Mehryar Mohri
Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 2
February 25, 2013
Due: March 11, 2013

**A. Rademacher complexity - properties**

Let $H$ be a hypothesis set reduced to two functions: $H = \{h_{-1}, h_{+1}\}$ and let $S = (x_1, \ldots, x_m) \subseteq \mathcal{X}$ be a sample of size $m$.

1. Assume that $h_{-1}$ is the constant function taking value $-1$ and $h_{+1}$ the constant function taking the value $+1$. What is the VC-dimension $d$ of $H$? Upper bound the empirical Rademacher complexity $\widehat{\mathfrak{R}}_S(H)$ (*hint*: express $\widehat{\mathfrak{R}}_S(H)$ in terms of the absolute value of a sum of Rademacher variables and apply Jensen's inequality) and compare your bound with $\sqrt{d/m}$.

   *Solution:* VCdim$(H) = 1$ since $H$ can shatter one point and clearly at most one. Observe that

$$\sup_{h \in H} \sum_{i=1}^{m} \sigma_i h(x_i) = \sup_{h \in H} \Big( \sum_{i=1}^{m} \sigma_i \Big) h(x_1) = \Big| \sum_{i=1}^{m} \sigma_i \Big|. \tag{1}$$

   Thus, by Jensen's inequality,

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(H) &= \frac{1}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \Big[ \Big| \sum_{i=1}^{m} \sigma_i \Big| \Big] \\
&\leq \frac{1}{m} \Big[ \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \Big[ (\sum_{i=1}^{m} \sigma_i)^2 \Big] \Big]^{1/2} \\
&= \frac{1}{m} \Big[ \mathop{\mathrm{E}}_{\boldsymbol{\sigma}} \Big[ \sum_{i=1}^{m} \sigma_i^2 \Big] \Big]^{1/2} \qquad (\mathrm{E}[\sigma_i \sigma_j] = 0 \text{ for } i \neq j) \\
&= \frac{1}{\sqrt{m}}.
\end{aligned}$$

   By the Khintchine inequality, the upper bound is tight modulo the constant $1/\sqrt{2}$. The upper bound coincides with $\sqrt{d/m}$. □

2. Assume that $h_{-1}$ is the constant function taking value $-1$ and $h_{+1}$ the function taking value $-1$ everywhere except at $x_1$ where it takes the value $+1$. What is the VC-dimension $d$ of $H$? Compute the empirical Rademacher complexity $\mathfrak{R}_S(H)$.

   *Solution:* VCdim$(H) = 1$ since $H$ can shatter $x_1$ and clearly at most one point. By definition,

$$\begin{aligned}
\mathfrak{R}_S(H) &= \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \Big[ \sup_{h \in H} \sum_{i=1}^{m} \sigma_i h(x_i) \Big] \\
&= \frac{1}{m} \operatorname*{E}_{\boldsymbol{\sigma}} \Big[ \sup_{h \in H} \sigma_1 h(x_1) - \sum_{i=2}^{m} \sigma_i \Big] \\
&= \frac{1}{m} \operatorname*{E}_{\sigma_1} \Big[ \sup_{h \in H} \sigma_1 h(x_1) \Big] && (\mathrm{E}[\sigma_i] = 0) \\
&= \frac{1}{m} \operatorname*{E}_{\sigma_1} \Big[ 1 \Big] = \frac{1}{m}.
\end{aligned}$$

   Here $\mathfrak{R}_S(H)$ is a clearly more favorable quantity than $\sqrt{d/m} = \sqrt{1/m}$.

   □

**B. Rademacher complexity bound**

Let $G$ be a family of functions mapping from $Z$ to $[0, 1]$. The general Rademacher complexity bound presented in class was based on the analysis of the function $\Phi$ defined by $\Phi(S) = \sup_{g \in G} \mathrm{E}[g] - \widehat{\mathrm{E}}_S[g]$ for any training sample $S = (z_1, \ldots, z_m)$ of size $m$, with $\widehat{\mathrm{E}}_S[g] = \frac{1}{m} \sum_{i=1}^{m} g(z_i)$. Instead, apply McDiarmid's inequality to $\Psi$ defined by $\Psi(S) = \sup_{g \in G} \mathrm{E}[g] - \widehat{\mathrm{E}}_S[g] - 2\widehat{\mathfrak{R}}_S(G)$ and try to obtain a slighty better generalization bound than the one obtained in class in terms of the empirical Rademacher complexity.

