ORIE 7790 High Dimensional Probability and Statistics Lecture 12–13 - 3/3,5/2020

Lecture 12–13: Statistical Learning Theory

Lecturer: Yudong Chen *Scribe: Miaolan Xie, Polina Alexeenko*

Reading:

- Sec 8.4 of Vershynin book,
- Sec 4.1 and 4.2 of Wainwright,
- Sec 3.3 of Duchi's notes.

In this lecture, we introduce statistical learning theory. In particular, e will study the theoretical foundations of statistical learning, including techniques for bounding the test error. We will formalize and generalize the idea that the accuracy with which we learn a function is proportional to the complexity of the function class.

1 Statistical Learning Theory

We denote by $f^* : \mathcal{X} \to [0,1]$ the unknown true regression function. We observe n data points $(X_i, f^*(X_i), i = 1, \dots, n$, where X_i is drawn i.i.d. from some unknown distribution μ. We would like to use the data to compute an estimator $\hat{f}$ of f^*.

For each function f, we define the **population risk** to be

$$L(f) = \mathbb{E}_{X \sim \mu} \left(f(X) - f^*(X)\right)^2.$$

Ideally, we want to find the population minimizer

$$f_0 \triangleq \arg\min_{f \in \mathcal{F}} L(f).$$

However, $L(f)$ is not computable. Instead, we consider the minimizing the **empirical risk**, defined as

$$\arg\min_{f \in \mathcal{F}} L_n(f) \triangleq \frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) - f^*(x_i)\right)^2.$$

This is called the **empirical risk minimization (ERM)** approach, where we use

$$\hat{f} = \arg\min_{f \in \mathcal{F}} L_n(f)$$

as our estimator of f^*.

Risk decomposition. We aim to bound the population risk of the empirical risk minimizer $\hat{f}$. The risk can be decomposed as follows:

$$\begin{aligned}
\underbrace{L\left(\hat{f}\right)}_{\text{test error}} &= \underbrace{\left[L\left(\hat{f}\right) - L_n\left(\hat{f}\right)\right]}_{\text{generalization error}} + \underbrace{L_n\left(\hat{f}\right)}_{\text{training error}} \\
&\le \left[L\left(f^*\right) - L_n\left(\hat{f}\right)\right] + L_n(f_0) \\
&= \underbrace{\left[L\left(\hat{f}\right) - L_n\left(\hat{f}\right)\right] + \left[L_n(f_0) - L(f_0)\right]}_{\text{estimation error}} + \underbrace{L(f_0)}_{\text{approximation error}} \\
&\le 2 \sup_{f \in \mathcal{F}} \left|L_n(f) - L(f)\right| + L(f_0).
\end{aligned}$$

Remark Note that the above bound, particularly the last step, may not be tight. We will consider more advanced techniques in later lectures.

Rearranging terms, we obtain the bound

$$\begin{aligned}\underbrace{L\left(\hat{f}\right) - L\left(f_0\right)}_{\text{excess risk}} &\leq 2 \sup_{f \in \mathcal{F}} \left| L_n\left(f\right) - L\left(f\right) \right| \\ &= 2 \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \left(f\left(X_i\right) - f^*\left(X_i\right)\right)^2 - \mathbb{E}\left(f\left(X\right) - f^*\left(X\right)\right)^2 \right|.\end{aligned}$$

2 Upper Bound by Rademacher Complexity

We can upper bound the right hand side by using Rademacher complexity. Assume that the functions in $\mathcal{F}$ are $[0,1]$-valued, and that f^* is also $[0,1]$ valued. Letting $g := \left(f - f^*\right)^2$, we define the function class

$$\mathcal{G} = \left\{ x \mapsto \left(f\left(x\right) - f^*\left(x\right)\right)^2 : f \in \mathcal{F} \right\}.$$

Then we can write

$$\sup_{f \in \mathcal{F}} \left| L_n\left(f\right) - L\left(f\right) \right| = \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^{n} g\left(X_i\right) - \mathbb{E}\left[g\left(X_i\right)\right] \right|. \tag{1}$$

We will focus on bounding the expectation of this supremum. The deviation of this supremum from its expectation can be bounded using concentration inequalities (McDiarmid, Lipschitz concentration, functional Hoeffding, etc.).

