Machine Learning Theory Notes

Gaotang Li

February 8, 2023

Contents

Part I

Generalization Theory

Chapter 1

Supervised Learning Framework

1.1 Basic Setups

In a supervised learning problem, we have a goal to predict a label given an input. Let S denote the dataset $\{(x_i, y_i)\}_{i=1}^n$ for

- $x_i \in \mathcal{X}$, the inputs in the input space.
- $y_i \in \mathcal{Y}$, the label associated with x_i in the label space.

We assume that the data are drawn **i.i.d.** from an unknown probability distribution $\mathcal{P}$ over $\mathcal{X} \times \mathcal{Y}$. We aim at learning a function mapping $h\colon \mathcal{X} \to \mathcal{Y}$ (aka *hypothesis*, *predictor*, *model*).

To evaluate the performance of h, we specify a loss function. A loss function $\ell\colon \mathcal{Y}, \mathcal{Y} \to \mathbb{R}$ measures the difference between the predicted label and the groundtruth label.

Definition 1.1.1 (population risk)**.** The *population risk* of a *hypothesis* h is its expected loss over the data distribution $\mathcal{P}$: $L_{\mathcal{P}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(h(x), y)]$

Example. Examples of Loss Functions

- Classification: 0-1 loss $\ell(\hat{y}, y) = \mathbb{1}(\hat{y} \neq y)$.
- Regression: squared loss $\ell(\hat{y}, y) = (\hat{y} - y)^2$

It is often impossible to consider all possible function mappings from $\mathcal{X} \to \mathcal{Y}$. We usually only consider a *hypothesis class* $\mathcal{H}$.

Example. Examples of $\mathcal{H}$.

- Linear function class: $\mathcal{H} = \{h_\theta | h_\theta(x) = \theta^T x, \theta \in \mathbb{R}\}$
- General parametric function class: $\mathcal{H} = \{h_\theta | h_\theta(x) = f(x, \theta), \theta \in \mathbb{R}^p\}$

1.2 Empirical Risk Minimization

Definition 1.2.1 (Empirical Risk)**.** The *empirical risk* of a hypothesis h is its average loss over the dataset S

$$L_s(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i)$$

Empirical risk minimization (ERM) is any algorithm that minimizes the empirical risk over the hypothesis class $\mathcal{H}$. We denote a hypothesis returned by ERM as $\hat{h}_{ERM}$, *i.e.*:

$$\hat{h}_{ERM} \in \text{argmin}_{h\in\mathcal{H}} L_S(h)$$

If we assume h is independent of S, then $\mathbb{E}_S[L_S(h)] = L_P(h)$. But in reality h and S are not independent.

1.3 Questions

In the supervised learning part of this course, we are mainly interested in the following two fundamental problems:

- **Statistical**: What guarantee do we have about $L_P(\hat{h}_{ERM})$?
- **Optimization**: When may ERM be achieved efficiently?

Chapter 2

Concentration Inequality

Concentration inequalities are a mathematical tool to study the relation between population and empirical quantities. Consider the following main question: for i.i.d. random variables $X_1, \ldots, X_n$, how does $\frac{1}{n}\sum_{i=1}^{n} X_i$ relate to $\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \mu$?

2.1 Chebyshev's Inequality

Lemma 2.1.1 (Markov's Inequality). Let X be a non-negative random variable, then for all $t > 0$,

$$\Pr(X \geq t) \leq \frac{\mathbb{E}(x)}{t}$$

Proof.

$$\mathbb{E}(X) \geq \Pr(X < t) * 0 + \Pr(X > t) * t$$

■

Theorem 2.1.1 (Chebyshev's Inequality). Let X be a random variable with finite expected value μ and finite non-zero variance σ^2. Then for any real number $t > 0$,

$$\Pr\left[|X - \mathbb{E}(X)| \geq t\right] \leq \frac{\mathrm{Var}(X)}{t}$$

Proof.

$$\begin{aligned}
\Pr\left[X - \mathbb{E}[X] \geq t\right] &= \Pr\left[(X - \mathbb{E}[X])^2 \geq t^2\right] \\
&= \frac{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}{t^2} \\
&= \frac{\mathrm{Var}[X]}{t^2}
\end{aligned}$$

■

Remark.

$$\Pr\left[|X - \mathbb{E}[X]| \geq t\right] \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^P]}{t^P}$$

Corollary 2.1.1. Let $x_1, \ldots, x_n$ be i.i.d. random variables such that $\mathbb{E}\left[x_i\right] = \mu$, $\mathrm{Var}[x_i] = \sigma^2$. Then:

$$\Pr\left[|\frac{1}{n}\sum_{i=1}^{n} x_i - \mu| \geq t\right] \leq \frac{\sigma^2}{nt^2}$$

2.2 Hoeffding Inequality

Lemma 2.2.1. If $X \in [0,1]$ a.s. Then,

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq e^{\frac{\lambda^2}{8}}$$

for all $\lambda \in \mathbb{R}$.