*Solution:* Let $S'$ be a sample differing from $S$ by one point, say $z_m$. Then, since a difference of suprema is upper bounded by the supremum of the differences, we can write

$$\Psi(S') - \Psi(S) = \sup_{g \in G}(\mathrm{E}[g] - \widehat{\mathrm{E}}_{S'}[g]) - \sup_{g \in G}(\mathrm{E}[g] - \widehat{\mathrm{E}}_S[g]) + \frac{2}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sup_{g \in G}\sum_{i=1}^m \sigma_i g(z_i) - \sup_{g \in G}\sum_{i=1}^m \sigma_i g(z_i')\right]$$
$$\leq \sup_{g \in G}(\mathrm{E}[g] - \widehat{\mathrm{E}}_{S'}[g]) - (\mathrm{E}[g] - \widehat{\mathrm{E}}_S[g]) + \frac{2}{m} \mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\sup_{g \in G}\sum_{i=1}^m \sigma_i g(z_i) - \sum_{i=1}^m \sigma_i g(z_i')\right]$$
$$= \sup_{g \in G} \frac{1}{m}(g(z_m) - g(z_m')) + 2 \mathop{\mathrm{E}}_{\boldsymbol{\sigma}}\left[\frac{1}{m}\sup_{g \in G} \sigma_m(g(z_m) - g(z_m'))\right] \leq \frac{3}{m}.$$

Thus, by McDiarmid's inequality, $\Pr[\Psi(S) - \mathrm{E}[\Psi(S)] > \epsilon] \leq \exp(-\frac{2}{9}m\epsilon^2)$. Thus, for any $\delta > 0$, with probability at least $1 - \delta$,

$$\forall g \in G, \Psi(S) - \mathrm{E}[\Psi(S) \leq 3\sqrt{\frac{\log \frac{1}{\delta}}{2m}}. \tag{2}$$

By definition, $\mathrm{E}[\Psi(S)] = \mathrm{E}[\Phi(S)] - 2\mathfrak{R}_m(G)$. In class, we showed that $\mathrm{E}[\Phi(S)] \leq 2\mathfrak{R}_m(G)$. Thus, with probability at least $1 - \delta$, $\Psi(S) \leq \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$, that is

$$\forall g \in G, \mathrm{E}[g] \leq \widehat{\mathrm{E}}_S[g] + 2\widehat{\mathfrak{R}}_S(G) + 3\sqrt{\frac{\log \frac{1}{\delta}}{2m}}. \tag{3}$$

**C. VC-dimension of union of $k$ intervals.**

What is the VC-dimension of subsets of the real line formed by the union of $k$ intervals?

*Solution:*

The VC-dimension of this class is $2k$. It is not hard to see that any $2k$ distinct points on the real line can be shattered using $k$ intervals; it suffices to shatter each of the $k$ pairs of consecutive points with an interval. Assume now that $2k + 1$ distinct points $x_1 < \cdots < x_{2k+1}$ are given. For any $i \in [1, 2k + 1]$, label $x_i$ with $(-1)^{i+1}$, that is alternatively label points with 1 or $-1$. This leads to $k + 1$ points labeled positively and requires $2k + 1$ intervals to shatter the set, since no interval can contain two consecutive points. Thus, no set of $2k + 1$ points can be shattered by $k$ intervals, and the VC-dimension of the union of $k$ intervals is $2k$. □

**D. Generalization bound based on covering numbers.**

Let $H$ be a family of functions mapping $\mathcal{X}$ to a subset of real numbers $\mathcal{Y} \subseteq \mathbb{R}$. For any $\epsilon > 0$, the *covering number* $\mathcal{N}(H, \epsilon)$ of $H$ for the $L_\infty$ norm is the minimal

$k \in \mathbb{N}$ such that $H$ can be covered with $k$ balls of radius $\epsilon$, that is, there exists $\{h_1, \ldots, h_k\} \subseteq H$ such that, for all $h \in H$, there exists $i \leq k$ with $\|h - h_i\|_\infty = \max_{x \in \mathcal{X}} |h(x) - h_i(x)| \leq \epsilon$. In particular, when $H$ is a compact set, a finite covering can be extracted from a covering of $H$ with balls of radius $\epsilon$ and thus $\mathcal{N}(H, \epsilon)$ is finite.