2.1 Symmetrization

The first step is a technique called symmetrization, which we have seen in this course and plays an important role in high-dimensional probability and statistics. Let

- $(Y_1, \ldots, Y_n)$ be an independent copy of $X_1, \ldots, X_n$,
- $(\epsilon_1, \ldots, \epsilon_n)$ be i.i.d. Rademacher random variables.

Then the expectation of the supremum in (1) can be bounded as follows:

$$\begin{aligned}\mathbb{E} \sup_{g \in \mathcal{G}} \left| \sum_{i=1}^{n} \Big[g\left(X_i\right) - \mathbb{E} g\left(X_i\right) \Big] \right| &= \mathbb{E}_X \sup_{g \in \mathcal{G}} \left| \sum_{i=1}^{n} \Big[g\left(X_i\right) - \mathbb{E}_Y[g\left(Y_i\right)] \Big] \right| \\ &\leq \mathbb{E}_X \sup_{g \in \mathcal{G}} \mathbb{E}_Y \left| \sum_{i=1}^{n} \Big[g\left(X_i\right) - g\left(Y_i\right) \Big] \right| && \text{(by Jensen's inequality)} \\ &\leq \mathbb{E}_X \mathbb{E}_Y \sup_{g \in \mathcal{G}} \left| \sum_{i=1}^{n} \Big[g\left(X_i\right) - g\left(Y_i\right) \Big] \right| && \text{(by Jensen's inequality)} \\ &= \mathbb{E}_X \mathbb{E}_Y \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left| \sum_{i=1}^{n} \epsilon_i \Big[g\left(X_i\right) - g\left(Y_i\right) \Big] \right| && \text{(by symmetry)} \\ &\leq 2 \mathbb{E}_X \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left| \sum_{i=1}^{n} \epsilon_i g\left(X_i\right) \right|. && \text{(by the triangle inequality)}\end{aligned}$$

The last right hand side is called the Rademacher complexity of $\mathcal{G}$.

Definition 1 (Rademacher Complexity)**.** *The **empirical Rademacher complexity** of $\mathcal{G}$ given X is*

$$R_n(\mathcal{G}|X) := \mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i g(X_i) \right| .$$

*The **Rademacher complexity** of $\mathcal{G}$ is*

$$R_n(\mathcal{G}) := \mathbb{E}_X \left[R_n(\mathcal{G}|X) \right] .$$

We have therefore proved the following.

Theorem 1. *We have*

$$\mathbb{E}_X \sup_{g\in\mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n \left[g\left(X_i\right) - \mathbb{E}\left[g\left(X_i\right)\right] \right] \right| \leq 2R_n\left(\mathcal{G}\right) .$$

2.2 Contraction

Our next step is to deal with the squared quantity inside the expectation, using a technique called contraction principles.

Theorem 2 (Ledoux-Talagrand Contraction Principle)**.** *Let $T \subset \mathbb{R}^n$. For each $i = 1, \ldots, n$, let $\phi_i : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz and centered (i.e., $\phi_i\left(0\right) = 0$). We then have*

$$\mathbb{E} \sup_{\theta\in T} \left| \sum_{i=1}^n \epsilon_i \phi_i\left(\theta_i\right) \right| \leq 2\mathbb{E} \sup_{\theta\in T} \left| \sum_{i=1}^n \epsilon_i \theta_i \right| ,$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables.

We will not prove Theorem 2 here, but we will prove its Gaussian analogue, where the ϵ_i's are replaced by Gaussian RVs.

Theorem 3 (Gaussian Contraction Principle)**.** *Let $T \subset \mathbb{R}^n$. For each $i = 1, \ldots, n$, let $\phi_i : \mathbb{R} \to \mathbb{R}$ be 1-Lipschitz. We then have*

$$\mathbb{E} \sup_{\theta\in T} \sum_{i=1}^n g_i \phi_i\left(\theta_i\right) \leq \mathbb{E} \sup_{\theta\in T} \sum_i g_i \theta_i ,$$

where $g_i \overset{\text{iid}}{\sim} N\left(0, 1\right)$.

Theorem 3 can be proved by Gaussian comparison inequalities.

