Math 547: Mathematical Foundations of Statistical Learning Theory Fall 2017

Lecture 29 — October 27

Instructor: Stas Minsker *Scribe: Mose Wintner*

29.1 Excess Risk Bounds

29.1.1 Adaboost

Theorem 1. Let $\mathcal{F}$ be a base class of VC dimension $V(\mathcal{F}) < \infty$. Let $\widehat{g}_T$ be the output of Adaboost after T steps. Then for any $\theta > 0$,

$$P(Y \neq \widehat{g}_T(X)) \leq \frac{1}{n}\sum_{j=1}^{n} I\{Y_j \widehat{g}_T(X_j) \leq \theta\} + K\left(\frac{1}{\theta}\sqrt{\frac{V}{n}} + \sqrt{\frac{t}{n}}\right).$$

with probability $\geq 1 - e^{-t}$ where $K > 0$ is an absolute constant.

Proof:

$$\widehat{g}_T(X) = \frac{\sum_j \alpha_j f_j(x)}{\sum_j \alpha_j}$$

belongs to the convex hull of $\mathcal{F}$.

Let $\varphi_\theta(x)$ be 1 if $x \leq 0$, 0 if $x \geq \theta$, and $1 - x/\theta$ otherwise. It is Lipschitz continuous with Lipschitz constant $1/\theta$. Note that $I\{X \leq 0\} \leq \varphi_\theta(x) \leq I\{X \leq \theta\}$. Then

$$\begin{aligned}
P(Y\widehat{g}_T(X) \leq 0) &= \mathbb{E} I\{Y\widehat{g}_T(X) \leq 0\} \\
&\leq \mathbb{E}\varphi_\theta(Y\widehat{g}_T(X)) \\
&= \frac{1}{n}\sum \varphi_\theta(Y_j\widehat{g}_T(X_j)) - \frac{1}{n}\sum \varphi_\theta(Y_j\widehat{g}_T(X_j)) - \mathbb{E}\varphi_\theta(Y\widehat{g}_T(X)) \\
&\leq \frac{1}{n}\sum I\{Y_j\widehat{g}_T(X_j) \leq \theta\} + \sup_{g \in CH\mathcal{F}} \left|\frac{1}{n}\sum \varphi_\theta(Y_j g(X_j)) - \mathbb{E}\varphi_\theta(Yg(X))\right| \\
&\leq \frac{1}{n}\sum I\{Y_j\widehat{g}_T(X_j) \leq \theta\} + (Z = \text{bounded difference with } c_j = 1/n)
\end{aligned}$$

Therefore $Z \leq \mathbb{E}Z + \sqrt{\frac{t}{n}}$ with probability $\geq 1 - e^{-2t}$. It remains to estimate $\mathbb{E}Z$.

$$\begin{aligned}
\mathbb{E}\sup_{g \in CH\mathcal{F}} \left|\frac{1}{n}\sum \varphi_\theta(Y_j g(X_j)) - \mathbb{E}\varphi_\theta(Yg(X))\right| &\leq 2\mathbb{E}\sup_{g \in CH\mathcal{F}} \left|\frac{1}{n}\sum \varepsilon_j \varphi_\theta(Y_j g(X_j))\right| \quad \text{(Symmetrization)} \\
&\leq \frac{4}{\theta}\mathbb{E}\sup_{g \in CH\mathcal{F}} \left|\frac{1}{n}\sum \varepsilon_j Y_j g(X_j)\right| \quad \text{(Contraction)} \\
&= \frac{4}{\theta}\mathbb{E}\sup_{g \in CH\mathcal{F}} \left|\frac{1}{n}\sum \tilde{\varepsilon}_j g(X_j)\right|
\end{aligned}$$

Fact: max and min of linear function over a convex set are attained at extreme points; therefore

$$\leq \frac{4}{\theta}\mathbb{E}\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum\varepsilon_j f(X_j)| \leq \frac{K}{\theta}\sqrt{\frac{V}{n}}.$$

The bound can be improved to something like

$$\frac{1}{\theta}\left(\frac{n}{V}\right)^{-\frac{1}{2}\frac{2+V}{1+V}} + \frac{t}{n}.$$

Assume that γ-weak learnability assumption holds, meaning that for all $w_1,\dots,w_n \geq 0$, $\sum w_i = 1$, there exists $f \in \mathcal{F}$ such that

$$\sum w_j I\{Y_j \neq f(X_j)\} \leq 1/2 - \gamma.$$

Theorem 2. Let $\widehat{g}_T$ be the output of Adaboost after T steps. Then for all $\theta > 0$,

$$\frac{1}{n}\sum_{j=1}^n I\{Y_j\widehat{g}_T(X_j) \leq \theta\} \leq 2^T((1-2\gamma)^{1-\theta}(1+2\gamma)^{1+\theta})^{T/2}.$$

Remark 1. When $\theta > 0$, the RHS converges to

$$[(1/2-\gamma)(1/2+\gamma)]^{T/2} \to 0$$

exponentially fast with T. Therefore for θ small enough, the RHS is bounded by $C(\theta,\gamma)^T$, where $C(\theta,\gamma) < 1$ (i.e. if training error is close to 0, generalization error will be close to 0).

Sketch of proof: Assume that $Y\widehat{g}_T(X) \leq \theta$. Then

$$Y\sum\alpha_j f_j(X) \leq \theta\sum\alpha_j$$
$$e^{-Y\sum\alpha_j f_j(X)+\theta\sum\alpha_j} \geq 1.$$

Next,

$$\frac{1}{n}\sum I\{Y_j\widehat{g}_T(X_j) \leq \theta\} = \frac{1}{n}\sum I\{Y_j\widehat{g}_T(X_j) - \theta \leq 0\} \leq \frac{1}{n}\sum_j e^{\theta\sum_i\alpha_i - Y_j\sum_i\alpha_i f_i(X_j)}$$

The rest proceeds as before in lecture on Adaboost. □

Assume that $(X,Y) \in S\times\mathbb{R}$ is a random couple from distribution P. Goal: predict Y based on X. Best possible predictor is $\eta(X) = \mathbb{E}(Y|X)$;

$$\eta(X) = \operatorname*{argmin}_{f:S\to\mathbb{R}} \mathbb{E}(Y-f(X))^2.$$