Covering numbers provide a measure of the complexity of a class of functions: the larger the covering number, the richer is the family of functions. The objective of this problem is to illustrate this by proving a learning bound in the case of the squared loss. Let $D$ denote a distribution over $\mathcal{X} \times \mathcal{Y}$ according to which labeled examples are drawn. Then, the generalization error of $h \in H$ for the squared loss is defined by $R(h) = \mathrm{E}_{(x,y)\sim D}[(h(x) - y)^2]$ and its empirical error for a labeled sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$ by $\widehat{R}(h) = \frac{1}{m} \sum_{i=1}^m (h(x_i) - y_i)^2$. We will assume that $H$ is bounded, that is there exists $M > 0$ such that $|h(x) - y| \leq M$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. The following is the generalization bound proven in this problem:

$$\Pr_{S \sim D^m} \left[ \sup_{h \in H} |R(h) - \widehat{R}(h)| \geq \epsilon \right] \leq \mathcal{N}\Big(H, \frac{\epsilon}{8M}\Big) 2 \exp\left(\frac{-m\epsilon^2}{2M^4}\right). \tag{4}$$

The proof is based on the following steps.

1. Let $L_S = R(h) - \widehat{R}(h)$, then show that for all $h_1, h_2 \in H$ and any labeled sample $S$, the following inequality holds:

$$|L_S(h_1) - L_S(h_2)| \leq 4M \|h_1 - h_2\|_\infty .$$

*Solution:* First split the term into two separate terms:

$$|L_S(h_1) - L_S(h_2)| \leq |R(h_1) - R(h_2)| + |\widehat{R}(h_1) - \widehat{R}(h_2)|$$
$$= \Big| \mathop{\mathrm{E}}_{x,y}[(h_1(x) - y)^2 - (h_2(x) - y)^2] \Big| + \Big| \frac{1}{m} \sum_{i=1}^m (h_1(x_i) - y_i)^2 - (h_2(x_i) - y_i)^2 \Big| .$$

Then, expanding the term

$$(h_1(x) - y)^2 - (h_2(x) - y)^2 = (h_1(x) - h_2(x))(h_1 + h_2 - 2y)$$
$$= (h_1(x) - h_2(x))\big((h_1 - y) + (h_2 - y)\big) \leq \|h_1 - h_2\|_\infty 2M ,$$

allows us to bound both the empirical and true error, resulting in a total bound of $4M\|h_1 - h_2\|_\infty$. □

2. Assume that $H$ can be covered by $k$ subsets $B_1, \ldots, B_k$, that is $H = B_1 \cup \ldots \cup B_k$. Then, show that, for any $\epsilon > 0$, the following upper bound holds:

$$\Pr_{S \sim D^m}\left[\sup_{h \in H} |L_S(h)| \geq \epsilon\right] \leq \sum_{i=1}^{k} \Pr_{S \sim D^m}\left[\sup_{h \in B_i} |L_S(h)| \geq \epsilon\right].$$

*Solution:* This follows by splitting the event into the union of several smaller events and then using the sum rule,

$$\Pr_S\left[\sup_{h \in H} |L_S(h)| \geq \epsilon\right] \\ = \Pr_S\left[\bigvee_{i=1}^{k} \sup_{h \in B_i} |L_S(h)| \geq \epsilon\right] \leq \sum_{i=1}^{k} \Pr_S\left[\sup_{h \in B_i} |L_S(h)| \geq \epsilon\right].$$

□

3. Finally, let $k = \mathcal{N}(H, \frac{\epsilon}{8M})$ and let $B_1, \ldots, B_k$ be balls of radius $\epsilon/(8M)$ centered at $h_1, \ldots, h_k$ covering $H$. Use part (a) to show that for all $i \in [1, k]$,

$$\Pr_{S \sim D^m}\left[\sup_{h \in B_i} |L_S(h)| \geq \epsilon\right] \leq \Pr_{S \sim D^m}\left[|L_S(h_i)| \geq \frac{\epsilon}{2}\right],$$

and apply Hoeffding's inequality to prove (4).

*Solution:* For any $i$ let $h_i$ be the center of ball $B_i$ with radius $\frac{\epsilon}{8M}$. Note that for any $h \in H$ we have $|L_S(h) - L_S(h_i)| \leq 4M\|h - h_i\|_\infty \leq \epsilon/2$. Thus, if for any $h \in B_i$ we have $|L_S(h)| \geq \epsilon$ it must be the case that $|L_S(h_i) \geq \epsilon_2|$, which shows the inequality.

To complete the bound, we use Hoeffding's inequality applied to the random variables $(h(x_i) - y_i)^2/m \leq M^2/m$, which guarantees

$$\Pr_S\left[|L_S(h_i)| \geq \frac{\epsilon}{2}\right] \leq 2\exp\left(\frac{-m\epsilon^2}{2M^4}\right).$$

□