Note on Machine Learning Theory

Zepeng CHEN
The HK PolyU

Date: March 8, 2023

1 Introduction

Definition 1.1 (Supervised Learning)

Definition 1.2 (Unsupervised Learning)

Definition 1.3 (Online Learning)

Definition 1.4 (Mathematical Machine Learning Model)

When we consider machine learning theory rather than classical statistics theory, it is because

- Statistics mostly focuses on asymptotic scenarios, and does not provide non-asymptotic guarantees.
- Statistics requires a well-behaved statistical model, e.g., distribution. However, real data, e.g., text and image, could be more complex.
- Statistics aims to find the entire prob distribution, which could be too costly to compute and analyze.

2 Probability Inequalities

Lemma 2.1 (The Union Bound)

Consider events $A_1, \cdots, A_t$, we have

$$\mathbb{P}(A_1 \cup \cdots \cup A_t) \leq \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_t).$$

Note on *This bound is very useful in machine learning theory.*

2.1 Chernoff Bounds

Theorem 2.1 (Chernoff Bounds)

Let X be a random variable with moment generating function $M(t) = \mathbb{E}[e^{tX}]$. Then for any $\epsilon > 0$

$$P\{X \geq \epsilon\} \leq e^{-t\epsilon}M(t) \leq \inf_{t>0} e^{-t\epsilon}M(t) \quad \forall t > 0$$
$$P\{X \leq \epsilon\} \leq e^{-t\epsilon}M(t) \quad \forall t < 0$$

Proof For $t > 0$, based on Markov's inequality, we have

$$P\{X \geq \epsilon\} = P\left\{e^{tX} \geq e^{t\epsilon}\right\} \leq E\left[e^{tX}\right] e^{-t\epsilon}$$

And similarly, we can get another bound. Since the Chernoff bounds hold for all t in either the positive or negative quadrant, we obtain the best bound by using the t that minimizes $e^{-t\epsilon}M(t)$. ■

Corollary 2.1 (Chernoff Bounds for i.i.d. Samples)

Suppose X_i are i.i.d., then for any $t > 0$,

$$\mathbb{P}(\hat{\mu}_n - \mu \geq \epsilon) \leq (M_{X-\mu}(t)e^{-t\epsilon})^n \leq (\inf_{t>0} M_{X-\mu}(t)e^{-t\epsilon})^n.$$

Proof The important result is that Chernoff bounds "play nicely" with summations Ng, 2022, which is a consequence of the moment generating function, i.e., if X_i are independent, then

$$M_{X_1+\cdots+X_n}(t) = \prod_{i=1}^{n} M_{X_i}(t).$$

Thus,

$$\begin{aligned}
\mathbb{P}(\hat{\mu}_n - \mu \geq \epsilon) &= \mathbb{P}(\sum_{i=1}^{n} X_i - n\mu \geq n\epsilon) \\
&\leq e^{-tn\epsilon} M_{X_1+\cdots+X_n-n\mu}(t) \quad \text{(Chernoff Bounds)} \\
&= e^{-tn\epsilon} \prod_{i=1}^{n} M_{X_i-\mu}(t) \\
&= (M_{X-\mu}(t)e^{-t\epsilon})^n
\end{aligned}$$

■

Note on *The exponetial decay shown by Chernoff's inequality will be much faster than the $O(\frac{1}{n})$ decay suggested by Chebyshev's inequality.*

Corollary 2.2 (Chernoff Tail Bound for Gaussians)

The optimized Chernoff tail bound for $X \sim \mathcal{N}(0, \sigma^2)$ will be

$$\mathbb{P}(X \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2\sigma^2}).$$

Proof

$$\begin{aligned}\mathbb{P}(X \geq \epsilon) &\leq e^{-t\epsilon} M(t) \quad \text{(Chernoff's Bounds)} \\ &\leq \inf_{t>0} \exp(\frac{t^2\sigma^2}{2} - t\epsilon) \\ &= \exp(\inf_{t>0} \frac{t^2\sigma^2}{2} - t\epsilon) \quad (\exp \text{ is increasing}) \\ &= \exp(-\frac{\epsilon^2}{2\sigma^2}) \quad (t^* = \frac{\epsilon}{\sigma^2})\end{aligned}$$

■

Corollary 2.3 (Chernoff Tail Bound for Gaussian Samples)

Given IID samples $x_1, \cdots, x_n \sim \mathcal{N}(\mu, \sigma^2)$ we have the following error bound for empirical mean $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n x_i$:

$$\mathbb{P}(\hat{\mu}_n - \mu \geq \epsilon) \leq \exp(-\frac{n\epsilon^2}{2\sigma^2}).$$

Proof A direct result of Corollary 2.1 and Corollary 2.1. ■

2.2 Hoffeding's Inequality for Bounded Random Variables

Lemma 2.2 (Hoffeding's Lemma)