Proof. Let $Z = X - \mathbb{E}[X]$, then $\mathbb{E}[Z] = 0$.
Define $\psi(\lambda) := \log \mathbb{E}\left[e^{\lambda Z}\right]$.
Using the Taylor expansion to get that $\psi(\lambda) = \psi(0) + \lambda\psi'(0) + \frac{\lambda^2}{2}\psi''(\lambda')$ where λ' is between 0 and λ.

Here the first term $\psi(0) = \log 1 = 0$, and the second term $\lambda\psi'(0) = \mathbb{E}[Z] = 0$. The only thing we need to is to compute the third term. The idea is to bound the third term by $1/4$.

Then

$$\begin{aligned}
\psi'(\lambda) &= \frac{\mathbb{E}\left[e^{\lambda Z} Z\right]}{e^{\lambda Z}} = \mathbb{E}[Y] \\
\psi''(\lambda) &= \frac{\mathbb{E}\left[e^{\lambda Z} Z^2\right]}{e^{\lambda Z}} - \left(\frac{\mathbb{E}\left[e^{\lambda Z} Z\right]}{\mathbb{E}[e^{\lambda Z}]}\right)^2 \\
&= \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 \\
&= \mathrm{Var}[Y]
\end{aligned}$$

Where we can think of Y as a reweighted version of Z, and we have that

$$dP_Y(x) = \frac{e^{\lambda x}}{\mathbb{E}\left[e^{\lambda Z}\right]} dP_Z(x)$$

not finished yet... ■

Remark. We also call such random variables **subgaussian** random variables. Another interpretation is that bounded random variables are subgaussian.

Another reminder is that the expectation is in the form of **Moment Generating Function**, where $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$.

Theorem 2.2.1 (Hoeffding Inequality)**.** Let $X_1, \ldots, X_n$ be i.i.d. random variables such that for each i, $X_i \in [0,1]$ a.s. Then for all $t > 0$:

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \geq t\right] \leq e^{-2nt^2}$$
$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \leq -t\right] \leq e^{-2nt^2}$$

Proof. Let us use $\bar{X}$ to denote $\frac{1}{n}\sum_{i=1}^n X_i$. Then we have that

$$\begin{aligned}
\Pr\left[\bar{X}-\mathbb{E}[\bar{X}]\geq t\right] &= \Pr\left[e^{\lambda(\bar{X}-\mathbb{E}[\bar{X}])}\geq e^{\lambda t}\right] && \lambda>0\\
&\leq \frac{\mathbb{E}\left[e^{\lambda(\bar{X}-\mathbb{E}[\bar{X}])}\right]}{e^{\lambda t}} && \text{Markov}\\
&= e^{-\lambda t}\cdot\mathbb{E}\left[e^{\frac{\lambda}{n}\sum_{i=1}^n(X_i-\mathbb{E}[X_i])}\right]\\
&= e^{-\lambda t}\cdot\mathbb{E}\left[\prod_{i=1}^n e^{\frac{\lambda}{n}(X_i-\mathbb{E}[X_i])}\right]\\
&= e^{-\lambda t}\prod_{i=1}^n\mathbb{E}\left[e^{\frac{\lambda}{n}(X_i-\mathbb{E}[X_i])}\right] && \text{Independence}\\
&\leq e^{-\lambda t}\left(e^{\frac{1}{8}(\frac{\lambda}{n})^n}\right)\\
&= e^{-\lambda t+\frac{\lambda^2}{8n}}\\
&= e^{-2nt^2} && \text{let } \lambda=4nt
\end{aligned}$$

By symmetry, we complete the proof. ■

Remark (Equivalent Definition of Hoeffding Inequality)**.** Let $X_1,\dots,X_n\in[0,1]$ a.s. and independent,

$$\forall\delta\in(0,1),\ \text{w.p.}\geq 1-\delta:\quad \frac{1}{n}\sum_{i=1}^n X_i-\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right]\leq\sqrt{\frac{\log\frac{1}{\delta}}{2n}}$$

$$\text{w.p.}\geq 1-\delta:\quad \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right]-\frac{1}{n}\sum_{i=1}^n X_i\leq\sqrt{\frac{\log\frac{1}{\delta}}{2n}}$$

