CSE522, Winter 2011, Learning Theory Lecture 5 - 01/24/2011

Infinite Hypothesis Classes: Rademacher complexity

Lecturer: Ofer Dekel *Scribe: Yanping Huang*

1 Review: empirical risk minimization

For a hypothesis class $\mathcal{H}$, we define the empirical risk minimizer $h_{ERM} = \operatorname{argmin}_{h\in\mathcal{H}} l(h;\mathcal{S})$ and the risk minimizer (the optimal hypothesis in the class) $h^\star = \operatorname{argmin}_{h\in\mathcal{H}} l(h;\mathcal{D})$. By the definition of ERM, we know that

$$l(h_{ERM};\mathcal{S}) - l(h^\star;\mathcal{S}) \;\le\; 0 \quad . \tag{1}$$

If we are able to prove:

$$\begin{aligned} l(h^\star;\mathcal{S}) - l(h^\star;\mathcal{D}) &\le \epsilon_1 &(2)\\ l(h_{ERM};\mathcal{D}) - l(h_{ERM};\mathcal{S}) &\le \epsilon_2 &(3)\end{aligned}$$

then we can simply sum Eqs.(1 - 3) and conclude that the excess risk is upper bounded by

$$l(h_{ERM};\mathcal{D}) - l(h^\star;\mathcal{D}) \;\le\; \epsilon_1 + \epsilon_2 \quad .$$

Since $h^\star$ is a deterministic function (it does not rely on the random sample S), we can prove Eq.(2) by directly applying Hoeffding's inequality. The same cannot be said for h_{ERM}, which is a random function that depends on the sample S. Our strategy is therefore to prove something more general than Eq.(3), namely,

$$\forall h \in \mathcal{H} \quad l(h_{ERM};\mathcal{D}) - l(h_{ERM};\mathcal{S}) \;\le\; \epsilon_2 \quad .$$

2 Generalization Bound for infinite hypothesis space

Theorem 1. *If the size of the hypothesis space is infinite, $|\mathcal{H}| = \infty$, the loss function $l \in [0,c]$, and $\mathcal{S}$ is the sample set drawn from distribution $\mathcal{D}$ with $|\mathcal{S}| = m$. Then $\forall\delta > 0$ and $\forall h \in \mathcal{H}$, with probability at least $1-\delta$,*

$$|l(h;\mathcal{S}) - l(h;\mathcal{D})| \le \epsilon(\delta) = \mathcal{R}(l\circ\mathcal{H}) + c\sqrt{\frac{\log(1/\delta}{2m}} \tag{4}$$

Proof. To apply Hoeffding's inequality, we define

$$f(S) = \max_{h\in\mathcal{H}}[l(h;\mathcal{D} - l(h;\mathcal{S})].$$

First we show that $f(\mathcal{S})$ is $\frac{c}{m}$ bounded. For all hypothesis $h \in \mathcal{H}$, we change one example in $\mathcal{S} \to \mathcal{S}'$. $l \in [0,c]$, thus $|l(h,\mathcal{S}') - l(h,\mathcal{S}| \le \frac{c}{m}$. $l(h,\mathcal{D})$ remains the same, we have $|f(\mathcal{S}) - f(\mathcal{S};)| \le \frac{c}{m}$.

Next we apply McDiarmid's inequality,

$$\begin{aligned} Pr(f(\mathcal{S}) - \mathsf{E}[f(\mathcal{S})| \ge \epsilon) &\le \exp(-\frac{2m\epsilon^2}{c^2}) = \delta \\ f(s) &\le E[f(\mathcal{S})] + c\sqrt{\log(1/\delta)/2m}\end{aligned}$$

where $\mathsf{E}[f(\mathcal{S})] = \mathsf{E}_{\mathcal{S}}\{\max_{h\in\mathcal{H}}[l(h;\mathcal{D} - l(h;\mathcal{S})]\}$. The expectation is taken over all possible samples.

Third, we show that $\mathsf{E}[f(\mathcal{S})] \le \mathcal{R}(l\circ\mathcal{H})$, where $\mathcal{R}(l\circ\mathcal{H})$ is the Rademacher complexity. We begin with the following tow lemmas.

