Math 547: Mathematical Foundations of Statistical Learning Theory **Fall 2017**

Lecture 5 — August 30

Instructor: Stas Minsker *Scribe: Mose Wintner*

$(X_1, Y_1), \ldots, (X_n, Y_n)$ iid from P.

Goal: find $g : \mathbb{S} \to \{\pm 1\}$ **such that** $P(Y \neq g(X)) = L(g)$ **is small.**

Let G be some "base class" of possible g.

Let

$$\widehat{g}_n = \operatorname*{argmin}_{g \in G} P_n I\{Y \neq g(X)\} := \operatorname*{argmin}_{g \in G} \frac{1}{n} \sum_{j=1}^{n} I\{Y_j \neq g(X_j)\} \approx P(Y \neq g(X)).$$

We usually are making empirical observations, so G can't be too large.

Example 1. "Decision stumps". $\mathbb{S} = \mathbb{R}$, $g_t^+(x) := I\{x \geq t\} - I\{x < t\}$, similar g_t^-, $G = \{g_t^+, g_t^- : t \in \mathbb{R}\}$. Then it is enough to consider $g_{X_{(i)}}^{\pm 1}$ for order statistics $X_{(i)}$. As $n \to \infty$ we get convergence.

Example 2. Higher-dimensional decision stumps. $\mathbb{S} = \mathbb{R}^d$. Consider decision stumps for each coordinate, e.g. $d = 2$. $g_{t,1}^+(x) = I\{x_1 \geq t\} - I\{x_1 < t\}$, $g_{t,2}^- = I\{x_2 \leq t\} - I\{x_2 > t\}$.

For binary classifiers, $P(Y \neq g(X)) = P(Yg(X) \leq 0)$, where $Yg(X)$ is called the margin. These are equal to $\mathbb{E}I\{Yg(x) \leq 0\} \leq \ell(Yg(X))$ for some functions ℓ. We'll choose $\ell(t) = e^{-t}$, the "classification-calibrated" loss, so we're now considering $\mathbb{E}e^{-Yg(X)}$.

Lemma 1. Let $\overline{g} = \operatorname{argmin}_g \mathbb{E}e^{-Yg(X)}$. Then

$$\overline{g} = \mathbb{E}[\mathbb{E}e^{-Yg(X)}|X] = \int e^{-g(X)} \frac{1 + \eta(x)}{2} + e^{g(X)} \frac{1 - \eta(x)}{2} \, d\Pi.$$

Let $g(x) = t$; then minimizing the integrand over $t \in \mathbb{R}$ gives

$$t = \frac{1}{2} \log \frac{1 + \eta(x)}{1 - \eta(x)}.$$

Then the sign of $g(x)$ is the same as the sign of $\eta(x)$. Therefore $\operatorname{sign}(\overline{g})$ is equal to the Bayes classifier. Next goal:

$$\mathbb{E}\ell(Yg(X)) = P\ell(yg(x)) \approx P_n \ell(yg(x)) = \frac{1}{n} \sum_{j=1}^{n} e^{-Y_j g(X_j)}.$$