2.3 Bounded Difference Concentration Inequality

We are concerned with diffeormorphism (*i.e.* change in one coordinate) and formally

$$f(X_1,\cdots,X_n)\mapsto\mathbb{E}\left[f(X_1,\cdots,X_n)\right]$$

Theorem 2.3.1 (Mcdiarmid's inequality)**.** Suppose $X_1,\dots,X_n$ are independent random variables taking values in a set A. Let $f\colon A^n\to\mathbb{R}$ be a function that satisfies the *bounded difference* condition:

$$\exists c_1,\dots,c_n>0 \text{s.t.} \forall\ x_1,\dots,x_n\in A, x_i'\in A|f(x_1,\dots,x_n)-f(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_n)|\leq c_i$$

Then, for all $t>0$,

$$\Pr\left[f(X_1,\dots,X_n)-\mathbb{E}\left[f(X_1,\dots,X_n)\right]\geq t\right]\leq e^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}}$$

Remark. If $f(X_1,\dots,X_n)=\frac{1}{n}\sum_{i=1}^n x_i$ and $A=[0,1]$, then $c_i=\frac{1}{n}$ and the bound recovers the Hoeffding inequality as e^{-2nt^2}.

Chapter 3

Rademacher Complexity

3.1 Uniform Convergence

Motivation: we want to study $L(\hat{h}_{\text{ERM}})$ and compare it against $h^* \in \text{argmin}_{h\in\mathcal{H}} L(h)$. We want to bound the difference $L(\hat{h}_{\text{ERM}}) - L(h^*)$, which is also referred to as the "**excess risk**".

$$L(\hat{h}_{\text{ERM}}) - L(h^*) = \left(L(\hat{h}_{\text{ERM}}) - L_S(\hat{h}_{\text{ERM}})\right) + \left(L_S(\hat{h}_{\text{ERM}}) - L_S(h^*)\right) + \left(L_S(h^*) - L(h^*)\right)$$

where the second term is smaller or equal to 0 by definition, and the third term can be bounded using the Hoeffding inequality as h^* does not depend on S.

Consequently, our aim becomes bounding the first term and we define the following **generalization gap**:

Definition 3.1.1 (Uniform Convergence).

$$L(\hat{h}_{\text{ERM}}) - L_S(\hat{h}_{\text{ERM}}) \leq \sup_{h\in\mathcal{H}} \left(L(h) - L_S(h)\right)$$

, where the bounded difference is called the generalization gap.

Theorem 3.1.1 (Generalization Bound for finite hypothesis class). If $\mathcal{H}$ is finite, then for any $\delta \in (0, 1)$, we have

$$\text{w.p.} \geq 1 - \delta, \sup_{h\in\mathcal{H}} \left(L(h) - L_S(h)\right) \leq \sqrt{\frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{2n}}$$

Remark. If $n >> \log|\mathcal{H}|$, excess risk $\to 0$.

What if $\mathcal{H}$ is infinite?

– Idea: Reduce infinite case to finite case.

3.2 Rademacher Complexity

Notation: Given $\mathcal{H}$ and ℓ, define the family of loss mappings:

$$\mathcal{G} = \{g_h \colon (x, y) \mapsto \ell(h(x), y), h \in \mathcal{H}\}$$

where $z = (x, y) \sim P$, $z_i = (x_i, y_i)$, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, and $L(h) = \mathbb{E}_{z\sim P}\left[g_h(z)\right]$, $L_S(h) = \frac{1}{n}\sum_{i=1}^{n} g_h(z_i)$.

$$\sup_{h\in\mathcal{H}} (L(h) - L_S(h)) = \sup_{g\in\mathcal{G}} \left(\mathbb{E}_{z\sim P}[g(z)] - \frac{1}{n}\sum_{i=1}^{n} g(z_i)\right)$$

Definition 3.2.1 (Empirical Rademacher Complexity)**.** Let $\mathcal{G}$ be a set of functions mapping $\mathcal{Z} \to \mathbb{R}$. Let $S = \{z_1, \cdots, z_n\} \subseteq \mathcal{Z}$.

