CSE522, Winter 2011, Learning Theory Lecture 7 - 01/25/2011

Linear Hypothesis Classes and Regularization

Lecturer: Ofer Dekel *Scribe: Amol Kapila*

1 Recap

Our goal is to get uniform convergence bounds on infinite hypothesis classes. Suppose that $\ell \in [0, c]$. We have proved two main results:

1. For every $\delta > 0$, with probability at least $1 - \delta$, for every $h \in H$,
$$\ell(h; \mathcal{D}) \leq \ell(h; S) + R_m(\ell \circ H) + \sqrt{\frac{\log(1/\delta)}{2m}}.$$
2. For every $\delta > 0$, with probability at least $1 - \delta$, for every $h \in H$,
$$\ell(h; \mathcal{D}) \leq \ell(h; S) + \widehat{R}_m(\ell \circ H, S) + 3\sqrt{\frac{\log(1/\delta)}{2m}},$$

where
$$R_m(\ell \circ H) = \frac{2}{m} \mathbb{E}_{S \sim \mathcal{D}} \mathbb{E}_\sigma \max_{h \in H} \sum_{i=1}^{m} \sigma_i \ell(h; (x_i, y_i))$$

and
$$\sigma_i = \begin{cases} 1 & \text{w.p. } 1/2 \\ 0 & \text{w.p. } 1/2 \end{cases}.$$

The empirical Rademacher average, which is much more useful than the Rademacher complexity because it does not depend on the unknown distribution, is
$$\widehat{R}_m(\ell \circ H, S) = \frac{2}{m} \mathbb{E}_\sigma \max_{h \in H} \sum_{i=1}^{m} \sigma_i \ell(h; (x_i, y_i)).$$

2 λ-Lipschitz Loss Functions

Definition 1. *The loss function ℓ is* λ-Lipschitz *if for any α and α', $|\ell(\alpha) - \ell(\alpha')| \leq \lambda \|\alpha - \alpha'\|$.*

Fact 2. *The Hinge loss is Lipschitz everywhere. Squared error loss is Lipschitz only on a bounded domain.*

Theorem 3. *If ℓ is a λ-Lipschitz function, then $R_m(\ell \circ H) \leq \lambda R_m(H)$, where*
$$R_m(H) = \frac{2}{m} \mathbb{E}_S \mathbb{E}_\sigma \max_{h \in H} \sum_{i=1}^{m} \sigma_i h(x_i).$$

Theorem 4. *More General Result: Consider sets of functions $\{g_i(\theta)\}$, $\{h_i(\theta)\}$. Assume that for every i, θ, θ', $|g_i(\theta) - g_i(\theta')| \leq |h_i(\theta) - h_i(\theta')|$. Then, for any $c(x, \theta)$ and any probability distribution over X,*
$$\mathbb{E}_\sigma \mathbb{E}_X \max_{\theta \in \Theta} \left(c(x, \theta) + \sum_{i=1}^{n} \sigma_i g_i(\theta) \right) \leq \mathbb{E}_\sigma \mathbb{E}_X \max_{\theta \in \Theta} \left(c(x, \theta) + \sum_{i=1}^{n} \sigma_i h_i(\theta) \right).$$

Observation 5. *This is a classic case of proving something more general in order to more easily prove a more specific result. Suppose $g_i(\theta) = \ell(yh(x_i))$ and $h_i(\theta) = \lambda h(x_i)$. If $\ell(h; (x_i, y_i))$ is λ-Lipschitz in $h(x)$ (for any fixed $y \in \mathcal{Y}$), then we can use Theorem 4 to prove Theorem 3.*

3 Linear Hypothesis Classes

Linear hypothesis classes consist of hypotheses of the form $h_w(x) = \langle w, x \rangle$. We can use this for confidence-rated binary classification. For classification, we consider loss functions of the form $\ell(yh(x))$, and we want the sign of $h(x)$ to equal y; i.e., $yh(x) > 0$. For regression problems, we want $h(x)$ to be close to y, so we might use $\ell(h(x) - y) = (h(x) - y)^2$.

3.1 Bounding the Rademacher Complexity of Linear Functions

Assume that $H = \{h_w(\cdot) \mid h_w(x) = \langle w, x \rangle, \|w\| \leq B\}$. Then,

$$\begin{aligned}
\widehat{R}_m(H,S) &= \frac{2}{m}\mathbb{E}_\sigma \max_{h\in H} \sum_{i=1}^m \sigma_i h(x_i) \\
&= \frac{2}{m}\mathbb{E}_\sigma \max_{\|w\|\leq B} \sum_{i=1}^m \sigma_i \langle w, x_i \rangle \\
&= \frac{2}{m}\mathbb{E}_\sigma \max_{\|w\|\leq B} \left\langle w, \sum_i \sigma_i x_i \right\rangle \\
&\leq \text{[This is an extremal case of the the Cauchy-Schwartz inequality, so we actually have equality]} \\
&\leq \frac{2}{m}\mathbb{E}_\sigma \max_{\|w\|\leq B} \|w\| \left\| \sum_{i=1}^m \sigma_i x_i \right\| \\
&= \frac{2B}{m}\mathbb{E}_\sigma \left\| \sum_{i=1}^m \sigma_i x_i \right\| \\
&= \frac{2B}{m}\mathbb{E}_\sigma \sqrt{\left\langle \sum_{i=1}^m \sigma_i x_i, \sum_{i=1}^m \sigma_i x_i \right\rangle} \\
&= \text{[linearity of the inner product]} \\
&= \frac{2B}{m}\mathbb{E}_\sigma \sqrt{\sum_{i=1}^m \sum_{j=1}^m \sigma_i \sigma_j \langle x_i, x_j \rangle} \\
&\leq \text{[Jensen's inequality and the concavity of the square root function]} \\
&\leq \frac{2B}{m} \sqrt{\mathbb{E}_\sigma \sum_{i=1}^m \sum_{j=1}^m \sigma_i \sigma_j \langle x_i, x_j \rangle} \\
&\leq \frac{2B}{m} \sqrt{\sum_{i=1}^m \sum_{j=1}^m \langle x_i, x_j \rangle \mathbb{E}_\sigma \sigma_i \sigma_j}.
\end{aligned}$$

