COMS 6998-4 Fall 2017 Presenter: Daniel Hsu
September 27, 2017 Scribe: Xingyu Niu

Statistical Learning

1 Introduction

On of the widely investigated problems in statistical learning is the binary classification problem. In binary classification, a sequence of n i.i.d. sample pairs $(X_i, Y_i) \in \mathcal{X} \times \{\pm 1\}$ are drawn from an known distribution P. The goal is to find a function $h : X \to Y$ which predicts Y from X. We use the criterion

$$\mathrm{err}_P(h) = \Pr(h(X) \neq Y)$$

to measure the *risk* of function h.

We usually have a Hypothesis class $\mathcal{H}$ and choose the target function from $\mathcal{H}$. For any candidate h, we define its *regret* as

$$\mathrm{Reg}_{\mathcal{H},P}(h) = \mathrm{err}_P(h) - \min_{h' \in \mathcal{H}} \mathrm{err}_P(h').$$

The function with minimal regret will be selected as our target function: $h^* \in \arg\min_{h \in \mathcal{H}} \mathrm{err}_P(h)$. In practice, however, we cannot measure $\mathrm{err}_P(h)$ directly since the distribution P is unknown. We approximate it by measuring the agreement of h with the sample pairs, which is also called *empirical risk*:

$$\mathrm{err}_{P_n}(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{h(X_i) \neq Y_i},$$

and choose the function with minimal empirical risk $h_n \in \arg\min_{h \in \mathcal{H}} \mathrm{err}_{P_n}(h)$.

To analyze the regret of Empirical Risk Minimizatin (ERM), we can rewrite the regret into three parts:

$$\begin{aligned} \mathrm{err}_P(h_n) - \mathrm{err}_P(h^*) = & \mathrm{err}_P(h_n) - \mathrm{err}_{P_n}(h_n) \\ & + \mathrm{err}_{P_n}(h_n) - \mathrm{err}_{P_n}(h^*) \\ & + \mathrm{err}_{P_n}(h^*) - \mathrm{err}_P(h^*). \end{aligned}$$

The second part is less than or equals to 0. We can disregard it when we aim at deriving an upper bound of the regret. Since the target function h^* is independent of the sample pairs, the third part can be bounded easily by analyzing the binomial distribution with success probability $\mathrm{err}_P(h^*)$ and n trials. To analyze the remaining first part, we bound a new term which is greater than the first part:

$$\mathrm{err}_P(h_n) - \mathrm{err}_{P_n}(h_n) \leq \max_{h \in \mathcal{H}} \mathrm{err}_P(h) - \mathrm{err}_{P_n}(h).$$

2 Rademacher complexity

Previous results have already shown that with high probability, for a finite hypothesis class of size N, with probability greater than $1-\delta$, the empirical measure error bound for *empirical measure* achieves

$$\max_{h\in\mathcal{H}}(\mathrm{err}_P(h)-\mathrm{err}_{P_n}(h))\leq\sqrt{\frac{\log N+\log\frac{1}{\delta}}{2n}},$$

where n is the number of sample pairs. This bound is sublinear with respect to hypothesis class size and the inverse of δ, which is impressive. However, it is not applicable to cases where there are infinte candidate functions in hypothesis class.

To tackle this problem, we incorporate a new concept called *Rademacher Complexity* measuring the "richness" of infinite hypothesis class. First we define the Rademacher radom vector as $\epsilon=(\varepsilon_1,\ldots,\varepsilon_n)$, where $\varepsilon_1,\ldots,\varepsilon_n$ are random variables on i.i.d. distribution ($\Pr(\varepsilon_i=1)=\Pr(\varepsilon_i=-1)=\frac{1}{2}$). This vector can be used to define the Rademacher complexity of a set $S\subseteq\mathbb{R}^n$:

$$\mathrm{Rad}(S)=\mathbb{E}_\varepsilon\max_{v\in S}\langle\varepsilon,v\rangle_n=\mathbb{E}_\varepsilon\max_{v\in S}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i v_i$$

To extend the concept of *Rademacher Complexity* to an infinite hypothesis class, a tricky idea is to establish a mapping which maps hypothesis class into a set. For a hypothesis class $F\subseteq[-1,+1]^{\mathcal{Z}}$, and samples $z_1,\ldots,z_n\in\mathcal{Z}$, we define the projection of F on $Z_{1:n}$ as $F|_{Z_{1:n}}=\{(f(z_1),\ldots,f(z_n)),f\in F\}$. By this definition, we can further generalize the Rademacher complexity:

$$\mathrm{Rad}_{n,P}(F)=\mathbb{E}[\mathrm{Rad}(F|_{Z_{1:n}})],$$

where $\mathbb{E}$ is the expection with respect to $z_1,\ldots,z_n\sim P$.