Suppose that r.v. X is bounded and satisfies $a \leq X \leq b$ for scalars $a, b \in \mathbb{R}$. Then, X is sub-Gaussian with parameter $\frac{(b-a)^2}{4}$, i.e., we have

$$\mathbb{E}[e^{t(X-\mathbb{E}[X])}] \leq \exp(\frac{(b-a)^2t^2}{8}).$$

Proof WLOG, by replacing Z by $X - \mathbb{E}[X]$, we can assume $\mathbb{E}[Z] = 0$, so that $a \leq 0 \leq b$. Since e^{tz} is convex, we have that for all $z \in [a, b]$,

$$e^{tz} \leq \frac{b-z}{b-a}e^{ta} + \frac{z-a}{b-a}e^{tb}.$$

Thus,

$$\begin{aligned}\mathbb{E}[e^{tZ}] &\leq \frac{b-\mathbb{E}[Z]}{b-a}e^{ta} + \frac{\mathbb{E}[Z]-a}{b-a}e^{tb} \\ &= \frac{b}{b-a}e^{ta} + \frac{-a}{b-a}e^{tb} \quad (\mathbb{E}[Z] = 0) \\ &= \exp(-\gamma u + \log(\gamma e^u + (1-\gamma))) = \exp(g(u)),\end{aligned}$$

where $u = t(b-a)$ and $\gamma = -\frac{a}{b-a}$ and the last equality can be established by solving $e^{g(u)} = \frac{b}{b-a}e^{ta} + \frac{-a}{b-a}e^{tb}$. Note that $g(0) = g'(0) = 0$ and $g''(u) \leq \frac{1}{4}$. By Taylor's theorem, we have

$$\exp(g(u)) = \exp(g(0) + ug'(0) + \frac{u^2}{2}g''(\varepsilon)) = \exp(\frac{u^2}{2}g''(\varepsilon)) \leq \exp(\frac{u^2}{8}) = \exp(\frac{(b-a)^2t^2}{8}).$$

■

Theorem 2.2 (Hoffeding's Inequality (liao_notes_2020))

Suppose that r.v. $X_1, \cdots, X_n$ *are independent and bounded as* $a_i \leq X_i \leq b_i$. *Then, defining the empirical mean* $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ *and underlying mean* $\mu = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i]$ *results in the following concentration inequality:*

$$\mathbb{P}(\mu - \hat{\mu}_n \geq \epsilon) \leq \exp(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}),$$
$$\mathbb{P}(\hat{\mu}_n - \mu \geq \epsilon) \leq \exp(-\frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}).$$

Particularly, if $a_i = a$ *and* $b_i = b$, *we have*

$$\mathbb{P}(\hat{\mu}_n - \mu \geq \epsilon) \leq \exp(-\frac{2n\epsilon^2}{(b-a)^2}).$$

Proof

■

Theorem 2.3 (McDiarmid's Inequality (liao_notes_2020))

Let $f : \mathcal{X}^n \to \mathbb{R}$ *be a function such that for every* $x_1, \cdots, x_n, x_1', \cdots, x_n' \in \mathcal{X}$ *the following bounded differences condition holds:*

$$\forall 1 \leq i \leq n : |f(x_1, \cdots, x_i, \cdots, x_n) - f(x_1, \cdots, x_i', \cdots, x_n)| \leq c_i.$$

Then, assuming $X_1, \cdots, X_n \in \mathcal{X}$ *are independent r.v., we have*

$$\mathbb{P}(f(X_1, \cdots, X_n) - \mathbb{E}[f(X_1, \cdots, X_n)] \geq \epsilon) \leq \exp(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}),$$
$$\mathbb{P}(\mathbb{E}[f(X_1, \cdots, X_n)] - f(X_1, \cdots, X_n) \geq \epsilon) \leq \exp(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}).$$

Note on *Hoeffding's inequality is a special case of McDiarmid's inequality for*

$$f(x_1, \cdots, x_n) = \frac{1}{n}\sum_{i=1}^n x_i.$$

3 Small Sample Theory

We summarize some useful definitions and theorems for limited sample theory which are not covered in large sample theory. For definitions and theorems in large sample theory, please refer to my notes on statistics.

3.1 Sub-Gaussian Random Variables

Definition 3.1 (Sub-Gaussian Random Variables)

X with mean μ is called as a sub-Gaussian r.v. with parameter σ^2 if the MGF of $X-\mu$ ($M_{X-\mu}$) satisfies the following inequality at every $t \in \mathbb{R}$:

$$M_{X-\mu}(t) := \mathbb{E}[e^{t(X-\mu)}] \leq \exp(\frac{\sigma^2 t^2}{2}).$$

Corollary 3.1 (Sum of Independent Sub-Gaussians)

If $X_1, \cdots, X_n$ are independent sub-Gaussian r.v. with parameters $\sigma_1^2, \cdots, \sigma_n^2$, then $\sum_{i=1}^n X_i$ will be sub-Gaussian with parameter $\sum_{i=1}^n \sigma_i^2$.

