Note for Bounds

October 2024

This note is focused on constructing a map for the learning theory that I have encountered. I think it is beneficial to understand the relationship of different inequalities and how they can be bridged to construct a bound for the learned estimator.

1 Framework

We want to prove the performance of the selected estimator is close to the performance of the best estimator. There are 2 objects to consider:

- Empirical evaluation function vs population evaluation function
- Empirical estimator vs oracle estimator vs ground truth

$$R\left(h_*\right) \leq R(\widehat{h}) \leq \widehat{R}_n(\widehat{h}) + \epsilon \leq \widehat{R}_n\left(h_*\right) + \epsilon \leq R\left(h_*\right) + 2\epsilon$$

The above inequality gives us an outline of how to give a bound on $R(\widehat{h})$ and $R(h_*)$ when we get the estimator based on an empirical evaluation function $\widehat{R}_n$.

2 Concentration Inequality

These inequality can bound the empirical estimator with its mean, i.e. population counterpart.

2.1 Basic Inequalities

Markov Inequality:

$$\mathbb{P}(Z > \epsilon) \leq \frac{\mathbb{E}(Z)}{\epsilon}$$

Chebyshev Inequality:

$$\mathbb{P}(|Z-\mu|>\epsilon)=\mathbb{P}\left(|Z-\mu|^2>\epsilon^2\right)\leq\frac{\mathbb{E}(Z-\mu)^2}{\epsilon^2}=\frac{\sigma^2}{\epsilon^2}$$

The above inequality does not decay exponentially fast as n increases. So we use Chernoff's method to get a sharper bounds:

1. take exponential on both sides of the inequality: $\mathbb{P}(Z>\epsilon)=\mathbb{P}\left(e^Z>e^\epsilon\right)=\mathbb{P}\left(e^{tZ}>e^{t\epsilon}\right)\leq e^{-t\epsilon}\mathbb{E}\left(e^{tZ}\right)$

2. minimize over t: $\mathbb{P}(Z>\epsilon)\leq\inf_{t\geq0}e^{-t\epsilon}\mathbb{E}\left(e^{tZ}\right)$

Then we want to bound the moment generating function e^{tZ}. The following lemma give us a bound on this object:

Let Z be a mean μ random variable such that $a\leq Z\leq b$. Then, for any t

$$\mathbb{E}\left(e^{tZ}\right)\leq e^{t\mu+t^2(b-a)^2/8}$$

2.2 Hoeffding Inequality:

If $Z_1,Z_2,\ldots,Z_n$ are independent with $\mathbb{P}\left(a\leq Z_i\leq b\right)=1$ and common mean μ then for any $t>0$

$$\mathbb{P}\left(\left|\bar{Z}_n-\mu\right|>\epsilon\right)\leq2e^{-2n\epsilon^2/(b-a)^2}$$

where $\bar{Z}_n=\frac{1}{n}\sum_{i=1}^n Z_i$.

2.3 McDiarmid Inequality:

McDiarmid Inequality is an extension of Hoeffding's inequality, making it useful for various evaluation functions.

Let $Z_1,\ldots,Z_n$ be independent random variables.Suppose that for $i=1,\ldots,n$

$$\sup_{z_1,\ldots,z_n,z_i'}\left|f\left(z_1,\ldots,z_{i-1},z_i,z_{i+1},\ldots,z_n\right)-f\left(z_1,\ldots,z_{i-1},z_i',z_{i+1},\ldots,z_n\right)\right|\leq c_i$$

Then we have that

$$\mathbb{P}\left(\left|f\left(Z_1,\ldots,Z_n\right)-\mathbb{E}\left(f\left(Z_1,\ldots,Z_n\right)\right)\right|\geq\epsilon\right)\leq2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

2.4 Bernstein Inequality:

Hoeffding's inequality does not use any information about the random variables except the fact that they are bounded. If the variance of X_i is small, then we can get a sharper inequality from Bernstein's inequality.

If $\mathbb{P}(|X_i| \leq c) = 1$ and $\mathbb{E}(X_i) = \mu$ then, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\bar{X}_n - \mu\right| > \epsilon\right) \leq 2\exp\left\{-\frac{n\epsilon^2}{2\sigma^2 + 2c\epsilon/3}\right\}$$

where $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}(X_i)$.

3 Measure of complexity for function space

Here we introduce some measure of complexity for function space, like covering number, VC dimension and Rademacher complexity. In learning theory, they can be used to characterize the gap between empirical evaluation function and the population evaluation function. Typically we can use them to upper bound this gap (when we want a uniform bound), and if we further upper bound the complexity term (i.e. $\mathrm{Rad}_n(\mathcal{F})$), we can get a bound explained in n, which will be more useful.