Here are some examples of *Rademacher complexity*.

Proposition 1. *Assume for any $f\in\{\pm1\}^{\urcorner}$, $f\in F$, and all samples $z_1,\ldots,z_n$ are distinct,*

$$F|_{Z_{1:n}}=\{\pm1\}^n.$$

In addition,

$$\mathrm{Rad}_{n,P}(F)=1.$$

Proposition 1 is straightforward. If F contains all functions in $\{\pm1\}^{\mathcal{Z}}$, then for any element $(a_1,\ldots,a_n)$ in $\{\pm1\}^n$, we can always find a function $f\in F$ such that $f(\mathcal{Z}_i)=a_i$ for all i. Thus $\{\pm1\}^n$ is its projection onto $\mathcal{Z}_{1:n}$ and

$$\mathrm{Rad}_{n,P}(F)=\mathbb{E}_\varepsilon[\max_{v\in\{\pm1\}^n}\langle\varepsilon,v\rangle_n]=\mathbb{E}_\varepsilon[\frac{1}{n}\sum_{i=1}^{n}|\varepsilon_i|]=1.$$

Proposition 2. *Assume $F=\{f_0\}$, $\mathrm{Rad}_{n,P}(F)=0$.*

Since there is only one function in the class, $F|_{\mathcal{Z}_{1:n}} = \{f_0(\mathcal{Z}_1), \ldots, f_0(\mathcal{Z}_n)\}$.

$$\mathrm{Rad}_{n,P}(F) = \mathbb{E}_\varepsilon[\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i f_0(\mathcal{Z}_i)] = \frac{1}{n}\sum_{i=1}^{n} f_0(\mathcal{Z}_i)\mathbb{E}_\varepsilon[\varepsilon_i] = 0.$$

Proposition 3. *Assume F is a finite hypothesis class with cardinality $|F|$ and $z_1, \ldots, z_n$ are the samples from distribution P, we have*

$$\mathrm{Rad}_{n,P}(F) = \sqrt{\frac{2\ln|F|}{n}}.$$

Proof. We want to set up the use of Hoeffding's Inequality. To begin with, we take the exponential of the empirical Rademacher complexity multiplied by a constant $s > 0$. By Jensen's Inequality, with the convex function $\exp(x)$, we have

$$\begin{aligned}
\exp\Big(s\mathbb{E}_\varepsilon\big[\max_{f\in F}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(z_i)\big]\Big) &\le \mathbb{E}_\varepsilon\Big[\exp\big(s\max_{f\in F}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(z_i)\big)\Big] \\
&= \mathbb{E}_\varepsilon\Big[\max_{f\in F}\Big(\exp\big(\frac{s}{n}\sum_{i=1}^{n}\varepsilon_i f(z_i)\big)\Big)\Big] \\
&\le \sum_{f\in F}\mathbb{E}_\varepsilon\Big[\exp\big(\frac{s}{n}\sum_{i=1}^{n}\varepsilon_i f(z_i)\big)\Big] \\
&= \sum_{f\in F}\mathbb{E}_\varepsilon\Big[\prod_{i=1}^{n}\exp\big(\frac{s}{n}\varepsilon_i f(z_i)\big)\Big] \\
&= \sum_{f\in F}\prod_{i=1}^{n}\mathbb{E}_\varepsilon\Big[\exp\big(\frac{s}{n}\varepsilon_i f(z_i)\big)\Big],
\end{aligned}$$

where the first equality follows from that $\exp(x)$ is an increasing function; the second inequality holds since F is a finite set; the last equality uses the fact that all ε_i are independent. Now we can apply Hoeffding's Inequality since $\mathbb{E}_\varepsilon[\varepsilon_i f(z_i)] = 0$ and $\varepsilon_i f(z_i) \in \{-1, 1\}$. This gives us

$$\begin{aligned}
\exp\Big(s\mathbb{E}_\varepsilon\big[\max_{f\in F}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(z_i)\big]\Big) &\le \sum_{f\in F}\prod_{i=1}^{n}\exp\Big(\frac{s^2}{2n^2}\Big) \\
&= \sum_{f\in F}\exp\Big(\frac{s^2}{2n}\Big) \\
&\le |F|\exp\Big(\frac{s^2}{2n}\Big).
\end{aligned}$$