Proof

$$\begin{aligned} M_{\sum_{i=1}^n X_i - \mathbb{E}[\sum_{i=1}^n X_i]}(t) &= \mathbb{E}[e^{t(\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i)}] \\ &= \prod_{i=1}^n \mathbb{E}[e^{t(Z_i-\mu_i}] \quad \text{(Independence)} \\ &\leq \prod_{i=1}^n \exp(\frac{\sigma_i^2 t^2}{2}) = \exp(\frac{\sum_{i=1}^n \sigma_i^2 t^2}{2}) \end{aligned}$$

■

Corollary 3.2 (Scalar Product of Sub-Gaussians)

If X is sub-Gaussian r.v. with parameter σ^2, then cX for scalar $c \in \mathbb{R}$ will be sub-Gaussian with parameter $c^2\sigma^2$.

Corollary 3.3 (Chernoff Tail Bound for Sub-Gaussians)

The optimized Chernoff tail bound for a sub-Gaussian X with parameter σ^2 and mean μ will be

$$\mathbb{P}(X - \mu \geq \epsilon) \leq \exp(-\frac{\epsilon^2}{2\sigma^2}).$$

Proof

$$\begin{aligned} \mathbb{P}(X-\mu \geq \epsilon) &\leq e^{-t\epsilon} M_{X-\mu}(t) \quad \text{(Chernoff's Bounds)} \\ &\leq \exp(\frac{t^2\sigma^2}{2} - t\epsilon) \quad \text{(Definition of Sub-Gaussians)} \\ &\leq \exp(\inf_{t>0} \frac{t^2\sigma^2}{2} - t\epsilon) \quad (\exp \text{ is increasing}) \\ &= \exp(-\frac{\epsilon^2}{2\sigma^2}) \quad (t^* = \frac{\epsilon}{\sigma^2}) \end{aligned}$$

■

Corollary 3.4 (Chernoff-based Concentration Inequality for Sub-Gaussians)

If $x_1, \cdots, x_n$ are i.i.d. samples for sub-Gaussian X with parameter σ^2 and mean μ, we have the following error bound on their empirical mean $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n x_i$:

$$\mathbb{P}(\hat{\mu}_n - \mu \geq \epsilon) \leq \exp(-\frac{n\epsilon^2}{2\sigma^2}).$$

Corollary 3.5 (Chernoff-based Concentration Inequality for Bounded Random Variables)

If $x_1, \cdots, x_n$ are i.i.d. samples for a bounded r.v. $a \leq X \leq b$ with mean μ, we have the following error bound on their empirical mean $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$:

$$\mathbb{P}(\hat{\mu}_n - \mu \geq \epsilon) \leq \exp(-\frac{2n\epsilon^2}{(b-a)^2}).$$

Proof A direct result of Corollary 3.1 and Hoffeding's Lemma. ■

4 Model-based Statistical Learning

This section covers exponential families, maximum likelihood, method of moments and maximum entropy principle, which overlaps statistics. For more detail, please refer to my note on Advanced Statistics directly.

4.1 Information Theory

Definition 4.1

Given a probability vector $\mathbf{q} = [q_1, \cdots, q_k]$ for a discrete random variable X, the (Shannon) entropy of X is defined as

$$H_{\mathbf{q}}(X) = \sum_{i=1}^{k} q_i log \frac{1}{q_i} = -\mathbb{E}_q log q.$$

Note on *Entropy is simply the negative expected loglikelihood.*

Note on *The entropy value is always non-negative and concave.*

Note on *The entropy measures the uncertainty in a given distribution. Moreover, the entropy is upper-bounded by $logk$ (Jensen's Inequality). Particularly, the upper-bound is achieved by the discrete uniform distribution, i.e., $q_1 = \cdots = q_k = \frac{1}{k}$. This can be proved by solving the entropy maximization problem:*

$$\begin{aligned} \max_{\mathbf{q}\in\mathbb{R}^k} \quad & \sum_{i=1}^{k} q_i log \frac{1}{q_i} \\ s.t. \quad & \sum_{i=1}^{k} q_i = 1, \\ & q_i \geq 0, i = 1, \cdots, k. \end{aligned}$$

Definition 4.2 (Conditional Entropy)

Definition 4.3 (Joint Entropy)

Note on *Cross-entropy is also the negative expected loglikelihood, and is not calculated under its truth, but under some other distribution.*

4.2 Maximum Entropy Principle

Definition 4.4 (Empirical Distribution)

Suppose we have n observations such as $x_1, \cdots, x_n$ from an unknown distribution p. The empirical distribution is defined as $\tilde{p} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x = x_i)$.

Definition 4.5 (Maximum Entropy Principle)

Given samples $\mathbf{x}_1, \cdots, \mathbf{x}_n$, we want to find a model in a set of probability distributions

$$M_\phi := \left\{ q \in \mathcal{P}_\mathcal{X} : \mathbb{E}_{\hat{\boldsymbol{\theta}}}[\phi(\mathbf{x})] = \frac{1}{n}\sum_{i=1}^{n} \phi(\mathbf{x}_i) \right\},$$

conduct the inference and base the decision on the distribution maximizing the Entropy function:

$$\underset{q \in M_\phi}{\operatorname{argmax}} H_q(\mathbf{X}) := \sum_{\mathbf{x} \in \mathcal{X}} q(\mathbf{x}) \log \frac{1}{q(\mathbf{x})}.$$

Note on *This principle chooses the most uncertain model based on the given set M.*

Theorem 4.1

The distribution that maximizes the entropy is an exponential family model with feature function ϕ.

