Concentration of Measure Reading Course: Lecture 9

11/12/2019

We introduce a classical result in empirical processes this week. The result is concerned with bounding the expected maximum magnitude of a subgaussian process of sincere interest in statistics. We will first state the main theorem and give an overview of the proof, then discuss some of its applications and consequences.

Theorem 1. Let \( F \) be a class of Boolean functions on \( \Omega \). Let \( X_1, \ldots, X_n \) be iid random variables with the same distribution as \( X \). Then,

\[
\mathbb{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X) \right| \leq C \sqrt{\frac{\text{vc}(F)}{n}},
\]

where \( \text{vc}(F) \) is the Vapnik-Chervonenkis (VC) dimension of \( F \) and defined by the cardinality of the largest subset \( A \) of \( \Omega \) that can be shattered by \( F \), i.e., restriction of the support of all functions in \( F \) to \( A \) yields the power set of \( P(A) \) of \( A \).

Proof. First note that by symmetrization,

\[
\mathbb{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X) \right| \leq 2 \mathbb{E} \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right|,
\]

where \( \varepsilon_i \) are independent Rademacher random variables. On the other hand, applying Khintchine’s inequality to the sum of independent subgaussians (bounded random variables are subgaussians),

\[
\left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right\|_{\psi_2} \lesssim \frac{1}{\sqrt{n}} \left( \frac{1}{n} \sum_{i=1}^{n} f(X_i)^2 \right)^{1/2},
\]

where \( \| \cdot \|_{\psi_2} \) is the subgaussian norm. Therefore, conditional to \( X_1, \ldots, X_n, \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \) is a \( 1/\sqrt{n} \)-subgaussian process indexed by \( F \) and with respect to the metric \( L_2(\mu_n) \), where \( \mu_n \) is the empirical measure given by \( X_1, \ldots, X_n \). Using Dudley’s chaining inequality (with \( f \equiv 0 \) added into \( F \)),

\[
\mathbb{E}_{\varepsilon_1, \ldots, \varepsilon_n} \sup_{f \in F} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \lesssim \frac{1}{\sqrt{n}} \int_{0}^{\infty} \sqrt{\log N(F, L_2(\mu_n), \varepsilon)} \, d\varepsilon,
\]

where \( N \) is the notation for covering number. The proof is complete by noting that the metric entropy term can be further bounded by the VC-dimension of \( F \) as follows:

\[
N(F, L_2(\mu_n), \varepsilon) \leq \left( \frac{2}{\varepsilon} \right)^{C \cdot \text{vc}(F)} , \tag{0.1}
\]

where \( C \) is some constant. \( \text{vc}(F) \) is a combinatorial attribute which does not depend on the metric. \(\square\)
Remark 2. The proof of (0.1) consists of several ingredients. We may assume in the general context with $\mu_n = \mu$ being any probability measure.

- $\mathcal{N}(\mathcal{F}, L_2(\mu), \varepsilon)$ is bounded by the cardinality of largest subset of $\mathcal{F}$ in which all elements are $\varepsilon$-separated, denoted by $\mathcal{F}_\varepsilon$.
- $\mu$ may have a continuous support. We may use a union bound argument to find an empirical measure whose support size is less than $C \varepsilon^{-d} \log |\mathcal{F}_\varepsilon|$ such that restriction of the elements in $\mathcal{F}_\varepsilon$ to $\text{supp}(\mu')$ is an injection, i.e., $|\mathcal{F}_\varepsilon| = |\mathcal{F}_\varepsilon \cap \text{supp}(\mu')|$.
- In the discrete world, we can bound $|\mathcal{F}_\varepsilon \cap \text{supp}(\mu')|$ by its shattering ability, which is measured by $vc(\mathcal{F})$:

$$|\mathcal{F}_\varepsilon \cap \text{supp}(\mu')| \leq \sum_{k=0}^{vc(\mathcal{F})} \binom{|\text{supp}(\mu')|}{k} \leq \left( \frac{e|\text{supp}(\mu')|}{vc(\mathcal{F})} \right)^{vc(\mathcal{F})}. \tag{0.2}$$

Inequality (0.2) is known as the Sauer-Shelah Lemma.

Remark 3. Taking $\Omega = \mathbb{R}^d$, $\mathcal{F} = \{I_{S,\sigma}\}_{\sigma=1}^{(\infty,x_1), \cdots, x_d} \in \mathbb{R}^d}$, and noting that $vc(\mathcal{F}) \leq d$ yields the classical Glivenko-Cantelli Theorem. Indeed, for $n$ points $z_1, \cdots, z_n$ in $\mathbb{R}^d$, $z_i$ can be separated from the rest points if and only if one of its coordinates is the smallest. But this cannot happen if $n \geq d + 1$.

Remark 4. The VC-dimension of the indicator functions on the sets of interest is as follows:

- Interval $[a, b]$ on real line: $vc(\mathcal{F}) = 2$;
- Disks in $\mathbb{R}^2$: $vc(\mathcal{F}) = 3$;
- Closed balls in $\mathbb{R}^d$: $vc(\mathcal{F}) \leq d + 2$;
- Rectangles in $\mathbb{R}^d$: $vc(\mathcal{F}) = 2d$;
- Half spaces in $\mathbb{R}^d$: $vc(\mathcal{F}) = d + 1$;
- Convex polygons with $d$ vertices: $vc(\mathcal{F}) = 2d + 1$.

