Stat 928: Statistical Learning Theory **Lecture: 9**

Empirical Process Theory and Oracle Inequalities

Instructor: Sham Kakade

1 Risk vs Risk

See Lecture 0 for a discussion on terminology.

2 The Union Bound / Bonferoni

Consider m events $E_1, \dots E_m$, we have $P(E_1 \cup \cdots \cup E_m) \leq P(E_1) + \cdots + P(E_m)$. In other words, with probability $1 - P(E_1) - \cdots - P(E_m)$, none of the events E_i ($i = 1, \dots, m$) occurs.

If we assume the probability $\sum_j P(E_j)$ is small. Union bound is relatively tight when the events E_j are independent.

$$P(E_1 \cup \cdots \cup E_m) \geq \sum_j P(E_j) - \sum_{j \neq k} P(E_j \cap E_k) \geq \sum_j P(E_j) - 0.5(\sum_j P(E_j))^2.$$

If E_j are correlated, then it is not tight. For example when they are completely correlated: $E_1 = \cdots = E_m$, then

$$P(E_1 \cup \cdots \cup E_m) = N^{-1} \sum_j P(E_j).$$

We will come back to this when we discuss chaining.

3 Motivation of Empirical Process

Consider learning problem with observations $Z_i = (X_i, Y_i)$, prediction rule $f(X_i)$ and loss function $L(f(X_i), Y_i)$. Assume further that f is parameterized by $\theta \in \Theta$ as $f_\theta(X_i)$.

Example, $f_\theta(x) = \theta^\top x$ be a linear function, and $L(f_\theta(x), y) = (\theta^\top x - y)^2$ is least squares loss. In the following, we introduce simplified notation $g_\theta(Z_i) = L(f_\theta(X_i), Y_i)$. We are interested in estimating $\hat{\theta}$ from training data. That is, $\hat{\theta}$ depends on Z_i.

Since we are using the training data as a surrogate of the test (true underlying) distribution, we hope training error is similar to test error. In learning theory, we are interested in estimating the following tail quantities for some $\epsilon > 0$:

$$P(n^{-1} \sum_{i=1}^n g_{\hat{\theta}}(Z_i) \geq E g_{\hat{\theta}}(Z) + \epsilon)$$

and

$$P(n^{-1} \sum_{i=1}^n g_{\hat{\theta}}(Z_i) \leq E g_{\hat{\theta}}(Z) - \epsilon).$$

The above two quantities can be bounded using the following two quantities:

$$P(n^{-1}\sum_{i=1}^{n} g_{\hat{\theta}}(Z_i) \geq Eg_{\hat{\theta}}(Z) + \epsilon) \leq P[\sup_{\theta\in\Theta}(n^{-1}\sum_{i=1}^{n} g_{\theta}(Z_i) - Eg_{\theta}(Z)) \geq \epsilon]$$

and

$$P(n^{-1}\sum_{i=1}^{n} g_{\hat{\theta}}(Z_i) \leq Eg_{\hat{\theta}}(Z) - \epsilon) \leq P[\sup_{\theta\in\Theta}(Eg_{\theta}(Z) - n^{-1}\sum_{i=1}^{n} g_{\theta}(Z_i)) \geq \epsilon].$$

Notation: in the above setting the collection of random variables $n^{-1}\sum_{i=1}^{n} g_{\theta}(Z_i)$ indexed by $\theta \in \Gamma$ is call an empirical process. We may also call $n^{-1}\sum_{i=1}^{n} g_{\theta}(Z_i) - Eg_{\theta}(Z)$ empirical process.

For each fixed θ, $n^{-1}\sum_{i=1}^{n} g_{\theta}(Z_i) - Eg_{\theta}(Z) \to 0$ in probability, by LLN. However, in empirical process, we are interested in uniform law of large numbers, that is the following supremum of empirical process defined as

$$\sup_{\theta\in\Theta}|n^{-1}\sum_{i=1}^{n} g_{\theta}(Z_i) - Eg_{\theta}(Z)|$$

converges to zero in probability. Given training data $Z_1^n = \{Z_1, \ldots, Z_n\}$, we may let $\hat{\theta}(Z_1^n)$ achieve the supremum above. Then

$$\sup_{\theta\in\Theta}|n^{-1}\sum_{i=1}^{n} g_{\theta}(Z_i) - Eg_{\theta}(Z)| = |n^{-1}\sum_{i=1}^{n} g_{\hat{\theta}(Z_1^n)}(Z_i) - Eg_{\hat{\theta}(Z_1^n)}(Z)|,$$

where $\hat{\theta}(Z_1^n)$ depends on the training data. This means that $\sum_{i=1}^{n} g_{\hat{\theta}(Z_1^n)}(Z_i)$ is not sum of independent random variable anymore. Supreme of empirical process is basically the worst case deviation of empirical mean (training error) and true mean (test error) for parameter θ that is chosen based on training data.

Conceptually, as long as you select $\hat{\theta}$ based on training data, you need to use empirical process and uniform law of large numbers. However, if you only consider fixed θ independent of training data, then you can use standard law of large numbers because $g_{\theta}(Z_i)$ are independent random variable.