Proof Consider the maximum entropy problem

$$\begin{aligned} \max_{\mathbf{q} \in \mathbb{R}^{|\mathcal{X}|}} \quad & \sum_{\mathbf{x} \in \mathcal{X}} q_\mathbf{x} log \frac{1}{q_\mathbf{x}} = -\sum_{\mathbf{x} \in \mathcal{X}} q_\mathbf{x} log q_\mathbf{x} \\ \text{s.t.} \quad & \sum_{\mathbf{x} \in \mathcal{X}} q_\mathbf{x} \phi(\mathbf{x}) = \hat{\boldsymbol{\mu}}, \\ & \sum_{\mathbf{x} \in \mathcal{X}} q_\mathbf{x} = 1, \\ & q_\mathbf{x} \geq 0, \mathbf{x} \in \mathcal{X}, \end{aligned}$$

as a problem without inequality constraints, i.e.,

$$\begin{aligned} \max_{\mathbf{q} \in \mathbb{R}^{|\mathcal{X}|}} \quad & -\sum_{\mathbf{x} \in \mathcal{X}} q_\mathbf{x} log q_\mathbf{x} \\ \text{s.t.} \quad & \sum_{\mathbf{x} \in \mathcal{X}} q_\mathbf{x} \begin{bmatrix} \phi(\mathbf{x}) \\ 1 \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{\mu}} \\ 1 \end{bmatrix}. \end{aligned}$$

Next we consider its Lagrangian problem

$$\mathcal{L}(\mathbf{q}, \boldsymbol{\gamma}) = \sum_{\mathbf{x} \in \mathcal{X}} q_\mathbf{x} \left(-log q_\mathbf{x} - \phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k} - \gamma_{k+1} \right) + \hat{\boldsymbol{\mu}}^\top \boldsymbol{\gamma}_{1:k} + \gamma_{k+1},$$

the stationary KKT condition

$$\nabla_{q_\mathbf{x}} \mathcal{L}(\mathbf{q}, \boldsymbol{\gamma}) = -log q^*_\mathbf{x} - \phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k} - \gamma_{k+1} + 1 = 0$$

leads to

$$q^*_\mathbf{x} = exp\left(-\phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k} - \gamma_{k+1} + 1 \right) \geq 0.$$

Thus, $q^*_{\mathbf{x}}$ is also the optimal solution to the original problem. Moreover,

$$q^*_{\mathbf{x}} \propto exp\left(-\phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k}\right)$$

leads to

$$q^*_{\mathbf{x}} = \frac{exp\left(-\phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k}\right)}{\sum_{x\in\mathcal{X}} exp\left(-\phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k}\right)}$$

due to the constraint that probability $q_{\mathbf{x}}$'s add up to 1. ■

Note on *Suppose $\Omega = \{0, 1\}$ and $\phi(x) = x$, then the maximum entropy principle leads to a binomial distribution. Suppose $\Omega = (-\infty, \infty)$ and $\phi(x) = [x, x^2]^\top$, then the maximum entropy principle leads to a Gaussian distribution.*

4.3 Maximum Relative Entropy Principle (Meila, 2012, Lec. 8)

Definition 4.6

Based on the logic of maximum entropy principle, suppose we also have a prior distribution q_0.

4.4 Minimum KL-Divergence

Definition 4.7 (Kullback-Leibler Divergence)

Let X be a random variable with possible outcomes $\mathcal{X}$ and let P and Q be two probability distributions on X. The KL-Divergence of P from Q is defined as:

$$KL[P||Q] = \sum_{x\in\mathcal{X}} p(x) log_b \frac{p(x)}{q(x)} = \mathbb{E}_p \log \frac{p}{q},$$

$$KL[P||Q] = \int_{\mathcal{X}} p(x) log_b \frac{p(x)}{q(x)} dx.$$

Note on *KL-Divergence captures how much a model distribution function differs from the true distribution of the data. However, since KL-Divergence is asymmetric in (p, q). We should not call it as the 'distance' between two distributions.*

Note on Perspective from Statistics (halvorsen, 2016) *If we have two hypothesis regarding which distribution is generating the data X, e.g., P and Q. Then $\frac{p(x)}{q(x)}$ is the likelihood ratio for testing H_0: Q against H_1: P. Since KL-Divergence $KL[P||Q]$ is the expected value of the loglikelihood ratio under the alternative hypothesis, so it is a measure of the difficulty of this test. The asymmetry of KL-Divergence actually reflects the asymmetry between null and laternative hypothesis.*

For example, let P be the t_1-distribution and Q be the standard normal distribution. Then

$$KL(P||Q) \approx \infty,$$

$$KL(Q||P) \approx 0.26.$$

That is, if the null model is normal but the data is generated from t-distribution, then it is quite

easy to reject the null! The logic here is that data from t-distribution do not look like normal. However, if the null is t and data is normal. Normal distributed data could look like t data.

