Math 547: Mathematical Foundations of Statistical Learning Theory Fall 2017

Lecture 25 — October 18

Instructor: Stas Minsker | Scribe: Mose Wintner

25.1 "Learnable" distributions

Theorem 1. $\mathcal{C}$ a VC class of VC dimension V, $\mathcal{A}$ a class of distributions that are "learnable" with $\mathcal{C}$ (see previous lecture). Then for any classifier $g_n(\cdot) = g_n(\cdot, (X_1, Y_1), \ldots, (X_n, Y_n))$, we have

$$\sup_{(X,Y)\sim P\in\mathcal{A}} P(Y \neq \hat{g}_n(X)) \geq \frac{V-1}{2en}(1 - \frac{1}{n}).$$

Proof: Let $x_1, \ldots, x_V$ be any set of points shattered by $\mathcal{C}$. Consider the following family of distributions: $X = x_i$ with probability $1/n$ for $i = 1, \ldots, V-1$, $X = x_V$ with probability $1-(V-1)/n$; and $Y = f_b(X) = b_i$ if $X = x_i$ for $i = 1, \ldots, V-1$ and -1 if $i = V$, where $b = (b_1, \ldots b_{V-1}) \in \{\pm 1\}^{V-1}$. This gives a family of 2^{V-1} distributions. Note that $\inf_{C\in\mathcal{C}} P(Y \neq g_C(X)) = 0$ where $g_C(X) = (-1)^{X\notin C}$. Then $\mathcal{A}' \subset \mathcal{A}$. Hence

$$\sup_{\mathcal{A}} P(Y \neq g_n(X)) \geq \sup_{\mathcal{A}'} P(Y \neq g_n(x)) = \sup_{b} P(f_b(X) \neq g_n(X)) \geq 2^{1-V} \sum_{b\in\{\pm1\}^{V-1}} P(f_b(X) \neq g_n(X))$$

by estimating the minimax risk by the Bayes risk from below (i.e. assuming b is random, estimating supremum from below by the average). This is then equal to $P = (f_B(X) \neq g_n(X))$, where $B \sim U\{\pm1\}^{V-1}$ a discrete uniform. This is equal then to

$$\int_{g_n(X)\neq\eta(X)} |\eta(X)|\, d\Pi.$$

Note that $\eta(x_1) = \cdots = \eta(x_{v-1}) = 0$ Then the probability of the risk is

$$\geq \frac{1}{2} P(X \neq x_1, \ldots, X \neq x_n, X \neq x_v),$$

because if we don't know how to classify X_i (i.e. it's not equal to any training data, and not equal to x_v, which we know to classify as -1), we set the probability of assigning 1 equal to 1/2...in other words, the probability of our g_n making an error is greater than half the probability of not knowing how to classify X. This is

$$\begin{aligned}
&= \frac{1}{2} P(\cup_{j=1}^{V-1} \{X = x_j, X_i \neq x_j, i = 1, \ldots, n\}) \\
&= \frac{1}{2} \sum_{j=1}^{V-1} P(X = x_j, X_1 \neq x_j, \ldots, X_n \neq x_j) \\
&= \frac{1}{2} \sum_{j=1}^{V-1} \frac{1}{n} \left(1 - \frac{1}{n}\right)^{v-1} \\
&= \frac{1}{2} \frac{V-1}{n} \left(1 - \frac{1}{n}\right)^{n-1} \left(1 - \frac{1}{n}\right) \\
&\geq \frac{V-1}{2en} \left(1 - \frac{1}{n}\right).
\end{aligned}$$

Bounded Difference inequality (McDiarmid's inequality): A uniform version of Hœffding's inequality. Let $Z = Z(x_1, \ldots, x_n)$ be a random variable. when does Z concentrate around $\mathbb{E}Z$?

Bounded difference condition (BDC): for all $1 \leq j \leq n$,

$$|Z(x_1, \ldots, x_j \ldots, x_n - Z(x_1, \ldots, x_j', \ldots, x_n)| \leq c_j$$

for any $x_1, \ldots, x_j, x_j', \ldots, x_n$.

Theorem 2. Assume that $X_1, \ldots, X_n$ are independent, and that the BDC holds. Then

$$P(|Z - \mathbb{E}Z| \geq t) \leq 2\exp\left(\frac{-2t^2}{\sum c_j^2}\right).$$

Example 1. $Z(x) = \sum x_j$,

$$Z(x_1, \ldots, x_j, \ldots, x_n) - Z(x_1, \ldots, x_j', \ldots, x_n) = x_j - x_j'.$$

If $|x_j| \leq M$, $|x_j - x_j'| \leq 2M$, which recovers Hœffding's inequality.

Example 2. $Z = ||P_n - P||_{\mathcal{C}} = \sup_{C \in \mathcal{C}} \left| \frac{1}{n} \sum_{j=1}^{n} I\{X_j \in C\} - P(C) \right|$. Then Z has the bounded difference property with $c_j = 2/n$. Therefore

$$P\left(||P_n - P||_{\mathcal{C}} - \mathbb{E}||P_n - P||_{\mathcal{C}} \geq \frac{t}{\sqrt{n}}\right) \leq \exp(-2t^2).$$