CS281B/Stat241B (Spring 2003) Statistical Learning Theory **Lecture: 16**

Applications of Rademacher averages in large margin classification

Lecturer: Peter L. Bartlett *Scribe: Zile Wei*

1 Consequences

The consequence of the upper bound of Rademacher average is that we can relate the minimum of empirical risk and the true risk. With probability larger than $1 - \delta$.

$$\begin{aligned} El_{\hat{f}} &\leq \inf_{f\in\mathcal{F}} El_f + \sqrt{\frac{2\log \pi_F(n)}{n}} + \sqrt{\frac{-c\log\delta}{n}} \\ &\leq \inf_{f\in\mathcal{F}} El_f + \sqrt{\frac{2d\log(en/d)}{n}} + \sqrt{\frac{-c\log\delta}{n}} \end{aligned} \tag{1}$$

where $d = d_{VC}(F)$ and $\hat{f} = \arg\min_{f\in\mathcal{F}} \hat{\mathbb{E}} l_f$. In fact, there is a refinement on 1) that $\sqrt{\frac{2d\log(en/d)}{n}}$ could be replaced by $c_1\sqrt{d/n}$. Asymptotically, the difference of empirical loss and true loss is dominated by $c_1\sqrt{d/n}$. Meanwhile, we can't do any better for some probability distribution, which the difference is about $c_1\sqrt{d/n}$. Here is the converse result with $c_1\sqrt{d/n}$ lower bound.

Theorem 1.1. For a function class $\mathcal{F}$ with $d_{VC}(\mathcal{F}) \geq d$, any decision function $f_n \in \mathcal{F} : (\mathcal{X}, \mathcal{Y})^n \times \mathcal{X} \to \mathcal{Y}$. There is a probability distribution P on $\mathcal{X} \times \mathcal{Y}$, s.t.

$$E[l(Y, f_n(X_1, Y_1, \cdots, X_n, Y_n; X))|D_n] - \inf_{f\in\mathcal{F}} El(f(X), Y)) \geq \min\{c_1, c_2\sqrt{d/n}\}$$

where $D_n = \{X_1, Y_1, \cdots, X_n, Y_n\}$.

Ideas of proof

The proof is quite similar to the converse example in the perceptron algorithm. We use probabilistic method and choose probability distribution P randomly from a finite set $\mathcal{P}$. $P \in \mathcal{P}$: P_x has support on a shattered set of $\mathcal{F}$ of size d. Set $P(Y = 1|X) = 1/2 \pm \epsilon$. We are forcing the algorithm to effectively guess the $\pm\epsilon$ in the condition probability $P(Y = 1|X)$ by observation D_n.

$\mathcal{F}$ contains all $\{\pm 1\}$-value function on $\{X_1, \cdots, X_n, X\}$. We can make estimate error close to Bayes risk, as n increases. Then f_n could correctly guess the $\pm\epsilon$ in $P(Y = 1|X)$ for most x, To solve this problem, the algorithm can't do better than $c/\sqrt{n}$.

(For the details, see Chapter 5 of M. Anthony and P. Bartlett. *Neural Network Learning: Theoretical Foundations.* Cambridge University Press. 1999.)

2 Properties of $R_n(F)$

(i) $F \subseteq G \Rightarrow R_n(F) \leq R_n(G)$
(ii) $R_n(cF) = |c|R_n(F)$, where, $cF = \{cf : f \in F\}$ and $c \in \mathbb{R}$.

(iii) $R_n(F+g) = R_n(F)$, where $F + g = \{f + g : f \in F\}$ and g is any uniformly bounded function.
(iv) $R_n(F) = R_n(Co(F)) = R_n(Co(F \cup -F))$, where $Co(F) = \{\sum \alpha_i f_i : \alpha_i \geq 0, \sum \alpha_i = 1, f_i \in F\}$ and $-F = \{-f : f \in F\}$.
(v) If $\phi_i : \mathbb{R} \to \mathbb{R}, i = 1, \cdots, n$ are Lipschitz with constant L, i.e., $|\phi_i(a) - \phi_i(b)| \leq L|a - b|$. Then

$$\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \phi_i(f(x_i)) \leq L\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i)$$

Proof.