Lemma 4.1

KL-Divergence is always non-negative. Moreover, $KL(p||q) = 0$ iff $p = q$.

Proof By Gibbs' Inequality. ■

Lemma 4.2

KL-Divergence is convex in the pair of probability distributions (p, q), i.e.,

$$KL[\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2] \leq \lambda KL[p_1||q_1] + (1-\lambda)KL[p_2||q_2],$$

where (p_1, q_1) and (p_2, q_2) are two pairs of probability distributions and $0 \leq \lambda \leq 1$. Moreover, KL-divergence attains the minimum 0 if $p = q$.

Proof Firstly, we show that KL-Divergence is bi-convex of p, q, i.e., it is a convex function of p for a fixed q and vice versa,

$$KL[\lambda p_1 + (1-\lambda)p_2||q_1] \leq \lambda KL[p_1||q_1] + (1-\lambda)KL[p_2||q_1],$$
$$KL[p_1||\lambda q_1 + (1-\lambda)q_2] \leq \lambda KL[p_1||q_1] + (1-\lambda)KL[p_1||q_2]$$

on the basis of $xlogx$'s convexity. Next we can prove the convexity from bi-convexity. ■

4.5 Connections

4.5.1 Maximum Entropy and MLE

Theorem 4.2 (Maximum Entropy v.s. MLE)

The maximum entropy problem over M_ϕ is the dual problem to the MLE for the exponential family with feature function ϕ.

Proof Suppose we replace the Lagrangian problem with $q^*_{\mathbf{x}}$, the dual of the maximum entropy problem can be written as

$$\min_{\boldsymbol{\gamma}} log\left(\sum_{x\in\mathcal{X}} exp\left(-\phi(\mathbf{x})^\top \boldsymbol{\gamma}_{1:k}\right)\right) + \hat{\boldsymbol{\mu}}^\top \boldsymbol{\gamma}_{1:k}.$$

■

4.5.2 Minimum KL-Divergence and MLE

Theorem 4.3

Let $\{q_\theta\}_{\theta\in\Theta}$ be a parametric family of distributions, and suppose $p_n(x)$ is the empirical pdf from n samples, MLE is minimizing KL-Divergence, i.e.,

$$\arg\min_{\theta\in\Theta} KL(p_n||q_\theta) = \arg\max_{\theta} p(x|\theta).$$

Proof

$$\begin{aligned}\arg\min_{\theta} KL(p||q) &= \arg\min_{\theta} \mathbb{E}_{x\sim p}\left[log\frac{p(x)}{q(x)}\right] \\ &\iff \arg\min_{\theta} \mathbb{E}_{x\sim p}\left[-logq(x)\right] \\ &\iff \arg\max_{\theta} \mathbb{E}_{x\sim p}\left[logq(x)\right]\end{aligned}$$

■

4.5.3 Minimum KL-Divergence and Maximum Entropy

Theorem 4.4 (Minimum KL-Divergence and Maximum Entropy)

The model with maximum entropy is equivalent to the minimum KL divergence to the uniform distribution.

5 Model-free Machine Learning

Compared to model-based learning, here the underlying distribution is unknown. We introduce the theory of supervised learning in this section.

Definition 5.1

A standard goal in supervised learning is to minimize the averaged prediction loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^{\geq 0}$ *where* $\ell(\hat{y}, y)$ *measures the loss suffered under prediction* $\hat{y}$ *for actual label* y.

Note on Example of Loss Function

- *Squared-error loss:* $\ell_2(\hat{y}, y) = (\hat{y} - y)^2$,
- *0-1 loss:* $\ell_{0/1}(\hat{y}, y) = I(\hat{y} \neq y)$.

Definition 5.2 (Population Risk Minimization)

Given ℓ, *the supervised learning goal is to find a prediction function* $f \in \mathcal{F}$ *to minimize the expected loss under distribution* $P_{X,Y}$, *i.e., population risk:*

$$\min_{f\in\mathcal{F}} L(f) = \mathbb{E}_{P_{X,Y}}[\ell(f(X), Y)].$$

Let f^* *denotes the function with minimum population risk, i.e.,* $f^* = \arg\min_{f\in\mathcal{F}} L(f)$.

Note on *This problem cannot be solved, since we do not know* $P_{X,Y}$.

Definition 5.3 (Empirical Risk Minimization)

Given loss function ℓ *and training data* $\{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)\}$, *the empirical risk minimization (ERM) approach finds the prediction rule* $\hat{f} \in \mathcal{F}$ *minimizing the empirical*

expected loss or empirical risk:

$$\min_{f\in\mathcal{F}} \hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), y_i).$$

Let $\hat{f}$ denotes the function with minimum empirical risk, i.e., $\hat{f} = \arg\min_{f\in\mathcal{F}} \hat{L}(f)$.

