DDA4210/AIR6002 Advanced Machine Learning
Lecture 03 Learning Theory

Tongxin Li

School of Data Science, CUHK-Shenzhen

Spring 2024

Overview

What is machine learning theory

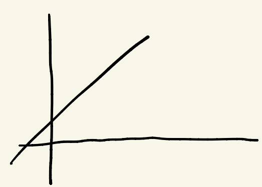
- Machine Learning Theory is also known as *Computational Learning Theory*.
- It aims to understand the fundamental principles of learning as a computational process and combines tools from Computer Science and Statistics.
  - Create mathematical models of machine learning and analyze the inherent ease or difficulty of different types of learning problems.
  - Proving guarantees for algorithms (e.g., under what conditions will they succeed, how much data and computation time is needed)
  - Developing machine learning algorithms that provably meet desired criteria.
  - Mathematically analyzing general issues (e.g., "When can one be confident about predictions made from limited data?", "What kinds of methods can learn even in the presence of large quantities of distracting information?")

A toy example: $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$ (Classification Task)

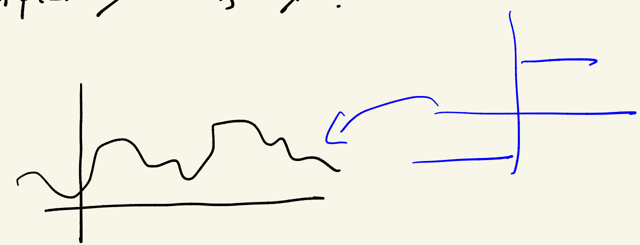
e.g. $\{(2, +1), (1, +1), (0.5, +1), (-2, -1), (-1, -1), (-0.5, -1)\}$

$(y = \mathrm{sign}(x))$

- Which function class does y belong to? e.g. linear y?
- How much data do we need? (sample complexity)
- How to measure the "complexity" of y?



simple ?



complicated ?

Basic notation

$\mathcal{X} \xrightarrow{f} \mathcal{Y}$

$x \in \mathcal{X}$, f, $x \to y$, $y \in \mathcal{Y}$

- Input space/feature space : $\mathcal{X}$
  - Feature is a numerical description for a sample or object.
  - Feature extraction is an art.
- Output space/label space: $\mathcal{Y}$
  - E.g.: $\{+1, -1\}$, $\{1, 2, \ldots, K\}$, $\mathbb{R}$-valued output, structured output.
- Loss function: $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$
  - E.g.: $0-1$ loss $\ell(y, \hat{y}) = 1\{y \neq \hat{y}\}$, square loss $\ell(y, \hat{y}) = (y - \hat{y})^2$, absolute loss $\ell(y, \hat{y}) = |y - \hat{y}|$, cross-entropy loss $\ell(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$.
  - It measures performance/cost per instance (e.g., inaccuracy or error of prediction).
- Model class/hypothesis class: $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ (or $\mathcal{H}$ or $\mathbb{H}$)
  - E.g.: $\mathcal{F} = \left\{x \mapsto f^\top x : \|f\|_2 \leq 1\right\}$, $\mathcal{F} = \left\{x \mapsto \operatorname{sign}\left(f^\top x\right)\right\}$

    $\mathcal{F} = \{\text{continuous } f\}$

(optional) $\mathcal{F}$ = RKHS (Reproducing kernel Hilbert space)

$\mathcal{F}$ = NTK (Neural Tangent kernel)

etc

Basic notation

$\mathcal{X} \xrightarrow{f} \mathcal{Y}$

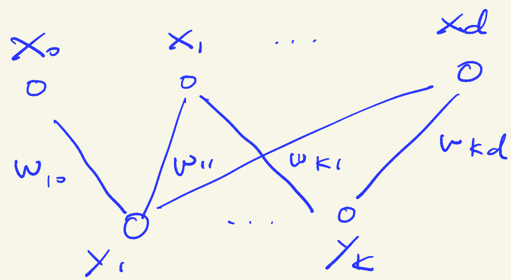
- Input space/feature space : $\mathcal{X}$
  - Feature is a numerical description for a sample or object.
  - Feature extraction is an art.
- Output space/label space: $\mathcal{Y}$
  - E.g.: $\{+1, -1\}$, $\{1, 2, \ldots, K\}$, $\mathbb{R}$-valued output, structured output.
- Loss function: $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$
  - E.g.: $0-1$ loss $\ell(y, \hat{y}) = 1\{y \neq \hat{y}\}$, square loss $\ell(y, \hat{y}) = (y - \hat{y})^2$, absolute loss $\ell(y, \hat{y}) = |y - \hat{y}|$, cross-entropy loss $\ell(y, \hat{y}) = -y \log \hat{y} - (1 - y) \log(1 - \hat{y})$.
  - It measures performance/cost per instance (e.g., inaccuracy or error of prediction).
- Model class/hypothesis class: $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ (or $\mathcal{H}$ or $\mathbb{H}$)
  - E.g.: $\mathcal{F} = \left\{x \mapsto f^\top x : \|f\|_2 \leq 1\right\}$, $\mathcal{F} = \left\{x \mapsto \text{sign}\left(f^\top x\right)\right\}$