Lemma 2. $\max_{i\in\mathcal{I}} \mathsf{E}(X_i) \leq \mathsf{E}(\max_{i\in\mathcal{I}} X_i)$

Proof. $\forall j \in \mathcal{I}$, $x_j \leq \max x_i$. $\therefore \mathsf{E}(X_j) \leq \mathsf{E}(\max_i X_i)$. It follows that $\max_j \mathsf{E}(X_j) \leq \mathsf{E}(\max_i X_i)$. $\square$

Lemma 3. *Let Z_1 and Z_2 be identical independent distributed,* $\mathsf{E}[f(Z_1, Z_2] = \mathsf{E}[f(Z_2, Z_1)]$

Proof. $\mathsf{E}[f(Z_1, Z_2)] = \int_{z_1,z_2} f(Z_1, Z_2)P(z_1, z_2)d_{z_1}d_{z_2} = \int_{z_2,z_1} f(Z_2, Z_1)P(z_2, z_1)d_{z_2}d_{z_1} = \mathsf{E}[f(Z_2, Z_1)]$. $\square$

Using lemma 2, we can show

$$\begin{aligned}
\mathsf{E}_{\mathcal{S}}[f(\mathcal{S})] &= \mathsf{E}_{\mathcal{S}}[\max_h [l(h;\mathcal{D}) - l(h;\mathcal{S})]] \\
&= \mathsf{E}_{\mathcal{S}}[\max_h [E_{\mathcal{S}'} - l(h;\mathcal{S})]] \qquad (\text{Define } \mathcal{S}' = \{(x_i', y_i'\}_{i=1}^m) \\
&\leq \mathsf{E}_{\mathcal{S}}\mathsf{E}_{\mathcal{S}'}[\max_h (l(h;\mathcal{S}') - l(h;\mathcal{S}))] \qquad (\text{Using Lemma2}) \\
&= \mathsf{E}_{\mathcal{S}}\mathsf{E}_{\mathcal{S}'}[\max_h [\frac{1}{m}\sum_{i=1}^{m}(l_i' - l_i)]] \qquad (5)
\end{aligned}$$

where $l_i = l(h; (x_i, y_i))$ and $l_i' = l(h; (x_i', y_i'))$.

Lemma 3 allows us to swap any pair of (l_i, l_i') we want. We can define $\vec{\sigma} = (\sigma_1 \ldots \sigma_m)^T \in \{\pm 1\}^m$, $\sigma_i = 1$ with probability $1/2$ and $\sigma_i = -1$ with probability $1/2$, for any $i = 1, \ldots, m$. We continues on inequality 5,

$$\begin{aligned}
\mathsf{E}[f(\mathcal{S})] &\leq \mathsf{E}_{\mathcal{S}}\mathsf{E}_{\mathcal{S}'}[\max_h [\frac{1}{m}\sum_{i=1}^{m}(l_i' - l_i)]] \\
&= \mathsf{E}_{\mathcal{S}}\mathsf{E}_{\mathcal{S}'}\mathsf{E}_{\vec{\sigma}}[\max_{h\in\mathcal{H}} [\frac{1}{m}\sum_{i=1}^{m}\sigma_i(l_i' - l_i)]] \\
&= \frac{1}{m}\mathsf{E}_{\mathcal{S}}\mathsf{E}_{\mathcal{S}'}\mathsf{E}_{\vec{\sigma}}[\max_{h\in\mathcal{H}} (\sum_{i=1}^{m}\sigma_i l_i' - \sum_{i=1}^{m}\sigma_i l_i)] \\
&\leq \frac{1}{m}\mathsf{E}_{\mathcal{S}}\mathsf{E}_{\mathcal{S}'}\mathsf{E}_{\vec{\sigma}}[\max_{h\in\mathcal{H}} (\sum_{i=1}^{m}\sigma_i l_i') + \max_{h\in\mathcal{H}} (\sum_{i=1}^{m}\sigma_i l_i)] \\
&= \frac{1}{m}\mathsf{E}_{\mathcal{S}'}\mathsf{E}_{\vec{\sigma}}[\max_{h\in\mathcal{H}} (\sum_{i=1}^{m}\sigma_i l_i')] + \frac{1}{m}\mathsf{E}_{\mathcal{S}}\mathsf{E}_{\vec{\sigma}}[\max_{h\in\mathcal{H}} (\sum_{i=1}^{m}\sigma_i l_i)] \\
&= \frac{2}{m}\mathsf{E}_{\mathcal{S}}\mathsf{E}_{\vec{\sigma}}[\max_{h\in\mathcal{H}} (\sum_{i=1}^{m}\sigma_i l(h; (x_i, y_i)))] \\
&\equiv \mathcal{R}_m(l \circ \mathcal{H}) \qquad (6)
\end{aligned}$$