Removing the exponential, we have

$$\mathrm{Rad}_{n,P}(F) = \mathbb{E}_\varepsilon\big[\max_{f\in F}\sum_{i=1}^{n}\varepsilon_i f(z_i)\big] \le \frac{\ln|F|}{s} + \frac{s}{2n},$$

where $\mathrm{Rad}_{n,P}(F)$ is minimized when $s = \sqrt{2n \ln |F|}$. Substituting this quantity back to the inequality we get

$$\mathrm{Rad}_{n,P}(F) \leq \sqrt{\frac{2 \ln |F|}{n}}.$$

□

3 Regret bound of empirical measure

Consider a binary value function class $F_{\mathcal{H}} \subseteq \{\pm 1\}^{\mathcal{Z}}$, where $\mathcal{Z} = \mathcal{X} \times \{0, 1\}$. Let

$$f_h(x, y) = \begin{cases} +1 & \text{if } h(x) \neq y \\ -1 & \text{otherwise.} \end{cases}$$

Proposition 4. *For $h \in \mathcal{H}$ and its corresponding $f_h(x, y)$,*

$$\mathrm{err}_P(h) - \mathrm{err}_{P_n}(h) = \frac{1}{2}(Pf_h - P_n f_h),$$

where

$$Pf = \mathbb{E}_{z \sim P} f(z)$$

and

$$P_n f = \mathbb{E}_{z \sim P_n} f(z) = \frac{1}{n} \sum_{i=1}^{n} f(z_i).$$

Proof. By direct computation:

$$\begin{aligned} Pf_n - P_n f_n &= E_{z \sim P} f(z) - \frac{1}{n} \sum_{i=1}^{n} f(z_i) \\ &= \Pr(h(x) \neq y) - (1 - \Pr(h(x) \neq y)) \\ &\quad - \frac{1}{n}[\sum_{i:h(x_i) \neq y_i} 1 - (n - \sum_{i:h(x_i) \neq y_i} 1)] \\ &= 2[\Pr(h(x) \neq y) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{h(x_i) \neq y_i}] \\ &= 2(\mathrm{err}_P(h) - \mathrm{err}_{P_n}(h)). \end{aligned}$$

□

Proposition 4 shows the relation between empirical regret of hypothesis h and empirical regret of constructed function f_h. Thus, the problem is transformed into bounding $Pf_h - P_n f_h$. As $h \in \mathcal{H}$, in the new problem we aims to bound the term $\max_{h \in \mathcal{H}} Pf - P_n f$, which is the greatest empirical regret among all hypotheses in $\mathcal{H}$. The main chanllenge of this problem is that the empirical regret depends on the sampled data. With different sets of samples, the

greatest empirical regret will be different and so is the corresponding hypothesis. A tricky idea is to bound the expectation of this term first using symmetrization, and then relates this bound back to the original term using concentration. Theorem 1 shows how the expectation of $\max_{h\in\mathcal{H}} Pf - P_n f$ can be bounded by Rademacher complexity.

Theorem 1. *For any function class* $F \subseteq [-1,+1]^{\mathcal{Z}}$,

$$\mathbb{E}\max_{f\in F} Pf - P_n f \leq 2\,\mathrm{Rad}_{n,P}(F).$$

Proof. We employ the symmetrization technique to proof Theorem 1. First we introduce a set of ghost samples $z_1^{'},\ldots,z_n{}'$ i.i.d. from the sample distribution P as $z_1,\ldots,z_n$. These ghost examples can be viewed as independent copies of $z_1,\ldots,z_n$, and they form a new empirical distribution $P_n^{'}$. Thus we have

$$P_n^{'} f = \mathbb{E}_{Z'\sim P_n'}[\frac{1}{n}\sum_{i=1}^{n} f(z_i^{'})].$$

A trival fact follows that $Pf = \mathbb{E}_{Z'} P_n^{'} f$, where the expectation is taken w.r.t. $z_1^{'},\ldots,z_n^{'}$. Using this fact, we have

$$\begin{aligned}
&\mathbb{E}_{Z_{1:n}} \max_{f\in F} Pf - P_n f \\
=\ &\mathbb{E}_{Z_{1:n}} \max_{f\in F} (\mathbb{E}_{Z'_{1:n}} P_n^{'} f) - P_n f \\
\leq\ &\mathbb{E}_{Z_{1:n}} \mathbb{E}_{Z'_{1:n}} \max_{f\in F} P_n^{'} f - P_n f,
\end{aligned}$$

where the inequality follows from the convexity of the maximum function. Further expanding the expectation of difference we have

$$\begin{aligned}
&\mathbb{E}\max_{f\in F} P_n^{'} f - P_n f \\
=\ &\mathbb{E}\max_{f\in F} \frac{1}{n}\sum_{i=1}^{n} f(z_i^{'}) - f(z_i) \\
=\ &\mathbb{E}\max_{f\in F} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i (f(z_i^{'}) - f(z_i)) \\
=\ &\mathbb{E}\max_{f\in F} \langle \varepsilon, f(Z_{1:n}^{'})\rangle_n - \langle \varepsilon, f(Z_{1:n})\rangle_n \\
\leq\ &2\mathbb{E}\max_{f\in F} \langle \varepsilon, f(Z_{1:n})\rangle = 2\,\mathrm{Rad}(F|_{Z_{1:n}}),
\end{aligned}$$