    $\mathcal{F} = \{\text{continuous } f\}$

    $\mathcal{F} =$ RKHS (Reproducing kernel Hilbert space)

    $\mathcal{F} =$ NTK (Neural Tangent kernel) etc

$\mathcal{F} = \{$ a set of single-layer NNs $\}$



$y = W \cdot x$

$\downarrow$

a K-by-$(d+1)$ matrix

(Recall: Hypothesis set $\mathcal{H}$)

Next: Three learning frameworks.

Probably approximately correct (PAC) learning

- Learner only observes training samples

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

ground truth

$\cdot$ $x_1, x_2, \ldots, x_n \sim D_X$, $y_i = f^\star(x_i)$, $i = 1, 2, \ldots, n$, where $f^\star \in \mathcal{F}$.

- Goal: find $\hat{f} \in \mathcal{Y}^{\mathcal{X}}$ to minimize

$$\mathbb{P}_{x \sim D_X}\left[\hat{f}(x) \neq f^\star(x)\right]$$

generalization error

- Probably approximately correct (PAC) [Valiant 1984] learning is a framework for mathematical analysis of machine learning.

Probably Approximately Correct (PAC) Learning

- In PAC learning, the learner receives samples and must select a generalization function (called the hypothesis) from a certain class of possible functions. The goal is that, with high probability ("probably"), the selected function will have low generalization error ("approximately correct"). The learner must be able to learn the concept given any arbitrary approximation ratio, probability of success, or distribution of the samples.
- Sample complexity (definition):
  Given $\delta > 0$, $\epsilon > 0$, *and sample complexity* $n(\epsilon, \delta)$ *is the smallest* n *such that we can always find forecaster* $\hat{f}$ *s.t. with probability at least* $1 - \delta$,

f

a

• f^* is optimal!

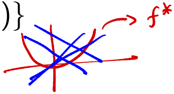
$$\mathbb{P}_{x \sim D_X} \left[\hat{f}(x) \neq f^\star(x) \right] \leq \epsilon$$

* The learner knows that there exists a perfect $f^\star$ that generates the label.

Statistical Learning (agnostic PAC)

- Learner only observes training samples $(x_i, y_i) \sim D$ IID.

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$



drawn iid from joint distribution D on $\mathcal{X} \times \mathcal{Y}$

- Goal: find $\hat{f}$ to minimize expected loss over future instances

$$\mathbb{E}_{(x,y)\sim D}[\ell(\hat{f}(x), y)] - \underbrace{\inf_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim D}[\ell(f(x), y)]}_{\text{the best one can do}} \quad \neq f^*$$

- Sample complexity (definition, denote $L(g) = \mathbb{E}[\ell(g, \cdot)]$):
  Given $\delta > 0$, $\epsilon > 0$, and sample complexity $n(\epsilon, \delta)$ is the smallest n such that we can always find a *forecaster $\hat{f}$ s.t. with probability at least $1 - \delta$,*