$\square$

Remarks on Rademacher's complexity:

- Since $\sigma_i \in \{\pm 1\}$, we can rewrite the Rademacher's complexity as:

$$\mathcal{R}_m(l \circ \mathcal{H}) = \frac{2}{m}\mathsf{E}_{\mathcal{S}}\mathsf{E}_{\vec{\sigma}}[\max_{h\in\mathcal{H}} (\sum_{i\in\{i:\sigma_i=1\}} l_i - \sum_{i\in\{i:\sigma_i=-1\}} l_i)]$$

  The $\vec{\sigma}$ partitioned the sample $\mathcal{S}$ into two disjoint sets. The Rademacher's complexity estimates how much difference between the total losses of two random-assigned disjoint sets can a hypothesis make.

- We can rewrite $\vec{l} = \{l_1, \ldots, l_m\}$. Then the inner product $< \vec{\sigma}, \vec{l} >$ is a measurement of the correlation between two vectors $\vec{\sigma}$ and $\vec{l}$. The Rademacher's complexity measures how well correlated the most-correlated hypothesis is to a random labeling of points in $\mathcal{S}$.
- The Rademacher's complexity depends on the distribution $\mathcal{D}$. We need to know $\mathcal{D}$ in order to compute $\mathcal{R}_m(l \circ \mathcal{H})$. This leads to the so-called empirical Rademacher's complexity.

3 Empirical Rademacher Average

We define the empirical Rademacher average as:

$$f'(\mathcal{S}) = \frac{2}{m} \mathsf{E}_{\vec{\sigma}}[\max_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i l(h; (x_i, y_i))] = \hat{\mathcal{R}}_m(l \circ \mathcal{H}, \mathcal{S}) \tag{7}$$

Notice that $f'(\mathcal{S})$ satisfies the $\frac{2c}{m}$ bounded difference property. Since $\mathsf{E}_{\mathcal{S}}[f'(\mathcal{S})] = \mathcal{R}_m(l \circ \mathcal{H})$, applying McDiarmid's inequality we have

Theorem 4. $\forall \delta \geq 0$, *with probability at least* $1 - \delta$,

$$\mathsf{E}_{\mathcal{S}}[f'(\mathcal{S})] - f'(\mathcal{S}) \leq 2c\sqrt{\frac{\log(1/\delta)}{2m}} \tag{8}$$

Define the set $\Omega = \{\mathcal{S} : f(\mathcal{S}) > E_{\mathcal{S}}[f(\mathcal{S})] + c\sqrt{\frac{1/\delta}{2m}}\}$, and $\Omega' = \{\mathcal{S} : E_{\mathcal{S}}[f'(\mathcal{S})] > f'(\mathcal{S}) + 2c\sqrt{\frac{1/\delta}{2m}}\}$. From Bole's inequality we have $P(\Omega \cup \Omega') \leq P(\Omega) + P(\Omega')$. We then have the following bound:

Theorem 5. $\forall \delta \geq 0$, *with probability at least* $1 - 2\delta$,

$$\forall h \in \mathcal{H}: \quad l(h; \mathcal{D}) - l(; \mathcal{S}) \leq \hat{\mathcal{R}}_m(l \circ \mathcal{H}, \mathcal{S}) + 3c\sqrt{\frac{1/\delta}{2m}} \tag{9}$$

3.1 Examples

Example 1 : Binary classification with 0–1 loss

In this example, $y \in -1, +1$, the 0–1 loss function $l(h; (x, y)) = \mathbf{1}_{h(x) \neq y}$. For a hypothesis class $\mathcal{H}$ and a training sample $\mathcal{S}$, assume that we have an algorithm returns the empirical risk minimizer $h_{ERM} = \mathrm{argmin}_{h \in \mathcal{H}} l(h; \mathcal{S})$. We would like to compute the upper bound of $l(h_{ERM}; \mathcal{D})$ using the uniform bound for the infinite hypothesis class.