where $\mathbb{E}$ is expectation with respect to $Z_{1:n}$, $Z_{1:n}^{'}$ and ε. The second equality holds because both $\frac{1}{n}\sum_{i=1}^{n} f(z_i^{'}) - f(z_i)$ and $\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i (f(z_i^{'}) - f(z_i))$ follows the same distribution: for $i = 1,\ldots,n$ and $r \in \{0, \pm 2\}$,

$$\begin{aligned}
&\Pr[\varepsilon_i (f(z_i^{'}) - f(z_i)) = r] \\
=\ &\Pr[\varepsilon_i = 1, f(z_i^{'}) - f(z_i) = r] + \Pr[\varepsilon_i = -1, f(z_i^{'}) - f(z_i) = -r] \\
=\ &2\cdot\frac{1}{2}\Pr[f(z_i^{'}) - f(z_i) = r] = \Pr[f(z_i^{'}) - f(z_i) = r].
\end{aligned}$$

The inequality uses the fact that $\langle \varepsilon, f(Z'_{1:n})\rangle_n$ and $\langle -\varepsilon, f(Z_{1:n})\rangle_n$ have the same distribution. □

Theorem 1 provides an upper bound of the expectation of greatest empirical regret in $\mathcal{H}$. What remains to do is to take out the expectation. A generalized version of Hoeffding's inequality, the McDiarmid's inequality, can be employed to achieve this task.

Theorem 2. *Assume for all* $i = 1, \ldots, n$,

$$sup_{z_1,\ldots,z_n,z_i'} |F(z_1, \ldots, z_i, \ldots, z_n) - F(z_1, \ldots, z_i', \ldots, z_n)| \leq c,$$

then with probability at least $1 - \delta$,

$$\Pr[|F - \mathbb{E}[F]|] \leq \sqrt{\frac{nc^2 \log \frac{1}{\delta}}{2}} \ .$$

We do not prove Theorem 2 in the presentation. Here we show how the empirical regret satisfies the prerequsite of this theorem. Suppose we change the empirical distribution P_n into P_n^i by replacing the i-th sample z_i by z_i', we have

$$\begin{aligned} & |\max_{f\in F}(Pf - P_n f) - \max_{f\in F}(Pf - P_n^i f)| \\ \leq\ & \max_{f\in F} |P_n^i f - P_n f| \\ =& \max_{f\in F} \frac{1}{n} |f(z_i') - f(z_i)| \leq \frac{2}{n}, \end{aligned}$$

where the last inequality holds because function f return ± 1. Therefore, changing any one sample in the empirical distribution leads to minor change to the empirical regret, and the difference of regrets is less than or equal to $c = \frac{1}{n}$.

Combining all the results above, we know with probability at least $1 - \delta$,

$$\begin{aligned} & \max_{h\in\mathcal{H}} err_P(h) - err_{P_n}(h) = \frac{1}{2}(\max_{f\in F_{\mathcal{H}}} Pf - P_n f) \\ \leq\ & \frac{1}{2}\mathbb{E}[\max_{f\in F_{\mathcal{H}}} Pf - P_n f] + \sqrt{\frac{\log \frac{1}{\delta}}{2n}} \leq \mathrm{Rad}_{n,P}(F) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}. \end{aligned}$$

This result gives an impressive upper bound on the empirical regret, which linear with the Rademacher complexity and sublinear with the probability term $\log \frac{1}{\delta}$. It is worthwhile to note that this bound takes a similar form as in finite cases, where the empirical measure error bound is $O(\sqrt{\frac{\log N + \log \frac{1}{\delta}}{2n}})$. The main idea of incorporating *Rademacher complexity* is the useage of mapping from infinte hypothesis class to finite sample-value pairs set.

Bibliographic notes

Hoeffding's inequality appears in Hoeffding. [1]. The bounded differences inequality was formulated explicitly by McDiarmid. [2], which gives a proof of the McDiarmid inequality. Bousquet, et al. [3] formulate the empirical measure error bound in statistical learning theory.

References

[1] W. Hoeffding. Probability inequalities for sums of bounded random variables *Journal of the American statistical association* 58(301):13–30, 1963.

[2] C. McDiarmid. On the method of bounded differences In: *Surveys in combinatorics*141(1):148–188, 1989.

[3] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to Statistical Learning Theory *Advanced lectures on machine learning*, 169–207, 2004.