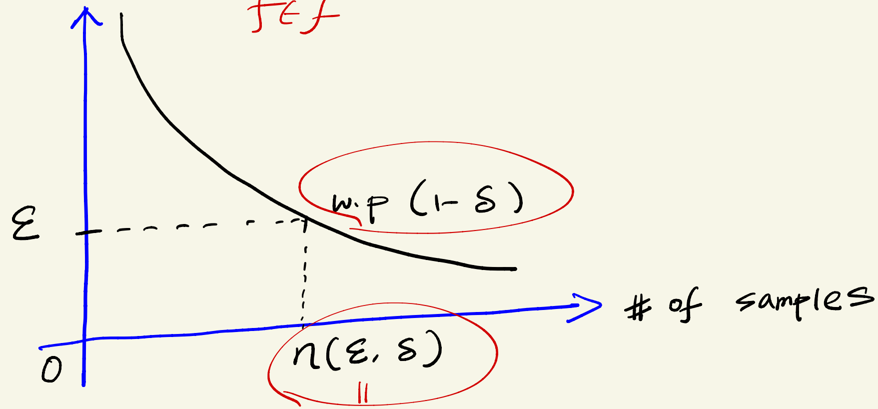
$$L_D(\hat{f}) - \inf_{f\in\mathcal{F}} L_D(f) \leq \epsilon$$

* The learner doesn't assume that $\mathcal{F}$ contains an error free hypothesis f.

Key difference: Agnostic to (f^*)?

((δ, ε)- notion for error analysis)

$$L_D(\hat{f}) - \inf_{f \in \mathcal{F}} L_D(f)$$



Online learning

(sequential input)

$\hat{y}_t$

t

- Online learning

  For $t = 1$ to n $\quad$ $\begin{pmatrix} x_1 & x_2 & \cdots & x_{t-1} & x_t \\ y_1 & y_2 & \cdots & y_{t-1} & ? \end{pmatrix}$ $\quad$ x_t

  Learner receives $x_t \in \mathcal{X}$

  Learner predicts output $\hat{y}_t \in \mathcal{Y}$, $\hat{y}_t = \hat{f}(x_t)$

  True output $y_t \in \mathcal{Y}$ is revealed

  EndFor

- Goal: minimize regret

fixed

$$\mathrm{Reg}_n(\mathcal{F}) := \frac{1}{n}\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \ell(f(x_t), y_t)$$

This course will only introduce the learning theory of offline and supervised learning.

(sequential input)

- Online learning

  For $t = 1$ to n
  Learner receives $x_t \in \mathcal{X}$
  Learner predicts output $\hat{y}_t \in \mathcal{Y}$, $\hat{y}_t = \hat{f}(x_t)$
  True output $y_t \in \mathcal{Y}$ is revealed
  EndFor

- Goal: minimize regret

fixed

$$\mathrm{Reg}_n(\mathcal{F}) := \frac{1}{n}\sum_{t=1}^{n} \ell\left(\hat{y}_t, y_t\right) - \inf_{f \in \mathcal{F}} \frac{1}{n}\sum_{t=1}^{n} \ell\left(f\left(x_t\right), y_t\right) \quad S$$

Q: What is the underlying assumption on $((x_t, y_t): t=1 \cdots n)$?
A: At each time t, (x_t, y_t) is sampled from a (stationary) distribution.

This course will only introduce the learning theory of offline and supervised learning.

what if D_t is changing over time?
(x_t, y_t) (Dynamic regret) adaptive regret

Previously, w.h.p. ... (stochastic models)

- How well does the best learning algorithm do in the worst case scenario? Imagine: There is an adversary attacking my model.

Minimax Rate = "Best Possible Guarantee"

- PAC framework

$\{(x_i, y_i) : i = 1 \cdots n\}$ there exists f^*!

$$\mathcal{V}_n^{PAC}(\mathcal{F}) := \inf_{\hat{f}} \sup_{D_X, f^\star \in \mathcal{F}} \mathbb{E}_{S:|S|=n} \left[\mathbb{P}_{x \sim D_x} \left(\hat{f}(x) \neq f^\star(x) \right) \right] \tag{1}$$

A problem is "PAC learnable" if $\mathcal{V}_n^{PAC} \to 0$ as $n \to \infty$.

- Statistical learning

Q: Can we bound $\mathcal{V}_n^{stat}(\mathcal{F})$? (under some assumptions)

$$\mathcal{V}_n^{stat}(\mathcal{F}) := \inf_{\hat{f}} \sup_{D} \mathbb{E}_{S:|S|=n} \left[L_D(\hat{f}) - \inf_{f \in \mathcal{F}} L_D(f) \right] \tag{2}$$

A problem is "statistically learnable" if $\mathcal{V}_n^{stat} \to 0$ as $n \to \infty$.