3.1 VC dimension:

$$\mathcal{F}_{z_1,\ldots,z_n} = \{(f(z_1), \ldots, f(z_n)) : f \in \mathcal{F}\}$$

shattering number: If the (Indicator function) function space is big, then it can map the n sample into 2^n different vectors, so it "shattered" the sample. But if the function space is small, the shattering number will be less than 2^n.

$$s(\mathcal{F}, n) = \sup_{z_1,\ldots,z_n} \left|\mathcal{F}_{z_1,\ldots,z_n}\right|.$$

VC dimension: The VC dimension of a class of binary functions $\mathcal{F}$ is

$$\mathrm{VC}(\mathcal{F}) = \sup\{n : s(\mathcal{F}, n) = 2^n\}$$

.

Sauer's Theorem: Suppose that $\mathcal{F}$ has finite VC dimension d. Then,

$$s(\mathcal{F}, n) \leq \sum_{i=0}^{d} \binom{n}{i}$$

and for all $n \geq d$,

$$s(\mathcal{F}, n) \leq \left(\frac{en}{d}\right)^d$$

3.2 Rademacher Complexity

Random variables $\sigma_1, \ldots, \sigma_n$ are called Rademacher random variables if they are iid and $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$.

Rademacher complexity of $\mathcal{F}$:

$$\mathrm{Rad}_n(\mathcal{F}) = \mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right|\right)$$

empirical Rademacher complexity of $\mathcal{F}$ by

$$\mathrm{Rad}_n(\mathcal{F}, Z^n) = \mathbb{E}_\sigma\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(Z_i)\right|\right)$$

where $Z^n = (Z_1, \ldots, Z_n)$ and the expectation is over σ only.

Intuitively, $Rad_n(\mathcal{F})$ is large if we can find functions $f \in \mathcal{F}$ that "look like" random noise, that is, they are highly correlated with the random noise.

4 Uniform Bounds

When we are analysing function space, we want to extend this type of tail bound—valid for a single function f—to a result that applies uniformly to all functions in a certain function class F. The compexity measure can be used to get a uniform bound by 2 steps:

1. Show the empirical term is close to its mean, using concentration inequality.
2. Bound the mean, by introducing ghost sample/ symmetrization and then using Rademacher complexity.

Symmetrization: The importance of symmetrization is that we have replaced $(P_n - P)f$ which can take any real value, with $(P_n - P_n')f$, which can take only finitely many values.

4.1 Vapnik-Chervonenkis bound

Let $\mathcal{F}$ be a class of binary functions. For any $t > \sqrt{2/n}$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}|(P_n - P)f| > t\right) \leq 4s(\mathcal{F}, 2n)e^{-nt^2/8}$$

and hence, with probability at least $1 - \delta$,

$$\sup_{f\in\mathcal{F}}|P_n(f) - P(f)| \leq \sqrt{\frac{8}{n}\log\left(\frac{4s(\mathcal{F}, 2n)}{\delta}\right)}$$

4.2 Rademacher bound

With probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}|P_n(f)-P(f)| \leq 2\,\mathrm{Rad}_n(\mathcal{F})+\sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}$$

and

$$\sup_{f\in\mathcal{F}}|P_n(f)-P(f)| \leq 2\,\mathrm{Rad}_n\left(\mathcal{F},Z^n\right)+\sqrt{\frac{4}{n}\log\left(\frac{2}{\delta}\right)}.$$

5 Uniform bounds with localization:

In order to provide sharp rates of convergence, an important step is to localize the deviations to a small neighborhood of the origin.

Localized Rademacher complexity:

$$\overline{\mathcal{R}}_n(\delta;\mathcal{F})=\mathbb{E}_{\varepsilon,x}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f\left(x_i\right)\right|\right]$$

Bounds given by critical radius δ:

Given a star-shaped and b-uniformly bounded function class $\mathcal{F}$, let δ_n be any positive solution of the inequality

$$\overline{\mathcal{R}}_n(\delta;\mathcal{F})\leq\frac{\delta^2}{b}.$$

Then for any $t\geq\delta_n$, we have

$$\left|\|f\|_n^2-\|f\|_2^2\right|\leq\frac{1}{2}\|f\|_2^2+\frac{t^2}{2}\quad\text{ for all } f\in\mathcal{F}$$

with probability at least $1-c_1e^{-c\frac{n2^2}{b^2}}$.

For nonparametric function class, the empirical critical inequality $\widehat{\mathcal{R}}_n(\delta;\mathcal{F})\leq\frac{\delta^2}{b}$ can be replaced by:

$$\frac{64}{\sqrt{n}}\int_{\frac{\delta^2}{2b}}^{\delta}\sqrt{\log N_n\left(t;\mathbb{B}_n(\delta;\mathcal{F})\right)}dt\leq\frac{\delta^2}{b}.$$