Math 547: Mathematical Foundations of Statistical Learning Theory

Lectures 3 — Week 5

Instructor: Stas Minsker *Scribe: S. Minsker*

1.1 Sub-Gaussian random variables

Let $(X_1, Y_1), ...(X_n, Y_n)$ be the training data, and let $\mathcal{F}$ be a collection of binary classifiers. The general question that we are trying to answer is the following: assume that for some binary classifier f, the training error $\frac{1}{n}\sum_{i=1}^n \mathbb{I}\{Y_i \neq f(X_i)\}$ is small. When can we conclude that $\Pr(Y \neq f(X))$ is also small? The problem is that g is usually random (training data-dependent), and we will require uniform bounds for the differences between the empirical errors and their population versions, namely, we need to construct general bounds for

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{Y_i\neq f(X_i)\}} - \Pr(Y \neq f(X))\right|.$$

Here is a more rigorous version of the statement. Define

$$\widehat{f}_n = \operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{j}^{n} I\left\{Y_j \neq f(X_j)\right\} \text{ and } \overline{f} = \operatorname*{argmin}_{f\in\mathcal{F}} \mathbb{E} I\left\{Y \neq f(X)\right\}.$$

Moreover, let $f_*(\cdot) = \mathrm{sign}(\eta(x))$ be the Bayes classifier, where $\eta(x) = \mathbb{E}(Y|X = x)$ is the regression function.

Recall that the excess risk of $f \in \mathcal{F}$ is $\mathcal{E}(f) = \Pr(Y \neq f(X)) - P(Y \neq f_*(X))$. Note that

$$\begin{aligned}
\mathcal{E}(\widehat{f}) &= \Pr\left(Y \neq \widehat{f}_n(x)\right) - \Pr(Y \neq f_*(x)) \\
&= \Pr\left(Y \neq \widehat{f}_n(x)\right) - \Pr\left(Y \neq \overline{f}(x)\right\} + \mathcal{E}(\overline{f}) \pm \frac{1}{n}\sum_{j=1}^n I\left\{Y_j \neq \overline{f}(X_j)\right\} \pm \frac{1}{n}\sum_{j=1}^n I\left\{Y_j \neq \widehat{f}_n(X_j)\right\}.
\end{aligned}$$

By the definition of $\hat{f}_n$,

$$\frac{1}{n}\sum_{j=1}^n I\left\{Y_j \neq \widehat{f}(X_j)\right\} - \frac{1}{n}\sum_{j}^n I\left\{Y_j \neq \overline{f}(X_j)\right\} \leq 0,$$

hence

$$
\begin{aligned}
\mathcal{E}(\widehat{f}) \leq \mathcal{E}(\overline{f}) &+ \left|\mathbb{E}\Pr\left(Y \neq \widehat{f}_n(X)\right) - \frac{1}{n}\sum_{j=1}^{n} I\left\{Y_j \neq \widehat{f}(X_j)\right\}\right| \\
&+ \left|\frac{1}{n}\sum_{j=1}^{n} I\left\{Y_j \neq \overline{f}(X_j)\right\} - \Pr\left(Y \neq \overline{f}(X)\right)\right| \\
\leq & \mathcal{E}(\overline{f}) + 2\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^{n} I\left\{Y_j \neq f(X_j)\right\} - P(Y \neq f(X))\right|.
\end{aligned}
$$

Previously, we saw that the control of the excess risk is related to the concept of "ε-representativeness" of the training data which itself implies that the supremum in the inequality above is small with high probability.

Next, we will introduce and study the class of *sub-Gaussian random variables* that turns to offer just the right tools for our problem.

Definition 1. A random variable X is sub-Gaussian with parameter σ^2 (we will write $X \in \mathrm{SG}(\sigma^2)$)) if

$$\mathbb{E}e^{\lambda x} \leq e^{\frac{\lambda^2\sigma^2}{2}}$$

for all $\lambda \in \mathbb{R}$.

Example 1. If $X \sim N(0, \sigma^2) \implies \mathbb{E}e^{\lambda x} = e^{\frac{\lambda^2\sigma^2}{2}}$, hence $X \in \mathrm{SG}(\sigma^2)$.

Lemma 1.

1. If X is $\mathrm{SG}(\sigma^2) \Rightarrow -X \in \mathrm{SG}(\sigma^2)$.
2. If $X \in \mathrm{SG}(\sigma^2)$, then $\mathbb{E}X = 0$ and $\mathrm{Var}(X) \leq \sigma^2$.

Proof. Indeed, $\varphi(\lambda) = \mathbb{E}e^{\lambda x}$, then $\mathbb{E}X = \varphi'(0)$,

$$\mathbb{E}X = \lim_{t\to 0}\frac{\varphi(t) - \varphi(0)}{t} \leq \lim_{t\to 0}\frac{e^{\frac{t^2\sigma^2}{2}} - 1}{t} = 0.$$

Similarly, since $-X$ is sub-Gaussian, $\mathbb{E}(-X) \leq 0 \Rightarrow \mathbb{E}X = 0$. The bound for the variance is left as an exercise. □

Example 2. Let X be a Rademacher random variable (a "random sign"), meaning that

$$X = \begin{cases} 1, & \text{with probability } 1/2, \\ -1, & \text{with probability } 1/2. \end{cases}$$

Then $x \in SG(1)$.