Empirical Risk Minimization

- Empirical Risk Minimization (ERM): pick the hypothesis from model class $\mathcal{F}$ that best fits the sample, i.e.,

$$\hat{f}_{\text{erm}} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \underbrace{\frac{1}{n} \sum_{t=1}^{n} \ell\left(f\left(x_t\right), y_t\right)}_{\text{empirical risk}} \triangleq R_{emp}(f) \tag{3}$$

- For a fixed function f, according to the law of large numbers, we have

$$R_{emp}(f) \longrightarrow R_f = \underbrace{\mathbb{E}[\ell(f(x), y)]}_{\text{true risk}} \qquad \text{for } n \longrightarrow \infty$$

randomness w.r.t. data distribution

Empirical Risk Minimization

- Empirical Risk Minimization (ERM): pick the hypothesis from model class $\mathcal{F}$ that best fits the sample, i.e.,

$$\hat{f}_{\mathrm{erm}} \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^{n} \ell\left(f\left(x_t\right), y_t\right) \triangleq R_{emp}(f) \tag{3}$$

- For a fixed function f, according to the law of large numbers, we have

$$R_{emp}(f) \longrightarrow R_f = \underbrace{\mathbb{E}[\ell(f(x), y)]}_{\text{true risk}} \qquad \text{for } n \longrightarrow \infty$$

randomness w.r.t. data distribution

- Bayes optimal function

$$f' := \operatorname*{argmin}_{f} \mathbb{E}[\ell(f(x), y)]$$

$$\Delta := \mathbb{E}[\ell(\hat{f}_{erm}(x), y)] - \mathbb{E}[\ell(f'(x), y)] \quad \text{"excess risk"}$$

In practice, f' is hard to get.

Just some error benchmarks so far.

1 Introduction

PAC learning
Statistical learning
Online learning

2 Empirical Risk Minimization

Error benchmarks
Sample complexity 1
Sample complexity 2
Regret

3 Growth Function and VC dimension

4 Rademacher Complexity

- There are many other interesting frameworks/models!
- You can define your own as long as it's consistent
- We will be focusing on the statistical learning setting.

Empirical Risk Minimization

- Empirical Risk Minimization (ERM): pick the hypothesis from model class $\mathcal{F}$ that best fits the sample, i.e.,

this what we do in practice!

$$\hat{f}_{\text{erm}} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \underbrace{\frac{1}{n} \sum_{t=1}^{n} \ell\left(f\left(x_t\right), y_t\right)}_{\text{empirical risk.}} \triangleq R_{emp}(f) \tag{3}$$

- For a fixed function f, according to the law of large numbers, we have

$$R_{emp}(f) \longrightarrow R_f = \underbrace{\mathbb{E}[\ell(f(x), y)]}_{\text{true risk}} \qquad \text{for } n \longrightarrow \infty$$

Remember this!

- Generalization error bound

f my model

$$\left| \underbrace{\mathbb{E}[\ell(f(x), y)]}_{\text{test error}} - \underbrace{\frac{1}{n} \sum_{t=1}^{n} \ell\left(f\left(x_t\right), y_t\right)}_{\text{training error}} \right| \leq ?$$

"generalizability" (It reflects the ability of "generalizing" from training to testing.)

* Connection with Statistical Learning?

Empirical Risk Minimization

- Hoeffding's inequality
  - Let $X_1, X_2, \ldots, X_n$ be independent random variables.
  - Suppose $S_n = X_1 + X_2 + \cdots + X_n$ and $a_i \le X_i \le b_i \ \forall i$.

$$P\left(|S_n - \underbrace{\mathbb{E}[S_n]}_{\sum_{i=1}^{n} \mathbb{E}[X_i]}| \ge \overset{1^{\circ}}{\epsilon}\right) \le 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

Example: $X_i \sim$ Bernoulli(p)

$b_i = 1$

$a_i = 0.$

$$X_i \begin{cases} p & 1 \\ 1-p & 0 \end{cases}$$

$$Pr\left(\left|\sum x_i - n \cdot p\right| \ge \varepsilon\right) \quad 2 \cdot \exp\left(-\frac{2\varepsilon^2}{n}\right)$$

Empirical Risk Minimization

- Hoeffding's inequality
  - Let $X_1, X_2, \ldots, X_n$ be independent random variables.
  - Suppose $S_n = X_1 + X_2 + \cdots + X_n$ and $a_i \le X_i \le b_i \ \forall i$.

$$P\left(|S_n - \mathbb{E}[S_n]| \ge \epsilon\right) \le 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

- Hoeffding's inequality for ERM

$X_t := \ell(f(x_t), y_t)$.