Definition 5.4 (Generalization Risk)

The generalization risk of a prediction function $\hat{f} \in \mathcal{F}$ is defined as the difference between its empirical and population risks:

$$\epsilon_{gen}(\hat{f}) = L(\hat{f}) - \hat{L}(\hat{f}).$$

Definition 5.5 (Excess Risk)

The excess risk of a prediction function $\hat{f} \in \mathcal{F}$ is defined as the difference between its population risk and the population risk of optimal f^*:

$$\begin{aligned}\epsilon_{excess}(\hat{f}) &= L(\hat{f}) - L(f^*) \\ &= \underbrace{L(\hat{f}) - \hat{L}(\hat{f})}_{\text{Gen. Risk for } \hat{f}} + \underbrace{\hat{L}(\hat{f}) - \hat{L}(f^*)}_{\leq 0} + \underbrace{\hat{L}(f^*) - L(f^*)}_{\text{Gen. Risk for } f^*}.\end{aligned}$$

Note on *The excess risk of every function in $\mathcal{F}$ is non-negative, but the generalization risk may be negative.*

Note on *While f^* is a deterministic function, $\hat{f}$ is a random function affected by the randomness of the training samples.*

Note on A sufficient condition for an ϵ-bounded excess risk

$$\begin{aligned}\epsilon_{excess}(\hat{f}) &\leq \underbrace{|L(\hat{f}) - \hat{L}(\hat{f})|}_{\text{Gen. Risk for } \hat{f}} + \underbrace{|\hat{L}(f^*) - L(f^*)|}_{\text{Gen. Risk for } f^*} \\ &\leq 2\sup_{f\in\mathcal{F}} |L(f) - \hat{L}(f)|.\end{aligned}$$

Thus, if $\sup_{f\in\mathcal{F}} |L(f) - \hat{L}(f)| \leq \frac{\epsilon}{2}$, then $\epsilon_{excess}(\hat{f}) \leq \epsilon$.

5.1 Finite Hypothesis Sets & Uniform Convergence Bounds

Theorem 5.1 (Population Risk Bound for Finite Function Sets)

Given

1. *0-1 loss: $\ell_{0/1}(\hat{y}, y) = I(\hat{y} \neq y)$,*
2. *a realizable scenario where $L(f^*) = 0$,*
3. *and a finite function set $\mathcal{F} = \{f_1, \cdots, f_t\}$ with t functions,*

the population risk bound holds for the ERM solution $\hat{f}$ with prob at least $1-\delta$:

$$L(\hat{f}) \leq \frac{\log t + \log\frac{1}{\delta}}{n}.$$

Proof Firstly, the realizability assumption implies that $L(f^*) = 0$ and $\hat{L}(\hat{f}) = 0$. Why?

For $\epsilon \geq 0$, define $F_\epsilon = \{f \in \mathcal{F} : L(f) \geq \epsilon\}$. Then our goal is to bound the probability $\mathbb{P}(\hat{f} \in F_\epsilon)$, i.e.,

$$\mathbb{P}(\hat{f} \in F_\epsilon) = \mathbb{P}().$$

TBD ■

Note on *The risk bound is based on two restrictive assumptions: the realizability condition and the finiteness of the hypothesis set $\mathcal{F}$.*

Theorem 5.2 (Excess Risk Bound for Finite Function Sets with Realizability Assumption)

Given

1. *0-1 loss: $\ell_{0/1}(\hat{y}, y) = I(\hat{y} \neq y)$,*
2. *a realizable scenario where $L(f^*) = 0$,*
3. *and a finite function set $\mathcal{F} = \{f_1, \cdots, f_t\}$ with t functions,*

the excess risk bound holds for the ERM solution $\hat{f}$ with prob at least $1 - \delta$:

$$\epsilon_{excess}(\hat{f}) \leq \frac{\log t + \log \frac{1}{\delta}}{n} = \mathcal{O}(\frac{\log(t/\delta)}{n}).$$

Proof TBD ■

Theorem 5.3 (Uniform Convergence Bound for the Excess Risk)

Given the best population and empirical risk functions $f^*, \hat{f} \in \mathcal{F}$, the probability of an ϵ-large excess risk is bounded as:

$$\mathbb{P}(L(\hat{f}) - L(f^*) \geq \epsilon) \leq \mathbb{P}(\sup_{f \in \mathcal{F}} |L(f) - \hat{L}(f)| \geq \frac{\epsilon}{2}).$$

Note on *If we can show that*

Theorem 5.4 (Excess Risk Bound for Finite Function Sets without Realizability Assumption)

Given

1. *0-1 loss: $\ell_{0/1}(\hat{y}, y) = I(\hat{y} \neq y)$,*
2. *and a finite function set $\mathcal{F} = \{f_1, \cdots, f_t\}$ with t functions,*

the excess risk bound holds for the ERM solution $\hat{f}$ with prob at least $1 - \delta$:

$$\epsilon_{excess}(\hat{f}) \leq \sqrt{\frac{2\log t + 2\log \frac{2}{\delta}}{n}} = \mathcal{O}(\sqrt{\frac{\log(t/\delta)}{n}}).$$

Proof TBD ■

Note on Difference between bounds with and without realizability assumption

- *The $O(\frac{1}{n})$ risk bound in the realizable case (noiseless setting) is vanishing faster than the $O(\frac{1}{\sqrt{n}})$ bound in the non-realizable case (noisy setting).*
- *In learning theory, the risk bounds that decay with $O(\frac{1}{n})$ are called fast rates bounds, which require extra assumptions on the learning setting, e.g. realizability or norm-based*

regularization.