Proof We have the suprema of two Gaussian processes indexed by $\theta \in T$:

$$X_\theta = \sum_{i=1}^n g_i \phi_i\left(\theta_i\right) \qquad \text{and} \qquad Y_\theta = \sum_{i=1}^n g_i \theta_i .$$

Comparing the increments of these two processes, we have

$$\begin{aligned} \mathbb{E}\left[\left(X_\theta - X_{\tilde{\theta}}\right)^2 \right] &= \sum_{i=1}^n \left(\phi_i\left(\theta_i\right) - \phi_i(\tilde{\theta}_i)\right)^2 && \text{because } \mathbb{E}\left[g_i^2\right] = 1 \\ &\leq \sum_{i=1}^n \left(\theta_i - \tilde{\theta}_i\right)^2 && \text{because } \phi_i \text{ is 1-Lipschitz} \\ &= \mathbb{E}\left(Y_\theta - Y_{\tilde{\theta}}\right)^2 . \end{aligned}$$

Applying the Sudakov-Fernique inequality (Lecture 7, Theorem 2), we obtain that

$$\mathbb{E} \sup_{\theta \in T} X_\theta \leq \mathbb{E} \sup_{\theta \in T} Y_\theta,$$

thereby proving the theorem. □

There is no Rademacher version of the Sudakov-Fernique inequality, so proving the Rademacher contraction is more involved. Luckily, we still have the Rademacher contraction inequality. Moreover, there are comparison inequalities between Rademacher and Gaussian processes, which we will not talk about here..

Returning to bounding the Rademacher complexity of $\mathcal{G}$, we recall that f, f^* are $[0,1]$-valued and that $g(X_i) = (f(X_i) - f^*(X_i))^2$. We now apply Theorem 2 by setting $\theta_i = f(X_i) - f^*(X_i)$ and $\phi_i(u) = u^2$. Since the domain of $f - f^*$ is $[-1, 1]$, the function ϕ_i restricted to in this domain is 2-Lipschitz. We therefore have

$$\begin{aligned} R_n(\mathcal{G}|X) &= \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \left[f(X_i) - f^*(X_i) \right]^2 \right| \\ &\leq 2\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \left[f(X_i) - f^*(X_i) \right] \right| \\ &\leq 4\mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \epsilon_i f(X_i) \right| \\ &= 4R_n(\mathcal{F}|X). \end{aligned}$$

Remark Note that this inequality is sometimes written as

$$R_n(\mathcal{F} \circ \phi) \leq 4R_n(\mathcal{F}).$$

2.3 Putting together

In summary, we have obtain an upper bound on the expected excess risk via the following steps:

$$\begin{aligned} \mathbb{E}\left[L_n\left(\hat{f}\right) - L(f_0)\right] &\lesssim \mathbb{E} \sup_{f \in \mathcal{F}} \left| L_n(f) - L(f) \right| && \text{(risk decomposition)} \\ &= \mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i}^{n} (f(X_i) - f^*(X_i))^2 - \mathbb{E}\left[(f(X) - f^*(X))^2 \right] \right| \\ &= \mathbb{E} \sup_{g \in \mathcal{G}} \left| \frac{1}{n} \sum_{i=1}^n \left[g(X_i) - \mathbb{E}\left[g(X_i) \right] \right] \right| && \text{(reparametrization)} \\ &\lesssim R_n(\mathcal{G}) && \text{(symmetrization)} \\ &\lesssim R_n(\mathcal{F}) && \text{(contraction)} \\ &:= \mathbb{E}_X \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right|. \end{aligned}$$

Note that we have bounded the supremum of one empirical process by that of another, where both processes are indexed by $f \in \mathcal{F}$. The second process, given in the definition of $R_n(\mathcal{F}) = \mathbb{E}_X[R_n(\mathcal{F}|X)]$, is often easier to control. In particular we can bound $R_n(\mathcal{F}|X)$ by conditioning on X, in which case we have a (canonical) Rademacher process

$$R_n(\mathcal{F}|X) = \mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} \left| \frac{1}{n} \langle \epsilon, f(X_1^n) \rangle \right|.$$

A variety of techniques may be used to bound the supremum of this process, including

- a union bound;
- Dudley's entropy integral bound, e.g., when $\mathcal{F}$ is Lipschitz (see Lecture 12);
- VC dimension, e.g., for binary functions, which we do not consider in this course;
- the Talagrand comparison inequality, i.e., $\mathbb{E}_\epsilon \sup_{f\in\mathcal{F}} |\langle \epsilon, f(X_1^n)\rangle| \lesssim \mathbb{E}_{g\sim N(0,I)} \sup_{f\in\mathcal{F}} |\langle g, f(X_1^n)\rangle|$, where the right hand side can be controlled using a rich range of techniques for Gaussian processes (e.g. Gaussian concentration, comparison and contraction).

