Math 547: Mathematical Foundations of Statistical Learning Theory Fall 2017

Lecture 11 — September 15

Instructor: Stas Minsker *Scribe: Mose Wintner*

11.1 Examples

The function $f \in C^0[0,\infty) \cap C^\infty(0,\infty)$ is called completely monotone if $(-1)^k f^{(k)}(r) \geq 0$ for all $r > 0$, $k \in \mathbb{N}$.

Lemma 1. f is a completely monotone function if and only if $f(||\cdot||_2^2)$ is positive definite. Then $K(x,y) = f(||x-y||^2)$ is a kernel.

Examples of kernels:

a) Gaussian kernel: $K(x,y) = e^{-\frac{||x-y||_2^2}{2\sigma^2}}$
b) Cauchy(?) kernel: $K(x,y) = \frac{1}{c+||x-y||_2^2}\alpha$, $\alpha > 0$, $c \neq 0$.
c) Linear kernel: $K(x,y) = \langle x,y\rangle$.
d) $K(x,y) = (a\langle x,y\rangle + 1)^d$, $a \in \mathbb{R}$, $d \in \mathbb{N}$.
e) Laplacian kernel: $e^{-\frac{||x-y||_2}{\sigma}}$, $\sigma > 0$

11.2 The subject coming after RKHS

Question: Assume that for some binary classifier $\tilde{f}$, the training error

$$\frac{1}{n}\sum_{j=1}^{n} I\{Y_j \neq \tilde{f}(X_j)\}$$

is small. When can we conclude that $P(Y \neq \tilde{f}(X))$ is also small? In other words, we want to construct general bounds for the difference between the generalization and training errors:

$$\left|\frac{1}{n}\sum_{j=1}^{n} I\{Y_j \neq \tilde{f}(X_j)\} - P(Y \neq \tilde{f}(X))\right|.$$

Usually $\tilde{f}$ itself is also random, so this doesn't necessarily go to 0 by LLN.

11.2.1 Sub-Gaussian random variables

Definition 1. X is a sub-Gaussian random variable with parameter σ^2 (written $X \in SG(\sigma^2)$) if $\mathbb{E}e^{\lambda X} \leq e^{\frac{\lambda^2\sigma^2}{2}}$ for all $\lambda \in \mathbb{R}$.

Remark 1. 1. If $X \sim N(0,\sigma^2)$, then $\mathbb{E}e^{\lambda X} = e^{\lambda^2\sigma^2/2}$.

2. If X is $SG(\sigma^2)$, then $-X \in SG(\sigma^2)$.

3. If $X \in SG(\sigma^2)$, then $\mathbb{E}X = 0$. Indeed, if $\phi(\lambda) = \mathbb{E}e^{\lambda X}$, then

$$\mathbb{E}X = \phi'(0) = \lim_{t\to 0} \lim \frac{\phi(t) - \phi(0)}{t} \leq \lim_{t\to 0} \frac{e^{t^2\sigma^2/2} - 1}{t} = 0.$$

   Similarly, $\mathbb{E}(-X) = 0$.

Example: Let X be a Rademacher random variable, meaning that $X = \pm 1$ with probabilities $1/2$. Then $X \in SG(1)$:

$$\mathbb{E}e^{\lambda x} = \frac{1}{2}e^{\lambda} + \frac{1}{2}e^{-\lambda} = \cosh(\lambda) = \frac{1}{2}\sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} = \sum \frac{\lambda^{2k}}{2^k}\frac{2^k}{(2k)!} \leq \left(\frac{\lambda^2}{2}\right)^k \frac{1}{k!} \leq \sum \frac{(\lambda^2/2)^k}{k!} = e^{\lambda^2/2}$$

since $2^k/[(k+1)\cdots(2k)] \leq 1$.

Example: Let X be such that $\mathbb{E}X = 0$, $a \leq X \leq b$ almost surely for some $a \leq 0$, $b \geq 0$. Then $X \in SG((b-a)^2/n)$. To see this, first reduce to a random variable that takes two values a and b. We know that $f(x) = e^{\lambda x}$ is convex; represent $x = \frac{b-x}{b-a}a + \frac{x-a}{b-a}b$ assuming WLOG $a > b$. Then

$$e^{\lambda x} = e^{\lambda(\alpha a + (1-\alpha)b} \leq \alpha e^{\lambda a} + (1-\alpha)e^{\lambda b}.$$

therefore

$$\mathbb{E}e^{\lambda x} \leq e^{\lambda a}\frac{b}{b-a} + \frac{-a}{b-a}e^{\lambda b},$$

which is the MGF of a random variable Z which is a with probability $b/(b-a)$ and is b with probability $-a/(b-a)$.

$$e^{\lambda a}\frac{b}{b-a} + e^{\lambda b}\frac{-a}{b-a} = e^{-\lambda(1-p)(b-a)}p + (1-p)e^{p(b-a)}.$$

Maximizing this function with respect to $p \in (0,1)$.