\mathbf{x}_i\right] \quad \text{(def. of abs. value)}\\
&= \Lambda' L\,\widehat{\mathfrak{R}}_S(\mathcal{H}').
\end{aligned}$$

3. Use the Cauchy-Schwarz inequality to show that

$$\widehat{\mathfrak{R}}_S(\mathcal{H}') = \frac{\Lambda}{m}\operatorname*{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^m \sigma_i\mathbf{x}_i\right\|_2\right].$$

Solution:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{H}') &= \frac{1}{m}\operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{u}\|_2\leq\Lambda, s\in\{-1,+1\}}\sum_{i=1}^m \sigma_i s\mathbf{u}\cdot\mathbf{x}_i\right]\\
&= \frac{1}{m}\operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{u}\|_2\leq\Lambda}\left|\sum_{i=1}^m \sigma_i\mathbf{u}\cdot\mathbf{x}_i\right|\right] \quad \text{(def. of abs. val.)}\\
&= \frac{1}{m}\operatorname*{E}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{u}\|_2\leq\Lambda}\left|\mathbf{u}\cdot\sum_{i=1}^m \sigma_i\mathbf{x}_i\right|\right]\\
&= \frac{\Lambda}{m}\operatorname*{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^m \sigma_i\mathbf{x}_i\right\|_2\right] \quad \text{(Cauchy-Schwarz eq. case).}
\end{aligned}$$

The last equality holds by setting $\mathbf{u} = \frac{\Lambda\sum_{i=1}^m \sigma_i\mathbf{x}_i}{\|\sum_{i1=1}^m \sigma_i\mathbf{x}_i\|}$.

4. Use the inequality $\mathrm{E}[\|\mathbf{X}\|_2] \leq \sqrt{\mathrm{E}[\|\mathbf{X}\|_2^2]}$, which holds by Jensen's inequality to upper bound $\widehat{\mathfrak{R}}_S(\mathcal{H}')$.

Solution:

$$
\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{H}') &= \frac{\Lambda}{m} \operatorname*{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^m \sigma_i \mathbf{x}_i\right\|_2\right] \\
&\leq \frac{\Lambda}{m}\sqrt{\operatorname*{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^m \sigma_i \mathbf{x}_i\right\|_2^2\right]} && \text{(Jensen's ineq.)} \\
&= \frac{1}{m}\sqrt{\sum_{i,j=1}^m \operatorname*{E}_{\boldsymbol{\sigma}}[\sigma_i\sigma_j]\,(\mathbf{x}_i \cdot \mathbf{x}_j)} \\
&= \frac{\Lambda}{m}\sqrt{\sum_{i,j=1}^m 1_{i=j}(\mathbf{x}_i \cdot \mathbf{x}_j)} && \text{(independence of } \sigma_i\text{s)} \\
&= \frac{\Lambda}{m}\sqrt{\sum_{i=1}^m \|\mathbf{x}_i\|_2^2}.
\end{aligned}
$$

5. Assume that for all $\mathbf{x} \in S$, $\|\mathbf{x}\|_2 \leq r$ for some $r > 0$. Use the previous questions to derive an upper bound on the Rademacher complexity of $\mathcal{H}$ in terms of r.

   Solution: In view of the previous questions,

$$
\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \Lambda' L\, \widehat{\mathfrak{R}}_S(\mathcal{H}') \leq \frac{\Lambda'\Lambda L}{m}\sqrt{\sum_{i=1}^m \|\mathbf{x}_i\|_2^2} \leq \frac{\Lambda'\Lambda L}{m}\sqrt{mr^2} = \frac{\Lambda'\Lambda L r}{\sqrt{m}}.
$$