5.2 Infinite Hypothesis Sets & Rademacher Complexity

Definition 5.6 (Worst-case Generalization Risk)

Given a random dataset of size n, the worst-case generalization risk is defined as

$$G_n := \sup_{f \in \mathcal{F}} L(f) - \hat{L}(f).$$

Lemma 5.1 (Concentration Bound for Worst-case Generalization Risk)

Suppose the loss function is bounded as $0 \leq \ell(y, \hat{y}) \leq c$. Consider the worst-case generalization risk G_n as a function of independent empirical samples $X_1, \cdots, X_n$. Then

$$\mathbb{P}(G_n \geq \mathbb{E}[G_n] + \epsilon) \leq \exp(-\frac{2n\epsilon^2}{c^2}).$$

Proof Note that the assumption implies that Why?

$$\forall 1 \leq i \leq n : |g(x_1, \cdots, x_i, \cdots, x_n) - g(x_1, \cdots, x_i', \cdots, x_n)| \leq \frac{c}{n}.$$

■

Note on Example *0-1 loss satisfies for $c = 1$.*

Definition 5.7 (Rademacher Random Variable)

A Rademacher r.v. σ is defined as uniformly distributing over $\{-1, +1\}$, i.e.,

$$\mathbb{P}(\sigma = 1) = \mathbb{R}(\sigma = -1) = \frac{1}{2}$$

.

Note on Application *To get rid of the virtual dataset, we use independent Rademacher r.v. $\sigma_1, \cdots, \sigma_n$. Due to the symmetry, the following equations hold:*

$$\begin{aligned} & X_i - X_i' \overset{dist.}{=} \sigma_i(X_i - X_i') \\ \Rightarrow & \frac{1}{n}\sum_{i=1}^{n}(X_i - X_i') \overset{dist.}{=} \frac{1}{n}\sum_{i=1}^{n}(\sigma_i X_i - \sigma_i X_i') \\ \Rightarrow & L_S(f) - L_{S'}(f) \overset{dist.}{=} \frac{1}{n}\sum_{i=1}^{n}(\sigma_i X_i - \sigma_i X_i'). \end{aligned}$$

Lemma 5.2 (x)

Introducing a virtual dataset $S' = \{X_1', \cdots, X_n'\}$ including n new samples independent from dataset S and denote $L(f) = \mathbb{E}[\hat{L}_{S'}(f)]$. The expected worst-case generalization risk can be bounded as

$$\begin{aligned} \mathbb{E}[G_n] &\leq \mathbb{E}_{S,S'}[\sup_{f \in \mathcal{F}} L_{S'}(f) - L_S(f)] \quad (\textit{Symmetrization Bound}) \\ &\leq 2\mathbb{E}_{S,\boldsymbol{\sigma}}[\sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i \ell(f(\boldsymbol{x}_i, y_i)]. \end{aligned}$$

Proof

■

Definition 5.8 (Rademacher Complexity)

For a function set $\mathcal{H}$ and Rademacher variables in $\boldsymbol{\sigma} = [\sigma_1, \cdots, \sigma_n]$, we define $\mathcal{H}$'s Rademacher complexity as

$$R_n(\mathcal{H}) := \mathbb{E}_{S,\boldsymbol{\sigma}}[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(X_i)].$$

Note on Motivation from binary classification **(chen_cos_2013)** *Suppose f is a classification function which maps data x_i to its label $\sigma_i \in \{-1, 1\}$. Since f is dependent on σ_i, to measure how well $\mathcal{H}$ can correlate with random noise, we take the expectation of the correlation over σ_i, i.e., Rademacher complexity. This intuitively measures the expressiveness of $\mathcal{H}$. For example, $|\mathcal{H}| = 1$ where we only have one choice for a hypothesis, our expectation equals 0 since the max term disappears; and $|\mathcal{H}| = 2^n$ where $\mathcal{H}$ shatters S, our expectation equals 1 since there always exists a hypothesis matching any set of σ_i's. That is, this measure must fall between 0 and 1.*

Corollary 5.1 (Basic Properties of Rademacher Complexity)

1. ***Monotonicity.*** *If $\mathcal{H}_1 \subseteq \mathcal{H}_2$, then $R_n(\mathcal{H}_1) \leq R_n(\mathcal{H}_2)$.*
2. ***Singleton Set.*** *If $\mathcal{H} = \{h\}$ contains only one function, then $R_n(\mathcal{H}) = 0$.*
3. ***Scalar Product.*** *If $c\mathcal{H} = \{ch : h \in \mathcal{H}\}$, then $R_n(\mathcal{H} = |c|R_n(\mathcal{H})$.*
4. ***Lipschitz Composition.*** *If $g : \mathbb{R} \to \mathbb{R}$ is a ρ-Lipschitz function, i.e.,*

   $$\forall z, z' \in \mathbb{R} : |g(z) - g(z') \leq \rho|z - z'|,$$

   then $R_n(g \circ \mathcal{H} \leq \rho R_n(\mathcal{H})$.
5. ***Convex Hull.*** *For a function set $\mathcal{H} = \{h_1, \cdots, h_2\}$, we define its convex hull:*

   $$\textit{convex-hull}(\mathcal{H}) := \{\sum_{i=1}^{t} \alpha_i h_i : \alpha_1, \cdots, \alpha_t \geq 0, \sum_{i=1}^{t} = 1\}.$$

   Then, $R_n(\textit{convex-hull}(\mathcal{H})) = R_n(\mathcal{H})$.