In what follows, we give an example of bounding $R_n(\mathcal{F})$ using the union bound.

3 Glivenko-Cantelli Uniform Law of Large Numbers

Consider the RVs $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} \mu$ with CDF $F(\theta) := \mathbb{P}(X_1 \le \theta)$. Note that

$$F(\theta) = \mathbb{E}[\mathbb{1}\{X_1 \le \theta\}].$$

Introduce the shorthand $g_\theta(X) := \mathbb{1}\{X_1 \le \theta\}$.

We estimate F using the empirical CDF:

$$\hat{F}(\theta) := \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i \le \theta\} = \frac{1}{n}\sum_{i=1}^n g_\theta(X_i).$$

Consider the set of step functions $G \triangleq \{g_\theta : \theta \in \mathbb{R}\}$. We are interested in bounding the distance between $\hat{F}$ and F in sup norm:

$$\begin{aligned} \|\hat{F} - F\|_\infty := \sup_{\theta\in\mathbb{R}} \left|\hat{F}(\theta) - F(\theta)\right| &= \sup_{g\in\mathcal{G}} \left| \frac{1}{n}\sum_{i=1}^n \big(g_\theta(X_i) - \mathbb{E}[g(X_i)]\big) \right| \\ &\lesssim R_n(\mathcal{G}) = \mathbb{E}_X[R_n(\mathcal{G}|X)] \qquad \text{(from Theorem 1)} \\ &= \mathbb{E}_X\left[\mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i g(X_i)\right|\right] \\ &= \frac{1}{n}\mathbb{E}_X\mathbb{E}_\epsilon \sup_{\theta\in\mathbb{R}} \left|\sum_{i=1}^n \epsilon_i g_\theta(X_i)\right|. \end{aligned}$$

We shall upper bound the empirical Rademacher complexity $R_n(\mathcal{G}|X)$ for each fixed $X = (X_1, \ldots, X_n)$. Without loss of generality, we can assume that $X_1 \le X_2 \le \cdots \le X_n$. Because step function is non-increasing, the vector $(g_\theta(X_1), \ldots, g_\theta(X_n)) \in [0,1]^n$ must be ordered and can take on at most $n+1$ possible values, i.e.,

$$\begin{gathered} (0,0,\ldots,0) \\ (1,0,\ldots,0) \\ (1,1,\ldots,0) \\ \vdots \\ (1,1,\ldots,1) \end{gathered}$$

Therefore, the quantity $\sup_{\theta\in\mathbb{R}} \left|\sum_{i=1}^{n} \epsilon_i g_\theta (X_i)\right|$ is the supremum of at most $n+1$ random variables. Moreover, for each θ, the random variable $\epsilon_i g_\theta (X_i) \in [-1,1]$ is zero-mean and bounded. It follows that the sum $\sum_i \epsilon_i g_\theta (X_i)$ is zero-mean and $O(n)$-sub-Gaussian (by Hoeffding). Using the sub-Gaussian maximum expectation inequality (Lemma 3 from Lecture 10), we have

$$\mathbb{E}_\epsilon \sup_{\theta\in\mathbb{R}} \left|\sum_i \epsilon_i g_\theta (X_i)\right| \lesssim \sqrt{\log n}.$$

Combining this with the previous result, we have the bound

$$\mathbb{E}_\epsilon \sup_{\theta\in\mathbb{R}} \left|\hat{F}(\theta) - F(\theta)\right| \lesssim \sqrt{\frac{\log n}{n}}$$

Using the bounded difference inequality (Theorem 4 from Lectures 5-6) to prove concentration, we obtain the classical Glivenko Cantelli theorem.

Theorem 4 (Classical Glivenko-Cantelli). *With probability at least* $1-\exp\left(-n\delta^2\right)$*, we have*

$$\sup_{\theta\in\mathbb{R}} \left|\hat{F}(\theta) - F(\theta)\right| \lesssim \sqrt{\frac{\log n}{n}} + \delta,$$

which implies that $\sup_{\theta\in\mathbb{R}} \left|\hat{F}(\theta) - F(\theta)\right| \to 0$ *almost surely.*

Remark Note that we can remove the $\sqrt{\log n}$ factor using Dudley and VC-dimension.