  - Suppose $\underline{\sup_{y,y' \in \mathcal{Y}} |\ell(y, y')|} \le 1$ assume $X_t \geq 0$. data pt

$$P\left(\left|\underbrace{\mathbb{E}[\ell(f(x), y)]}_{\text{true risk}} - \underbrace{\frac{1}{n}\sum_{t=1}^{n} \ell\left(f\left(x_t\right), y_t\right)}_{\text{empirical risk}}\right| \ge \epsilon\right) \le 2\exp\left(-\frac{\epsilon^2 n}{2}\right) \quad (4)$$

* What's the drawback of this bound?

$$\frac{1}{n} \mathbb{E}\left[\sum_{t=1}^{n} \ell(f(x_t), y_t)\right] = \mathbb{E}[\ell(f(x), y)].$$

since the IID assumption

Empirical Risk Minimization

Cardinality : $\mathcal{F} = \{ f(x) = ax,\ x \in \mathbb{R} \mid a = \{1, 2, 3\} \}$ $|\mathcal{F}| = 3$

- Assume $((x_t, y_t) \sim P : t = 1, \dots, n)$ are generated i.i.d from a distribution P

- ERM with finite class

Proposition 1

Consider the case when the hypothesis $\mathcal{F}$ has finite cardinality, that is $|\mathcal{F}| < \infty$. For any loss ℓ satisfies $\sup_{y, y' \in \mathcal{Y}} |\ell(y, y')| \le 1$, we have

$$\mathcal{V}_n^{stat}(\mathcal{F}) \le \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}[\ell(f(x), y)] - \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right| \right] \le 8 \sqrt{\frac{\log(n|\mathcal{F}|^2)}{n}}$$

The minimax rate is $O\left(\sqrt{\frac{\log |\mathcal{F}|}{n}}\right)$.

$|\mathcal{F}|$:= cardinality of a function class $\mathcal{F}$.

Empirical Risk Minimization

Cardinality: $\mathcal{F} = \{ f(x) = ax, \ x \in \mathbb{R} \mid a = \{1, 2, 3\} \}$ $\quad |\mathcal{F}| = 3$

- Assume $((x_t, y_t) \sim P : t = 1, \dots, n)$ are generated i.i.d from a distribution P

- ERM with finite class

Proposition 1

Consider the case when the hypothesis $\mathcal{F}$ has finite cardinality, that is $|\mathcal{F}| < \infty$. For any loss ℓ satisfies $\sup_{y,y' \in \mathcal{Y}} |\ell(y, y')| \le 1$, we have

$$\mathcal{V}_n^{stat}(\mathcal{F}) \le \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}[\ell(f(x), y)] - \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right| \right] \le 8 \sqrt{\frac{\log(n|\mathcal{F}|^2)}{n}}$$

The minimax rate is $O\left(\sqrt{\frac{\log |\mathcal{F}|}{n}}\right)$. assuming $|\mathcal{F}|$ is large

minimax rate in statistical learning.

- The iid assumption can be replaced by Martingales.
- In practice, iid assumption is very unrealistic $\longrightarrow$ causal learning week 9.

$P(A \cup B) = P(A) + P(B).$

$\inf_{\hat{f} \in \mathcal{F}} \sup_{P} [\quad]$

- ERM with finite class

Proposition 1

Consider the case when the hypothesis $\mathcal{F}$ has finite cardinality, that is $|\mathcal{F}| < \infty$. For any loss ℓ satisfies $\sup_{y,y' \in \mathcal{Y}} |\ell(y, y')| \leq 1$, we have

$$\mathcal{V}_n^{stat}(\mathcal{F}) \leq \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}[\ell(f(x), y)] - \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t) \right| \right] \leq 8 \sqrt{\frac{\log n|\mathcal{F}|^2}{n}}$$

The minimax rate is $O\left(\sqrt{\frac{\log |\mathcal{F}|}{n}}\right)$.

Proof sketch:

$\mathcal{V}_n^{stat}(\mathcal{F}) \leq \sup_{f \in \mathcal{F}} (\text{Generalization Gap}(f)) \leq$

Empirical Risk Minimization

Proof (part I):

$$\mathbb{E}_S\left[L_D(\hat{f}_{erm}) - \inf_{f\in\mathcal{F}} L_D(f)\right]$$

$$=\mathbb{E}_S\left[L_D(\hat{f}_{erm})\right] - \inf_{f\in\mathcal{F}} \mathbb{E}_S\left[\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right), y_t\right)\right]$$

def of $L_D(\cdot)$

$\inf \mathbb{E} \geq \mathbb{E} \inf$

$$\leq\mathbb{E}_S\left[L_D(\hat{f}_{erm}) - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right), y_t\right)\right]$$

Goal: bound V_n^{stat}.