We next mention an application of Theorem 1 in statistical learning. Consider the problem of learning the relationship between the response variable $Y$ and the covariate $X$ given independent samples $(X_1, Y_1), \cdots, (X_n, Y_n)$, where $(X_i, Y_i) \sim (X, Y)$. When $Y$ is a binary, the problem reduces to a classification problem in supervised machine learning. In such context, we further assume that $Y = \mathbb{E}[Y|X] := T(X)$, which is an unknown Boolean function on $\Omega$. Our goal is to learn $T$ within a given subset $\mathcal{F}$ of all Boolean functions on $\Omega$, in the sense of solving

$$\min_{f \in \mathcal{F}} \mathbb{E}[f(X) - T(X)]^2. \tag{0.3}$$

Since the expectation is generally unknown, it is usually approximated by the one calculated under the empirical measure and leads to the following problem:

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} |f(X_i) - T(X_i)|^2. \tag{0.4}$$

Let $f^*$ and $f_*$ be the minimizers of (0.3) and (0.4), respectively. We wish to know how large $\mathbb{E}_X |f_*(X) - T(X)|^2$ ($f_*$ depends on $X_1, \cdots, X_n$) is compared to the theoretical bound $\mathbb{E}|f^*(X) - T(X)|^2$, which is determined by $\mathcal{F}$. Note that including too few functions in $\mathcal{F}$ has little predictive power (small variance but large bias), and including too many functions in $\mathcal{F}$ can cause overfitting (small bias but large variance). To strike a balance, we assume that $\mathcal{F}$ has reasonable combinatorial complexity and this is manifested by $vc(\mathcal{F}) < \infty$. The following theorem tells us that average empirical risk is bounded by the theoretical one plus an VC-dimension term.
Theorem 5.

\[ \mathbb{E}_{X_1, \ldots, X_n} [ f_* (X) - T(X)]^2 \leq \mathbb{E} [ f_* (X) - T(X)]^2 + C \sqrt{\frac{\text{vc}(\mathcal{F})}{n}}. \]

Proof. Note that for every \( f \in \mathcal{F} \), \((f - T)^2\) is a Boolean function on \( \Omega \). We therefore consider the set of Boolean functions given by \( \mathcal{L} = \{ (f - T)^2, f \in \mathcal{F} \} \). Directly applying Theorem 1 to \( \mathcal{L} \) is hard since an upper bound on \( \text{vc}(\mathcal{L}) \) does not follow from the fact that \( \text{vc}(\mathcal{F}) \) is finite. To get around this, we note that

\[ \mathbb{E}_{X_1, \ldots, X_n} [ f_* (X) - T(X)]^2 \leq \mathbb{E}_{X_1, \ldots, X_n} \frac{1}{n} \sum_{i=1}^{n} (f_* (X_i) - T(X_i))^2 + \mathbb{E}_{X_1, \ldots, X_n} \sup_{l \in \mathcal{L}} \left| \frac{1}{n} \sum_{i=1}^{n} l(X_i) - \mathbb{E} l(X) \right| \]

\[ \leq \mathbb{E}_{X_1, \ldots, X_n} \frac{1}{n} \sum_{i=1}^{n} (f^* (X_i) - T(X_i))^2 + \mathbb{E}_{X_1, \ldots, X_n} \sup_{l \in \mathcal{L}} \left| \frac{1}{n} \sum_{i=1}^{n} l(X_i) - \mathbb{E} l(X) \right| \]

\[ \leq \mathbb{E} [ f^* (X) - T(X)]^2 + 2 \mathbb{E}_{X_1, \ldots, X_n} \sup_{l \in \mathcal{L}} \left| \frac{1}{n} \sum_{i=1}^{n} l(X_i) - \mathbb{E} l(X) \right|. \]

The proof of Theorem 1 tells us that the second term in the last inequality above can be bounded by

\[ \mathbb{E}_{X_1, \ldots, X_n} \sup_{l \in \mathcal{L}} \left| \frac{1}{n} \sum_{i=1}^{n} l(X_i) - \mathbb{E} l(X) \right| \lesssim \frac{1}{\sqrt{n}} \mathbb{E}_{X_1, \ldots, X_n} \int_{0}^{1} \sqrt{\log \mathcal{N}(\mathcal{L}, L_2(\mu_n), \varepsilon)} \, d\varepsilon \]

\[ \lesssim \frac{1}{\sqrt{n}} \mathbb{E}_{X_1, \ldots, X_n} \int_{0}^{1} \sqrt{\log \mathcal{N}(\mathcal{F}, L_2(\mu_n), \varepsilon/4)} \, d\varepsilon \]

\[ \lesssim \sqrt{\frac{\text{vc}(\mathcal{F})}{n}}, \]

where \( \mu_n \) is the empirical measure given by \( X_1, \ldots, X_n \), and the second inequality can be checked by a direct estimate. The proof is complete. \( \square \)

Remark 6. Using a high-dimensional probabilistic version of Dudley’s inequality, it can be shown that with probability at least \( 1 - 2e^{-u^2} \) that

\[ \mathbb{E} [ f_* (X) - T(X)]^2 \leq \mathbb{E} [ f^* (X) - T(X)]^2 + C \sqrt{n (\text{vc}(\mathcal{F}) + C_1 u)}, \]

for some \( C, C_1 > 0 \). Setting \( \delta = 2e^{-u^2} \) and \( \frac{C}{\sqrt{n}} (\text{vc}(\mathcal{F}) + C_1 u) = \varepsilon \), we have

\[ n = O \left( \frac{\text{vc}(\mathcal{F}) + \log(1/\delta)}{\varepsilon^2} \right). \]

This bound is called the agnostic sample complexity of the estimator \( f_* \) for the class \( \mathcal{F} \).

VC-dimension plays a crucial role in providing a upper bound for the learnability of the empirical risk minimizer in the binary classification model in statistical learning. The concept itself can be extended to the multi-class learning to derive a similar bound for the learnability of the empirical risk minimizer using the graph dimension and the Natarajan dimension. More details can be found in [1].

References