(i) It's immediate from definition.

$$R_n(F) = E \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i).$$

supremum over a larger set will achieve higher upper bound than that of a smaller set.
(ii)

$$R_n(cF) = E \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i c f(x_i) = |c| E \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) = |c| R_n(F)$$

(iii)

$$R_n(F+g) = \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (f(x_i) + g(x_i)) = \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i)$$

for ϵ_i is a uniformly random variable over $\{\pm 1\}$, so $E(\epsilon_i g(x_i)|x_i) = 0$.
(iv) From (i), we know $R_n(F) \leq R_n(Co(F)) \leq Rn(co(F \cup -F))$. And

$$\begin{aligned}
R_n(Co(F \cup -F)) =& \mathbb{E}\Big(\sup_m \sup_{f_1, \cdots, f_m} \sup_{\|\alpha\|=1} \sum_{j=1}^{n} \alpha_j (\frac{1}{n} \sum_{i=1}^{n} \epsilon_i f_j(x_i)\Big) \\
=& \mathbb{E}\Big(\sup_m \sup_{f_1, \cdots, f_m} \sup_{\|\alpha\|} \alpha^t S\Big) \\
=& \mathbb{E}\Big(\sup_m \sup_{f_1, \cdots, f_m} \max_j s_j\Big) \\
=& \mathbb{E}\Big(\sup_m \sup_{f_1, \cdots, f_m} \max_j \frac{1}{n} \epsilon_i f_j(x_i)\Big) \\
=& R_n(F)
\end{aligned} \tag{2}$$

where $S = \{s_1, \cdots, s_n\}$ and $s_j = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f_j(x_i)$.
(v) It's called Ledoux-Talagrand contraction inequality (see Ledoux and Talagrand. Probability in Banach spaces: Isoperimetry and processes. Springer-Verlag. http://www.math.ohio-state.edu/∼talagran/book.ps.gz, p118–119) □

3 Applications of Rademacher averages

Suppose $\forall y : \hat{y} \mapsto l(y, \hat{y})$ is Lipschitz. $\phi(\alpha) = l(y_i, \alpha)$. Define Rademacher averages over loss function as

$$R_n(l_F) = \mathbb{E}\Big[\mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum \epsilon_i l(Y_i, f(X_i)) | X_1, Y_1, \cdots, X_N, Y_N\Big]$$

for all $l_f = \phi(y f(x))$.

3.1 Lipschitz function class

Suppose a function class $\mathcal{F} \subseteq \mathbb{R}^x$. Cost function $\phi : \mathbb{R} \to [0,1]$ is Lipschitz with constant 1. $|\phi(a) - \phi(b)| \leq |a - b|$.

Consider loss class as $l_F = \{(x,y) \mapsto \phi(yf(x)) : f \in \mathcal{F}\}$. e.g.

$$\begin{aligned} \phi(\alpha) =& \min(1, \max(0, 1-\alpha)) \text{ or} \\ \phi(\alpha) =& \min(1, (\max(0, 1-\alpha))^2) \text{ or} \\ \phi(\alpha) =& \min(1, e^{-\alpha}) \end{aligned} \tag{3}$$

Then with probability at least $1 - \delta$.

$$\begin{aligned} \sup_{f \in \mathcal{F}} \Big(\mathbb{E}\phi(Yf(X)) - \hat{\mathbb{E}}\phi(Yf(X) \Big) \leq& 2R_n(l_F) + \sqrt{\frac{-\log \delta}{2n}} \\ =& 2\mathbb{E}\mathbb{E}\Big[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \phi(Y_i, f(X_i)) | D_n \Big] + \sqrt{\frac{-\log \delta}{2n}} \\ \leq& 2R_n(F) + \sqrt{\frac{-\log \delta}{2n}} \end{aligned} \tag{4}$$

For a specific class: $\mathcal{F} = \lambda Co(G)$, $G = \{\pm 1\}^x$, $d_{VC}(G) = d$.