Proof. Homework exercise. □

Example 3. Let X be such that $\mathbb{E}X = 0$, and $a \leq X \leq b$ almost surely for some $a \leq 0, b \geq 0$. Then $X \in SG(\frac{(b-a)^2}{4})$.

Proof. First, we reduce the problem to a r.v. that takes two values a and b. Note that $f(x) = e^{\lambda x}$ is convex. Since $X = \underbrace{\frac{b-X}{b-a}}_{=\alpha} \cdot a + \underbrace{\frac{X-a}{b-a}}_{=1-\alpha} \cdot b$,

$$e^{\lambda X} = e^{\lambda(\alpha a + (1-\alpha)b)} \le \alpha e^{\lambda a} + (1-\alpha)e^{\lambda b},$$

and $\mathbb{E}e^{\lambda x} \le e^{\lambda a} \cdot \frac{b}{b-a} + \frac{-a}{b-a} e^{\lambda b}$. That is exactly the MGF of a r.v. Y s.t.

$$Y = \begin{cases} a, \text{ with probability } \frac{b}{b-a}, \\ b, \text{ with probability } \frac{-a}{b-a}. \end{cases}$$

Let $p = \frac{b}{b-a}$, $1 - p = \frac{-a}{b-a}$ and $h = \lambda(b-a)$. It follows that

$$\begin{aligned} \mathbb{E}e^{\lambda x} &\le p e^{-h(1-p)} + (1-p)e^{ph} \\ &= e^{ph}((1-p) + pe^{-h}) = e^{F(h)}, \end{aligned}$$

where $F(h) = ph + \log(1 - p + pe^{-h})$. Note that

$$\begin{aligned} F'(h) &= p + \frac{-pe^{-h}}{1-p+pe^{-h}}, \quad F'(0) = 0, \\ F''(h) &= \frac{pe^{-h}(1-p+pe^{-h}) - (pe^{-h})^2}{(1-p+pe^{-h})^2} \\ &= \frac{pe^{-h}}{1-p+pe^{-h}} \left(1 - \frac{pe^{-h}}{1-p+pe^{-h}}\right) \le \frac{1}{4}, \end{aligned}$$

since $z(1-z) \le 1/4$ for $z \in [0,1]$. Hence it follows from Taylor's expansion at 0 that $F(h) \le \frac{h^2}{8}$, and result easily follows. $\square$

The following *tail bound* is one of the key properties of sub-Gaussian random variables.

Proposition 1. Assume that $X \in \mathrm{SG}(\sigma^2)$. Then $\Pr(X \ge t) \le e^{\frac{-t^2}{2\sigma^2}}$ and $\Pr(X \le -t) \le e^{\frac{-t^2}{2\sigma^2}}$ for any $t \ge 0$.

Proof. For any $\lambda > 0$,

$$\begin{aligned} \Pr(X > t) = \Pr(\lambda X > \lambda t) &= \Pr(e^{\lambda X} > e^{\lambda t}) \\ &\le \frac{\mathbb{E}e^{\lambda X}}{e^{\lambda t}} \le e^{\frac{\lambda^2\sigma^2}{2} - \lambda t}. \end{aligned}$$

Since $\lambda > 0$ was arbitrary, we can minimizing $\frac{\lambda^2\sigma^2}{2} - \lambda t$ to minimum over λ to get

$$\Pr(X > t) \le e^{-\frac{t^2}{2\sigma^2}}.$$

$\square$

Corollary 1. $\Pr(|X| \geq t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$.

Lemma 2. Let $X_1 \in \mathrm{SG}(\sigma_1^2)$, $X_2 \in \mathrm{SG}(\sigma_2^2)$,..., $X_n \in \mathrm{SG}(\sigma_n^2)$ be independent. Then $\sum_{j=1}^n X_j \in \mathrm{SG}(\sigma_1^2 + \sigma_2^2 + ...\sigma_n^2)$.

Proof. $\mathbb{E}e^{\lambda \sum_{j=1}^n X_j} = \mathbb{E}\Pi_{j=1}^n e^{\lambda X_j} = \Pi_{j=1}^n \mathbb{E}e^{\lambda X_j} \leq e^{\frac{\lambda^2}{2} \sum_{j=1}^n \sigma_j^2}$. $\square$

Theorem 1. (Hoeffding's inequality) Let $X_1, ..., X_n$ be independent variables such that $a_j \leq X_j - \mathbb{E}X_j \leq b_j$ a.s. for all j, Then

$$\Pr\left(\left|\sum_{j=1}^n (X_j - \mathbb{E}X_j)\right| \geq t\right) \leq e^{-\frac{2t^2}{\sum_{j=1}^n (b_j - a_j)^2}}.$$

Proof. Result immediately follows from Lemma 2 and Corollary 1. Indeed,

$$\sum_{j=1}^n (X_j - \mathbb{E}X_j) \in \mathrm{SG}(\Sigma^2),$$

where $\Sigma^2 = \frac{\sum_{j=1}^n (b_j - a_j)^2}{4}$. $\square$