Note that

$$\mathbb{E}_\sigma \sigma_i \sigma_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}.$$

Hence, we have

$$\widehat{R}_m(H,S) \leq \frac{2B}{m} \sqrt{\sum_{i=1}^m \|x_i\|^2},$$

and therefore, for a λ-Lipschitz loss function,

$$\widehat{R}_m(\ell \circ H, S) \leq \lambda \frac{2B}{m} \sqrt{\sum_i \|x_i\|^2}.$$

Furthermore, if $\|x_i\| \leq X$ for all i, then,

$$\widehat{R}_m(H, S) \leq \frac{2B}{m}\sqrt{mX^2} = \frac{2BX}{\sqrt{m}} \Rightarrow R_m(H) \leq \frac{2BX}{\sqrt{m}}.$$

Combining this with previous results, we have that for any $\delta > 0$, with probability at least $1-\delta$, (assuming ℓ is λ-Lipschitz, $\ell \in [0, c]$, $\mathcal{D}$ is such that $\|x_i\| \leq X$), for any w such that $\|w\| \leq B$,

$$\ell(h_w; \mathcal{D}) \leq \ell(h_w; S) + \lambda\frac{2BX}{\sqrt{m}} + c\sqrt{\frac{\log(1/\delta)}{2m}} = \ell(h_w; S) + O(1/\sqrt{m}).$$

Observation 6. *Note that the dimensionality of the input space has no role in this analysis. The curse of dimensionality has nothing to do with the dimension of the feature space, but instead, the Rademacher complexity of the hypothesis class.*

Rademacher complexity is a recent topic (mostly post-2000). Later, we will turn to "old-school" V-C theory, from the 1990s.

4 Structural Risk Minimization

Vapnik introduced SRM in the context of V-C theory. This is the old way of talking about things. Nowadays, we talk about *regularization*.

We want $h \in H$ with small risk $\ell(h; \mathcal{D})$. How do we choose H? If $H \subseteq H'$, then clearly $\min_{h \in H} \ell(h; S) \geq \min_{h \in H'} \ell(h; S)$. So, a bigger hypothesis class fits the data S better, but can hurt our estimation due to overfitting.

Let $H = \{w \mid \|w\| \leq 1\}$ and $H' = \{w \mid \|w\| \leq 2\}$. To be concrete, we will assume that we are dealing with the Hinge loss, although this is not strictly necessary. Then, for any $\delta > 0$, with probability at least $1-\delta$, S is such that

1. For all $\|w\| \leq 1$, $\ell(h_w; \mathcal{D}) \leq \ell(h_w, S) + \frac{\lambda 2X}{\sqrt{m}} + X\sqrt{\frac{\log(1/\delta)}{2m}}$.

2. For all $\|w\| \leq 2$, $\ell(h_w; \mathcal{D}) \leq \ell(h_w, S) + \frac{\lambda 4X}{\sqrt{m}} + 2X\sqrt{\frac{\log(1/\delta)}{2m}}$.

Lemma 7. $\sum_{i=1}^{\infty} 1/(i(i+1)) = 1$.

Proposition 8. *For all $B \in \mathbb{N}$, for all $\|w\| \leq B$, with probability at least $1 - \frac{\delta}{B(B+1)}$,*

$$\ell(h_w; \mathcal{D}) \leq \ell(h_w; S) + \frac{2\lambda BX}{\sqrt{m}} + BX\sqrt{\frac{\log(B(B+1)/\delta)}{2m}}.$$

Hence, for any $w \in \mathbb{R}^n$,

$$\ell(h_w; \mathcal{D}) \leq \ell(h_w; S) + \frac{2\lambda X(\|w\|+1)}{\sqrt{m}} + (\|w\|+1)\sqrt{\frac{\log(B(B+1)/\delta)}{2m}}.$$

4.1 Regularized ERM

In regularized ERM, we minimize the regularized empirical risk:

$$h_{RERM} = \arg\min_{w \in \mathbb{R}^n} \ell(h_w; S) + \beta \|w\|.$$

If we choose ℓ to be the Hinge loss and we use $\|w\|^2$ as our regularizer, then this reduces to the SVM. This is the justification for SVMs, which Vapnik invented.

5 Ensemble Methods (Boosting Algorithms)

Boosting incrementally adds another hypothesis to the sum, so that at the end of the day, you end up with $\tilde{h}_k = \sum_{i=1}^{k} \alpha_i h_i$, where $h_i \in H$. How does the complexity behave as you add more and more hypotheses to the space?

Definition 9. *If H is a hypothesis space, then the* convex hull *of H is*

$$\overline{H} = \left\{ \sum_{i=1}^{k} \alpha_i h_i | k \in \mathbb{N}, \alpha_i \geq 0, \sum_{i=1}^{k} \alpha_i = 1, h_i \in H \right\}.$$

Theorem 10. *$R(\overline{H}) = R(H)$, where $\overline{H}$ is the convex hull of H, as defined above.*