CSE522, Winter 2011, Learning Theory Lecture 6 - 01/31/2011

Linear Hypothesis Classes

Lecturer: Ofer Dekel *Scribe: Yanping Huang*

1 Review: Rademacher's Complexity

Theorem 1. *Let the loss function $l \in [0, c]$, and $\mathcal{S}$ be 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 $1 - \delta$, we have*

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

where the Rademacher complexity

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

1.1 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 random vector $\vec{\sigma}$ 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 $\mathcal{S}$.

- When the loss function is a constant independent of examples, $l = 1$. We have $\mathsf{E}_{\vec{\sigma}} \sum_i \sigma_i \times 1 = 0$. In this case, $\mathcal{R}_m(l \circ \mathcal{H}) = 0$.

- If $\mathcal{H} = \{h\}$, then $\mathcal{R}_m(l \circ \mathcal{H}) = 0$

- In literature, sometimes the definition of Rademacher's complexity is written as

$$\mathcal{R}_m^{ori}(l \circ \mathcal{H}) = \frac{2}{m} \mathsf{E}_{\mathcal{S}} \mathsf{E}_{\vec{\sigma}}[\max_{h \ in \mathcal{H}} \left| \sum_{i=1}^{m} \sigma_i l_i \right|] \tag{3}$$

  However, this definition is inferior since it is a higher upper bound than the definition in Eq 2. In some special cases such as $\mathcal{H} = \{h\}$ and $l = 1$, $\mathcal{R}_m^{ori}(l \circ \mathcal{H}) > 0$. And the absolute value in the definition is generally harder to work with.

- Gaussian complexity is a similar complexity with similar physical meanings, and can be obtained from the previous complexity using with $\sigma_i \sim N(0, 1)$.

1.2 Special Case: Binary Classification

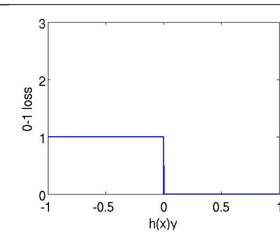
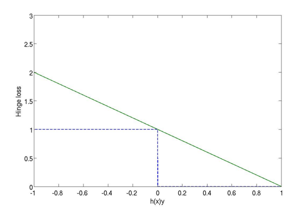
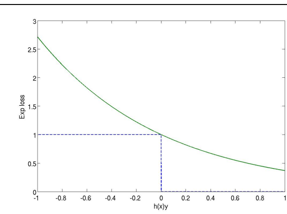
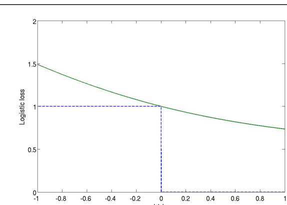
In this case, $y \in \{+1, -1\}$, l is 0–1 loss. $\vec{\sigma} = \{\sigma_1, \ldots, \sigma_m\}$ is a random vector with $Pr(\sigma_i = 1)$ with probability 1/2, and $Pr(\sigma_i = 0) = 0$ with probability 1/2. $\mathcal{S}' = \{(x_i, \sigma_i)\}_{i=1}^{m}$. Then $\forall \delta > 0$, with probability at least $1 - \delta$, we have $\hat{\mathcal{R}}_m(l \circ \mathcal{H}, \mathcal{S}) \leq 1 - 2 \min_{h \in \mathcal{H}} l(h; \mathcal{S}')$.

Note that $\hat{\mathcal{R}}_m(l \circ \mathcal{H}, \mathcal{S})$ becomes minimum when $l(\bar{h}; \mathcal{S}') = 1/2$ for some $\bar{h} \in \mathcal{H}$. That means that $\bar{h}$ can only predict random labels with probability 1/2. In the worse case where $\hat{\mathcal{R}}_m(l \circ \mathcal{H}, \mathcal{S})$ becomes maximum, we have $l(\bar{h}, \mathcal{S}') = 0$, when $\bar{h}$ can perfectly predict any random labels. In the average case, we expect a "good" hypothesis class $\mathcal{H}$ has the property that $\hat{\mathcal{R}}_m(l \circ \mathcal{H}, \mathcal{S}) \sim O(\frac{1}{m})$.

2 Linear hypothesis classes

In these classes, the hypotheses are parametrized by a linear vector w such that $h_w(x) =< w, x >$ where $w \in \mathbb{R}^n$ and $x \in \mathbb{R}^n$.

