Math 547: Mathematical Foundations of Statistical Learning Theory Fall 2017

Lecture 30 — October 30

Instructor: Stas Minsker *Scribe: Mose Wintner*

30.1 Regression

$(X,Y) \in S \times \mathbb{R}$ has distribution P; $(X_1,Y_1),\dots,(X_n,Y_n)$ are iid copies of (X,Y). Goal: predict Y based on X

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

Assume that $|Y| \le 1$ almost surely and let $\mathcal{F} = \{f : S \to [-1,1]\}$ be a convex set of functions. Now suppose we wish to minimize over $\mathcal{F}$ and consider

$$\widehat{f}_n = \operatorname*{argmin}_{f\in\mathcal{F}} P_n(y-f(X))^2 = \operatorname*{argmin}_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{j=1}^n (Y_j - f(X_j))^2\right].$$

What is a bound for $\mathcal{E}(\widehat{f}_n)$? Note that

$$\mathcal{E}(\widehat{f}_n) = \mathbb{E}(Y-\widehat{f}_n(X))^2 - \mathbb{E}(Y-\eta(X))^2 = \mathbb{E}\widehat{f}_n^2(X) - \mathbb{E}\eta^2(X) - 2\mathbb{E}[Y\widehat{f}_n(X)] + 2\mathbb{E}[Y\eta(X)].$$

Next,

$$\mathbb{E}[Y\widehat{f}_n(X)] = \mathbb{E}\mathbb{E}[Y\widehat{f}_n(X)|X] = \mathbb{E}\widehat{f}_n(X)\eta(X)$$
$$\mathbb{E}Y\eta(X) = \mathbb{E}\mathbb{E}[Y\eta(X)|X] = \mathbb{E}\eta^2(X)$$

so that, reasonably,

$$\mathcal{E}(\widehat{f}_n) = \mathbb{E}(\eta(X) - \widehat{f}_n(X))^2.$$

Furthermore, let $\overline{f} = \operatorname{argmin}_{f\in\mathcal{F}} \mathbb{E}(Y-f(X))^2$. Then

$$\mathcal{E}(\widehat{f}_n) = \mathbb{E}(\widehat{f}_n(X) - \overline{f})^2 + \mathbb{E}(\eta(X) - \overline{f}(X))^2.$$

The second term depends only on $\mathcal{F}$. Recall P to denote the expectation against distribution P and P_n the empirical expectation. Then the first term is equal to

$$\begin{aligned}\mathcal{E}(\widehat{f}_n) - \mathcal{E}(\overline{f}) &= P[y-\widehat{f}_n(x))]^2 - P_n[y-\widehat{f}_n(x))^2\\ &\quad + P_n[y-\widehat{f}_n(x)]^2 + P_n(y-\overline{f}(x))^2\\ &\quad - P[y-\overline{f}(x)]^2 - P_n[y-\overline{f}(x)]^2\\ &\le 2\sup_{f\in\mathcal{F}} |[P-P_n](y-f(x))^2| = 2Z((X_1,Y_1),\dots,(X_n,Y_n)).\end{aligned}$$

First, Z satisfies the bounded difference property with $c_j = 8/n$.

$$|(y_1 - f(x_1))^2 - (y_2 - f(x_2))^2| \leq 8.$$

Therefore

$$Z \leq \mathbb{E}Z + \sqrt{\frac{t}{n}}$$

with probability $\geq 1 - e^{-t^2/32}$. It remains to estimate $\mathbb{E}Z$. Note that by contraction inequality

$$\mathbb{E}\phi\left(\frac{1}{2}\sup_t\left|\frac{1}{n}\sum_{j=1}^n \varepsilon_j\varphi_j(t_j)\right|\right) \leq \mathbb{E}\phi\left(\sup_t\left|\frac{1}{n}\sum_{j=1}^n \varepsilon_j t_j\right|\right).$$

Then we estimate

$$\begin{aligned}
\mathbb{E}Z &= \mathbb{E}\sup_{f\in\mathcal{F}}|[P - P_n](y - f(x))^2| \\
&\leq 2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^n \varepsilon_j(Y_j - f(X_j))^2\right| \qquad \text{(Symmetrization)}
\end{aligned}$$

For $\varphi_j(f) := (y_j - f)^2 - y_j^2$, we have

$$|\varphi_j(f) - \varphi_j(g)| \leq 4|f - g|.$$

Then continuing,

$$2\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^n \varepsilon_j(Y_j - f(X_j))^2 - Y_j^2 + Y_j^2\right| \leq 16\mathbb{E}\sup_f\left|\frac{1}{n}\sum_{j=1}^n \varepsilon_j f(X_j)\right| + 2\mathbb{E}\left|\frac{1}{n}\sum_{j=1}^n \varepsilon_j Y_j^2\right| = I + II.$$

To estimate II, note it's

$$\leq 2\sqrt{\mathbb{E}\left(\frac{1}{n}\sum \varepsilon_j Y_j^2\right)^2} = 2\sqrt{\mathbb{E}_Y\left(\frac{1}{n^2}\sum Y_j^4\right)} \leq \frac{2}{\sqrt{n}}.$$

Then

$$I = 16\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum \varepsilon_j f(X_j)\right|.$$

Assume that either (a) $\mathcal{F}$ is the convex hull of G where G is a VC subgraph of VC dimension V (e.g. a finite collection of functions) or (b) $\mathcal{F}'$ is a d dimensional space of functions containing $\mathcal{F}$. Then

$$16\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum \varepsilon_j f(X_j)\right| \leq 16\mathbb{E}\sup_{g\in G}\left|\frac{1}{n}\sum \varepsilon_j g(X_j)\right| \leq K\sqrt{\frac{V(G)}{n}}$$

in case (a) and $\leq K'\sqrt{\frac{d}{n}}$ in case (b). Finally, we obtain the bound

$$\mathcal{E}(\widehat{f}_n) - \mathcal{E}(\overline{f}) = O(n^{-1/2})$$

which is frequently $O(n^{-1})$.