$$\begin{aligned} R_n(F) \leq& \lambda R_n(Co(G)) \\ =& \lambda R_n(G) \\ \leq& c\lambda\sqrt{d/n} \end{aligned} \tag{5}$$

i.e. with probability $\geq 1 - \delta$, any $f \in \mathcal{F}$ has

$$\mathbb{E}\phi(Y(f(X))) \leq \hat{\mathbb{E}}\phi(Y(f(X))) + c\lambda\sqrt{d/n} + \sqrt{\frac{-\log \delta}{2n}}$$

Hence, let $\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{\mathbb{E}}\phi(Yf(X))$, then

$$\mathbb{E}\phi(Y\hat{f}(X)) \leq \inf_{f \in \mathcal{F}} \mathbb{E}\phi(Yf(X)) + c\Big(\lambda\sqrt{d/n} + \sqrt{\frac{-\log \delta}{2n}}\Big)$$

We can think of λ as a regularization parameter. As λ becomes larger, we are considering a richer class and would achieve better approximation with price on $c\Big(\lambda\sqrt{d/n}\Big)$. It's estimation tradeoff in λ.

3.2 Estimate Rademacher average from data

Computing estimates of $R_n(G)$. Recall: with probability $\geq 1 - \delta$.

$$\begin{aligned} R_n(G) \leq& \mathbb{E}\Big[\sup_{g \in G} \frac{1}{n} \sum \epsilon_i g(x_i) | D_n \Big] + c\sqrt{\frac{-\log \delta}{n}} \\ \leq& \sup_{g \in G} \frac{1}{n} \sum \epsilon_i g(x_i) + c_2\sqrt{\frac{-\log \delta}{n}} \end{aligned} \tag{6}$$

But for $G \subseteq \{\pm 1\}^x$, if $l(y, \hat{y}) = \mathbb{1}[y \neq \hat{y}]$. We have

$$\begin{aligned} \sup_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i g(x_i) =& \sup_{g \in G} \frac{1}{n} \sum (1 - 2l(\epsilon_i, g(x_i))) \\ =& 1 - 2 \inf_{g \in G} \frac{1}{n} \sum_{i=1}^{n} l(\epsilon_i, g(x_i)) \end{aligned} \tag{7}$$

The last term is to minimize empirical risk over G with $y_i > \epsilon_i$

3.3 Computer Rademacher average for kernel classes

Optimization problem in kernel methods, e.g. SVM, is in this form:

$$\min_{f \in \mathcal{H}} \|f\|_{\mathcal{H}}^2 + c\hat{\mathbb{E}}\phi(Y(f(X)))$$

This is equivalent to: for some λ to

$$\begin{aligned} \text{minimize } & \hat{\mathbb{E}}\phi(Yf(X)) \\ \text{s.t. } & \|f\|_{\mathcal{H}} \leq \lambda \end{aligned} \tag{8}$$

Consider function class in Hilbert space $F = \{f \in H : \|f\|_H \leq \lambda\}$, we have, with probability $\geq 1 - \delta$

$$\sup_{f \in F} \left(\mathbb{E}\phi(Yf(X)) - \hat{\mathbb{E}}\phi(Yf(X)) \right) \leq 2R_n(F) + \sqrt{\frac{-\log \delta}{2n}} \leq 2\hat{R}_n(F) + c\sqrt{\frac{-\log \delta}{n}})$$

In kernel class, $\hat{R}_n(F) \leq \lambda\sqrt{Trace(K)}/n$, where $K_{ij} = k(x_i, x_j)$. K is the reproducing kernel of H. $Trace(K) = \sum_{i=1}^{n} k(x_i, x_i)$ is the sum of eigenvalues.

Also, $R_n(F) = E\hat{R}_n(F) \leq \lambda\sqrt{\frac{\sum_{j=1}^{\infty} \theta_j}{n}}$, where $\theta_1 \geq \theta_2 \geq \cdots$ are the sorted eigenvalues of the integral operator

$$T_k : f \mapsto \int_X k(\cdot, x) f(x) dx.$$

Recall: in Mercer theory, $k(x_1, x_2)$ can be decomposed into $\sum_{j=1}^{\infty} \theta_j \Psi_j(x_1)\Psi_j(x_2)$, where Ψ_j are orthonormal eigenfunctions of T_k.