The **empirical Rademacher complexity** of $\mathcal{G}$ with respect to the simple set S is:

$$R_S(\mathcal{G}) = \mathbb{E}_{\sigma_1,\cdots,\sigma_n}\left[\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i g(z_i)\right]$$

where $\sigma_i = \begin{cases} +1 & \text{w.p.} \frac{1}{2} \\ -1 & \text{w.p.} \frac{1}{2} \end{cases}$ i.i.d (called Rademacher random variables).

Remark. Rademacher complexity measures the ability of a function class to fit random noise

$$R_S(\mathcal{G}) = \mathbb{E}_{\vec{\sigma}}\left[\sup_{g\in\mathcal{G}} \frac{1}{n} < \vec{\sigma}, \vec{g}_s >\right]$$

Definition 3.2.2 (Rademacher Complexity)**.** Let P be a distribution over $\mathcal{Z}$.

For an integer $n \geq 1$, the **Rademacher complexity** of $\mathcal{G}$ is

$$R_n(\mathcal{G}) = \mathbb{E}_{S\sim P^n}\left[R_S(\mathcal{G})\right]$$

Theorem 3.2.1 (Generalization Bound using Rademacher Complexity)**.** Let $\mathcal{G}$ be a function class mapping $\mathcal{Z}$ to $[0,1]$, $S = \{z_1, \ldots, z_n\} \sim P^n$. Then for any $\delta \in (0,1)$:

$$\text{w.p.} \geq 1-\delta,\ \sup_{g\in\mathcal{G}} \left(\mathbb{E}_{z\sim P}\left[g(z)\right] - \frac{1}{n}\sum_{i=1}^{n} g(z_i)\right) \leq 2R_n(\mathcal{G}) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}$$

$$\text{w.p.} \geq 1-\delta,\ \sup_{g\in\mathcal{G}} \left(\mathbb{E}_{z\sim P}\left[g(z)\right] - \frac{1}{n}\sum_{i=1}^{n} g(z_i)\right) \leq 2R_S(\mathcal{G}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

Proof. **Step 1: Relate the sup terms to the expectation of sups using Mcdiarmid's ineq**

Define $f(z_1, \cdots, z_n) = \sup_{g\in\mathcal{G}}\left(\mathbb{E}_{z\sim P}\left[g(z)\right] - \frac{1}{n}\sum_{i=1}^{n} g(z_i)\right)$.

Consider $\{z_1, \ldots, z_n\}$ and $\{z'_1, \ldots, z'_n\}$ that only differs by 1 point (*i.e.* $z_k \neq z'_k, z_i = z'_i\ \forall i \neq k$).

$$\begin{aligned}
f(z_1, \ldots, z_n) &= \sup_{g\in\mathcal{G}} \left(\mathbb{E}\left[g(z)\right] - \frac{1}{n}\sum_{i=1}^{n} g(z'_i) + \frac{1}{n}\sum_{i=1}^{n} g(z'_i) - \frac{1}{n}\sum_{i=1}^{n} g(z_i)\right) \\
&\leq \sup_{g\in\mathcal{G}} \left(\mathbb{E}\left[g(z)\right] - \frac{1}{n}\sum_{i=1}^{n} g(z'_i)\right) + \sup_{g\in\mathcal{G}} \left(\frac{1}{n}\sum_{i=1}^{n} g(z'_i) - \frac{1}{n}\sum_{i=1}^{n} g(z_i)\right) \\
&= f(z'_1, \ldots, z'_n) + \sup_{g\in\mathcal{G}} \left(\frac{1}{n} g(z'_k) - \frac{1}{n} g(z_k)\right) \\
&\leq f(z'_1, \ldots, z'_n) + \frac{1}{n}
\end{aligned}$$

Similarly, $f(z'_i, \ldots, z'_n) - f(z_1, \ldots, z_n) \leq \frac{1}{n}$. Combining them we can get that $|f(z_1, \ldots, z_n) - f(z'_1, \ldots, z'_n)| \leq \frac{1}{n}$.