Proof

■

Corollary 5.2 (Excess Risk Bound via Rademacher Complexity)

For a hypothesis set $\mathcal{F}$, define $\mathcal{H} = \{[x, y] \to \ell(f(x), y) : f \in \mathcal{F}\}$ to be the composition of loss function ℓ with the hypotheses in $\mathcal{F}$. Then, with probability at least $1 - \delta$,

$$L(\hat{f}) - L(f^*) \leq 4R_n(\mathcal{H}) + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Proof

■

Definition 5.9 (Empirical Rademacher Complexity)

For a function set $\mathcal{H}$, Rademacher variables in $\boldsymbol{\sigma} = [\sigma_1, \cdots, \sigma_n]$ and a fixed dataset $S = \{x_1, \cdots, x_n\}$, we define $\mathcal{H}$'s empirical Rademacher complexity as

$$\hat{R}_n(\mathcal{H}) := \mathbb{E}_{\boldsymbol{\sigma}}[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i)].$$

In other words,

$$R_n(\mathcal{H}) = \mathbb{E}_S[\hat{R}_n(\mathcal{H})].$$

Lemma 5.3 (Massart Lemma)

Suppose that $\mathcal{H} = \{h_1, \cdots, h_t\}$ is a finite set of t functions. Also, suppose that for every $h \in \mathcal{H}$ and dataset $S = \{x_1, \cdots, x_n\}$ the following holds:

$$\frac{1}{n} \sum_{i=1}^{n} h(x_i)^2 \leq M.$$

Then, the following bound on the empirical Rademacher complexity holds:

$$\hat{R}_n(\mathcal{H}) \leq \sqrt{\frac{2M \log t}{n}}$$

.

Proof

■

Note on *Massart lemma shows that the Rademacher complexity of a finite function set of size t is bounded by $\mathcal{O}(\sqrt{\frac{\log t}{n}})$.*

Corollary 5.3 (Empirical Rademacher Complexity of ℓ_2-Norm-bounded Linear Functions)

Consider the following set of ℓ_2-norm-bounded linear functions:

$$\mathcal{H} = \{h_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} : ||\boldsymbol{w}||_2 \leq M\}.$$

Then for a dataset $S = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$, we have the following bound on the empirical Rademacher complexity:

$$\hat{R}_n(\mathcal{H}) \leq \frac{M \max_i ||\boldsymbol{x}_i||_2}{\sqrt{n}}.$$

Corollary 5.4 (Empirical Rademacher Complexity of ℓ_1-Norm-bounded Linear Functions)

Consider the following set of ℓ_1-norm-bounded linear functions on a d-dimensional $\boldsymbol{x} \in \mathbb{R}^d$:

$$\mathcal{H} = \{h_{\boldsymbol{w}}(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} : ||\boldsymbol{w}||_1 \leq M\}.$$

Then for a dataset $S = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$, we have the following bound on the empirical Rademacher complexity:

$$\hat{R}_n(\mathcal{H}) \leq M \max_i ||\boldsymbol{x}_i||_\infty \sqrt{\frac{2 \log(2d)}{n}}.$$

5.2.1 Rademacher Complexity of ReLU-based Neural Nets

Definition 5.10 (Frobenius Norm)

Given a matrix $W \in \mathbb{R}^{d \times t}$, we define its Frobenius norm as

$$||W||_F = \sqrt{\sum_{i=1}^{d} \sum_{j=1}^{t} w_{ij}^2}.$$

Definition 5.11 (ReLU Function)

The ReLU function is defined as $\psi_{ReLU}(x) = \max\{0, x\}$.

Corollary 5.5 (Rademacher Complexity of ReLU-based Nueral Nets)

Consider the following set of L-layer neural nets with ReLU activation function:

$$\mathcal{H} = \{h_{\boldsymbol{w}}(\boldsymbol{x}) = W\}$$

TBD

5.3 VC Dimension

5.4 Covering Numbers

6 Theory of Representation

6.1 Kernel Functions and Methods

6.2 Approximation in Deep Learning

7 Theory of Convergence

Bibliography

halvorsen, kjetil b (Jan. 2016). *Answer to "Intuition on the Kullback–Leibler (KL) Divergence"*.

Meila, Marina (2012). *STAT 538 Statistical Learning: Modeling, Prediction and Computing*.

Ng, Andrew (2022). *CS229: Machine Learning*.