The empirical Rademacher average can be written as:

$$\begin{aligned}
\mathcal{R}_m(l \circ \mathcal{H}) &= \frac{2}{m} \mathsf{E}_{\vec{\sigma}} \max_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i l(h; (x_i, y_i)) \\
&= \frac{2}{m} \mathsf{E}_{\vec{\sigma}} \max_{h \in \mathcal{H}} [\sum_{i=1}^{m} l(h; (x_i, \sigma_i y_i)) + \sum_{i=1}^{m} (\sigma_i l(h; (x_i, y_i) - l(h; (x_i, \sigma_i y_i))] \\
&= \frac{2}{m} \mathsf{E}_{\vec{\sigma}} \max_{h \in \mathcal{H}} [\sum_{i=1}^{m} l(h; (x_i, \sigma_i y_i)) - \sum_{i: \sigma_i = -1} 1]
\end{aligned} \tag{10}$$

The above equation 10 can be verified by different combinations of l_i and σ_i: As shown in the above Table 1, the difference $[\sigma l(h; (x, y)) - l(h; (x, \sigma y)] = 0$ when $\sigma = 1$, and -1 when $\sigma = -1$.

σ	$(h(x), y)$	$\sigma l(h; (x, y))$	$(h(x), \sigma y)$	$l(h; (x, \sigma y)$	$\sigma l(h; (x, y)) - l(h; (x, \sigma y)$
1	$h(x) = y$	0	$h(x) = \sigma y$	0	0
1	$h(x) \neq y$	1	$h(x) \neq \sigma y$	1	0
-1	$h(x) \neq y$	-1	$h(x) = \sigma y$	0	-1
-1	$h(x) = y$	0	$h(x) \neq \sigma y$	1	-1

Continue on the above derivation 10, we have

$$\begin{aligned}
\mathcal{R}_m(l \circ \mathcal{H}) &= \frac{2}{m} \mathsf{E}_{\vec{\sigma}} \max_{h \in \mathcal{H}} [\sum_{i=1}^{m} l(h; (x_i, \sigma_i y_i)) - \frac{2}{m} \frac{m}{2} \\
&= \frac{2}{m} \mathsf{E}_{\vec{\sigma}} \max_{h \in \mathcal{H}} [\sum_{i=1}^{m} [1 - l(h; (x_i, -\sigma_i y_i))] - 1 \\
&= 1 + \frac{2}{m} \mathsf{E}_{\vec{\sigma}} \max_{h \in \mathcal{H}} [\sum_{i=1}^{m} -l(h; (x_i, -\sigma_i y_i))] \\
&= 1 - \frac{2}{m} \mathsf{E}_{\vec{\sigma}} \min_{h \in \mathcal{H}} [\sum_{i=1}^{m} l(h; (x_i, -\sigma_i y_i))] \\
&= 1 - 2\mathsf{E}_{\vec{\sigma}} \min_{h \in \mathcal{H}} \frac{1}{m} [\sum_{i=1}^{m} l(h; (x_i, \sigma_i))] \qquad (11)
\end{aligned}$$

Again we define $f''(\vec{\sigma}) = \min_{h \ in \mathcal{H}} \frac{1}{m} [\sum_{i=1}^{m} l(h; (x_i, \sigma_i))]$, $f''(\vec{\sigma})$ satisfies $\frac{2}{m}$ bounded difference property. Thus we have:

$$\mathsf{E}[f''(\vec{\sigma})] \leq f''(\vec{\delta}) + 2\sqrt{\frac{\log(1/\delta)}{2m}} \qquad (12)$$

with prability at least $1 - \delta$.

Corollary 6. $\forall \delta \geq 0$, *with probability at least* $1 - 3\delta$,

$$\forall h \in \mathcal{H}, l(h; \mathcal{D}) \leq l(h; \mathcal{S}) + (1 - 2\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} l(h; (x_i, \sigma_i)) + 5\sqrt{\frac{\log(1/\delta)}{2m}} \qquad (13)$$

If some hypothesis $h \in \mathcal{H}$ manges to “explain” the random labels such that $\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} l(h; (x_i, \sigma_i) = 0$, then the complexity for $\mathcal{H}$ would reach the maximum. A hypothesis can be considered a “good” hypothesis if $l(h; (x_i, \sigma_i)) = 0$ with probability 0.5, the expected loss with respect to random labels is just 0.5.