Applying the Mcdiarmid's inequality, we can get the following bound:

$$\text{w.p.} \geq 1 - \delta, f(z_1, \ldots, z_n) - \mathbb{E}\left[f(z_1, \ldots, z_n)\right] \leq \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

Step 2: Bound $\mathbb{E}_S\left[\sup_{g \in \mathcal{G}}\left(\mathbb{E}_{z \sim P}\left[g(z)\right] - \frac{1}{n}\sum_{i=1}^n g(z_i)\right)\right]$ by Rademacher Complexity

Draw a fresh set of n samples $S' = \{z'_1, \ldots, z'_n\} \sim P^n$. Fix S, we have

$$\begin{aligned}
\sup_{g \in \mathcal{G}}\left(\mathbb{E}_{z \sim P}\left[g(z)\right] - \frac{1}{n}\sum_{i=1}^n g(z_i)\right) &= \sup_{g \in \mathcal{G}}\left(\mathbb{E}_{S'}\left[\frac{1}{n}\sum_{i=1}^n g(z_i)\right] - \frac{1}{n}\sum_{i=1}^n g(z_i)\right) \\
&= \sup_{g \in \mathcal{G}}\left(\mathbb{E}_{S'}\left[\frac{1}{n}\sum_{i=1}^n \left(g(z'_i) - g(z_i)\right)\right]\right) \\
&\leq \mathbb{E}_{S'}\left[\sup_{g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^n \left(g(z'_i) - g(z_i)\right)\right]
\end{aligned}$$

Taking expectation over S on both sides generate that

$$\begin{aligned}
\mathbb{E}_S\left[\sup_{g \in G}\left(\mathbb{E}_{z \sim P}\left[g(z)\right] - \frac{1}{n}\sum_{i=1}^n g(z_i)\right)\right] &\leq \mathbb{E}_{S,S'}\left[\sup_{g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^n \left(g(z'_i) - g(z_i)\right)\right] \\
&= \mathbb{E}_{S,S',\vec{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i\left(g(z'_i) - g(z_i)\right)\right] \qquad ? \\
&\leq \mathbb{E}_{S,S',\vec{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^n \sigma_i g(z'_i)\right] + \mathbb{E}_{S,S'\vec{\sigma}}\left[\sup_{g \in \mathcal{G}} \frac{1}{n}\sum_{i=1}^n -\sigma_i g(z_i)\right] \\
&= 2R_n(\mathcal{G})
\end{aligned}$$

Combining the result from **step 1** and **step 2**, we prove the first inequality in the theorem.

Step 3: Prove $R_n(\mathcal{G})$ and $R_S(\mathcal{G})$ are close Similar to step 1, we can verify that $R_S(\mathcal{G})$ satisifies the bounded difference property.

Apply Mcdiarmid's inequality, we can get that

$$\text{w.p.} \geq 1 - \delta, R_n(\mathcal{G}) \leq R_S(\mathcal{G}) + \sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

Combining the outputs from step 1 - 3 and replacing δ with $\delta/2$ gives the second inequality. ■

Chapter 4

VC-Dimension

In this chapter, we only consider the binary classification case with the 0-1 loss, *i.e.* $y = \{\pm 1\}$ and $\mathcal{G} = \{(x, y) \mapsto \mathbb{1}[h(x) \neq y] : h \in \mathcal{H}\}$.

4.1 Using VC-Dimension to bound generalization bound

Lemma 4.1.1. $R_n(\mathcal{G}) = \frac{1}{2} R_n(\mathcal{H})$

Proof. Given $S = \{(x_i, y_i)\}_{i=1}^n$, we have

$$\begin{aligned}
R_S(\mathcal{G}) &= \mathbb{E}_{\vec{\sigma}}\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sigma_i \mathbb{1}[h(x_i) \neq y_i]\right] \\
&= \mathbb{E}_{\vec{\sigma}}\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \frac{1 - y_i h(x_i)}{2}\right] \\
&= \frac{1}{2}\mathbb{E}_{\vec{\sigma}}\left[\frac{1}{n} \sum_{i=1}^n \sigma_i + \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i(-y_i) h(x_i)\right] \\
&= \frac{1}{2}\mathbb{E}_{\vec{\sigma}}\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i h(x_i)\right] \\
&= \frac{1}{2} R_S(\mathcal{H})
\end{aligned}$$

■

Remark. It then becomes natural to bound $R_n(\mathcal{H})$.

Definition 4.1.1 (Growth Function)**.** The growth function $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ for a hypothesis class $\mathcal{H}$ that maps to $y = \{\pm 1\}$ is defined as

$$\Pi_{\mathcal{H}}(n) = \sup_{x_1, \ldots, x_n \in \mathcal{X}} |\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\}|$$

Remark. This definition defines the set of all possible predictions on a given set of inputs.

Theorem 4.1.1 (Generalization bound using growth function)**.** Let $\mathcal{H}$ be a hypothesis class taking values $y = \{\pm 1\}$. Then

$$R_n(\mathcal{H}) \leq \sqrt{\frac{2 \log \Pi_{\mathcal{H}}(n)}{n}}$$