- Regression problems, $y \in \mathbb{R}$. The loss function is a function of the difference between prediction and y: $l(h; (x, y)) = l(h(x) - y)$. For square loss, $l(h; (x, y)) = (h(x) - y)^2$. In general $l(h; (x, y)) = |h(x) - y|^p$ for $p > 0$.
- Confidence rated binary classification (margin based confidence). Here $y \in \mathbb{R}$, $sign(h(x))$ represents the binary label of the example x, and $|h(x)|$ represents the corresponding confidence.
- Binary classification $y = \{+1, -1\}$. In this case, the loss function $l(h(x)y)$ is in general a function of $h(x)y$. Some popular choices of loss functions are:

0–1 loss	$\mathbb{I}_{\{h(x) \neq y\}}$	
Hinge loss	$[1 - h(x)y]_+$	
Exponential Loss(Ada Boost)	$\exp(-h(x)y)$	
Logistic loss	$\log(c + \exp(-yh(x)))$	

To help our analysis, the desired loss function should 1) be not less than the 0–1 loss function, 2) be convex, 3) and be Lipschitz. A function $l(.)$ is called $\lambda-$Lipschitz iff $|l(\alpha) - l(\beta)| \leq \lambda|\alpha - \beta|$.

Theorem 2. *If the loss function is $\lambda-$Lipschitz, we have*

$$\mathcal{R}_m(l \circ \mathcal{H}) \leq \lambda \mathcal{R}_m(\mathcal{H}) \tag{4}$$

$$\tag{5}$$

where

$$\mathcal{R}_m(\mathcal{H}) = \frac{2}{m} \mathsf{E}_{\vec{\sigma}} \mathsf{E}_{\mathcal{S}} \max_{h \in \mathcal{H}} \sum_{i=1}^{m} \sigma_i h(x_i) \tag{6}$$

The same inequality also holds for $\hat{\mathcal{R}}_m(l \circ \mathcal{H}, \mathcal{S})$

Theorem 2 can be shown be the following lemma,

Lemma 3. *Let $g_i(\theta)$ and $f_i(\theta)$ be sets of functions such that $\forall i, \theta, \theta'$,*

$$|g_i(\theta) - g_i(\theta')| \leq |f_i(\theta) - f_i(\theta')| \tag{7}$$

Then for any function $c(x, \theta)$ and any distribution over $\mathbb{X}$,

$$\mathsf{E}_{\vec{\sigma}} \mathsf{E}_x \sup_{\theta} [c(x, \theta) + \sum_i \sigma_i g_i(\theta)] \leq \mathsf{E}_{\vec{\sigma}} \mathsf{E}_x \sup_{\theta} [c(x, \theta) + \sum_i \sigma_i f_i(\theta)] \tag{8}$$

Proof. We are going to show it by induction. The lemma obviously holds for $n = 0$. Then suppose the lemma holds for $n = k$, for $n = k + 1$:

$$\begin{aligned}
& \mathsf{E}_{\sigma_1 \ldots \sigma_{k+1}} \mathsf{E}_x \sup_{\theta} [c(x, \theta) + \sum_{i=1}^{k+1} \sigma_i g_i(\theta)] \\
= \quad & \mathsf{E}_{\sigma_1 \ldots \sigma_k} \mathsf{E}_x \sup_{\theta_1, \theta_2} [\frac{c(x, \theta_1) + c(x, \theta_2)}{2} + \sum_{i=1}^{k} \sigma_i (\frac{g_i(\theta_1) + g_i(\theta_2)}{2}) + \frac{g_{k+1}(\theta_1) - g_{k+1}(\theta_2)}{2}] \\
= \quad & \mathsf{E}_{\sigma_1 \ldots \sigma_k} \mathsf{E}_x \sup_{\theta_1, \theta_2} [\frac{c(x, \theta_1) + c(x, \theta_2)}{2} + \sum_{i=1}^{k} \sigma_i (\frac{g_i(\theta_1) + g_i(\theta_2)}{2}) + \frac{|g_{k+1}(\theta_1) - g_{k+1}(\theta_2)|}{2}] \\
\leq \quad & \mathsf{E}_{\sigma_1 \ldots \sigma_k} \mathsf{E}_x \sup_{\theta_1, \theta_2} [\frac{c(x, \theta_1) + c(x, \theta_2)}{2} + \sum_{i=1}^{k} \sigma_i (\frac{g_i(\theta_1) + g_i(\theta_2)}{2}) + \frac{|f_{k+1}(\theta_1) - f_{k+1}(\theta_2)|}{2}] \\
= \quad & \mathsf{E}_{\sigma_1 \ldots \sigma_k} \mathsf{E}_x \sup_{\theta_1, \theta_2} [\frac{c(x, \theta_1) + c(x, \theta_2)}{2} + \sum_{i=1}^{k} \sigma_i (\frac{g_i(\theta_1) + g_i(\theta_2)}{2}) + \frac{f_{k+1}(\theta_1) - f_{k+1}(\theta_2)}{2}] \\
= \quad & \mathsf{E}_{\sigma_1 \ldots \sigma_{k+1}} \mathsf{E}_x \sup_{\theta} [c(x, \theta) + \sum_{i=1}^{k} \sigma_i g_i(\theta)) + \sigma_{k+1} f_{k+1}(\theta)] \\
\leq \quad & \mathsf{E}_{\sigma_1 \ldots \sigma_{k+1}} \mathsf{E}_x \sup_{\theta} [c(x, \theta) + \sigma_{k+1} f_{k+1}(\theta) + \sum_{i=1}^{k} \sigma_i f_i(\theta))]
\end{aligned}$$

Let $c(x, \theta) = 0$, $g_i(\theta) = l(h_w(x)y)$ and $f_i(\theta) = \lambda h_w(x)y$, we apply the above lemma and prove Theorem 2 $\square$

Theorem 4. *A linear hypothesis class $\mathcal{H}$ such that $\forall h \in \mathcal{H}$, $h_w(x) =< w, x >\in [-1, +1]$, where $w \in \mathbb{R}^n$ $\|w\|_2 \leq \mathcal{B}$, and $x \in \mathbb{R}^n$, $\|x\|_2 \leq \mathcal{X}$, we have*

$$\hat{\mathcal{R}}_m(\mathcal{H}, \mathcal{S}) \leq \frac{2\mathcal{B}\mathcal{X}}{\sqrt{m}} \tag{9}$$

Proof.

$$\begin{aligned}
\hat{\mathcal{R}}_m(\mathcal{H}, \mathcal{S}) &= \frac{2}{m}\mathsf{E}_{\vec{\sigma}} \max_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i h(x_i) \\
&= \frac{2}{m}\mathsf{E}_{\vec{\sigma}} \max_{\|w\|_2\leq\mathcal{B}} \sum_{i=1}^{m} \sigma_i < w, x_i > \\
&= \frac{2}{m}\mathsf{E}_{\vec{\sigma}} \max_{\|w\|_2\leq\mathcal{B}} < w, \sum_{i=1}^{m} \sigma_i x_i > \\
&\leq \frac{2}{m}\mathsf{E}_{\vec{\sigma}} \max_{\|w\|_2\leq\mathcal{B}} \|w\| \left\| \sum_{i=1}^{m} \sigma_i x_i \right\| \quad \text{(CauchySchwarz inequality)} \\
&= \frac{2\mathcal{B}}{m}\mathsf{E}_{\vec{\sigma}} \left\| \sum_{i=1}^{m} \sigma_i x_i \right\| \\
&= \frac{2\mathcal{B}}{m}\mathsf{E}_{\vec{\sigma}} \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} \sigma_i\sigma_j < x_i, x_j >} \quad \text{(linearity of inner product)} \\
&\leq \frac{2\mathcal{B}}{m} \sqrt{\mathsf{E}\sum_{ij} \sigma_i\sigma_j < x_i, x_j >} \quad \text{(Jensen's inequality)} \\
&= \frac{2\mathcal{B}}{m} \sqrt{\sum_{ij} < x_i, x_j > \mathsf{E}\sigma_i\sigma_j} \\
&\leq \frac{2\mathcal{B}}{m} \sqrt{\sum_{i} \|x_i\|^2} \\
&\leq \frac{2\mathcal{B}}{m} \sqrt{m}\mathcal{X} \\
&= \frac{2\mathcal{B}\mathcal{X}}{\sqrt{m}}
\end{aligned}$$

□