Note that $V_n^{stat}(\mathcal{F}) = \inf_{\hat{f}} \sup_{D} \boxed{\mathbb{E}_S\left[L_D(\hat{f}) - \inf_{f\in\mathcal{F}} L_D(f)\right]}$

Let's focus on this!

[Let's consider using $\hat{f}_{erm}$!]

Proof (part I):

def of $L_D(\cdot)$

$$
\begin{aligned}
&\mathbb{E}_S\left[L_D(\hat{f}_{erm}) - \inf_{f\in\mathcal{F}} L_D(f)\right] \\
=&\mathbb{E}_S\left[L_D(\hat{f}_{erm})\right] - \inf_{f\in\mathcal{F}} \mathbb{E}_S\left[\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right), y_t\right)\right] \\
\text{(why?)}\le&\mathbb{E}_S\left[L_D(\hat{f}_{erm}) - \inf_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right), y_t\right)\right]
\end{aligned}
$$

Empirical Risk Minimization

Proof (part I):

$$\mathbb{E}_S\left[L_D(\hat{f}_{erm}) - \inf_{f\in\mathcal{F}} L_D(f)\right]$$

$$=\mathbb{E}_S\left[L_D(\hat{f}_{erm})\right] - \inf_{f\in\mathcal{F}} \mathbb{E}_S\left[\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right), y_t\right)\right]$$

$$\leq\mathbb{E}_S\left[L_D(\hat{f}_{erm}) - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right), y_t\right)\right]$$

← from last slide

$$\leq\mathbb{E}_S\left[\mathbb{E}\left[\ell(\hat{f}_{erm}(x), y)\right] - \frac{1}{n}\sum_{t=1}^{n}\ell\left(\hat{f}_{erm}\left(x_t\right), y_t\right)\right]$$

def of L_D def of $\hat{f}_{erm}$

(why?)

$$\leq\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left[\mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right), y_t\right)\right]\right]$$

$$\leq\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right), y_t\right)\right|\right]$$

Empirical Risk Minimization

Proof (part I):

$$
\begin{aligned}
&\mathbb{E}_S\left[L_D(\hat{f}_{erm}) - \inf_{f\in\mathcal{F}} L_D(f)\right] \\
=&\mathbb{E}_S\left[L_D(\hat{f}_{erm})\right] - \inf_{f\in\mathcal{F}} \mathbb{E}_S\left[\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right), y_t\right)\right] \\
\leq&\mathbb{E}_S\left[L_D(\hat{f}_{erm}) - \inf_{f\in\mathcal{F}} \frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right), y_t\right)\right] \quad \leftarrow \text{from last slide} \\
\leq&\mathbb{E}_S\left[\underbrace{\mathbb{E}\left[\ell(\hat{f}_{erm}(x), y)\right]}_{\text{def of } L_D} - \frac{1}{n}\sum_{t=1}^{n}\ell\left(\underbrace{\hat{f}_{erm}}_{\text{def of } \hat{f}_{erm}}\left(x_t\right), y_t\right)\right] \\
\leq&\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left[\mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right), y_t\right)\right]\right] \quad \text{since } \hat{f}_{erm}\in\mathcal{F} \\
\leq&\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left|\overbrace{\mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right), y_t\right)}^{Z}\right|\right]
\end{aligned}
$$

Empirical Risk Minimization

Proof (part II):

$$\mathcal{V}_n^{stat}(\mathcal{F}) = \inf_{\hat{f}} \sup_{D} \mathbb{E}_S \left[L_D(\hat{f}) - \inf_{f \in \mathcal{F}} L_D(f) \right]$$

$$\leq \sup_{D} \mathbb{E}_S \left[L_D(\hat{f}_{erm}) - \inf_{f \in \mathcal{F}} L_D(f) \right]$$

from last slide

$$\leq \sup_{D} \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left| \underbrace{\mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n} \sum_{t=1}^{n} \ell\left(f\left(x_t\right), y_t\right)}_{Z} \right| \right]$$

D.

$$\leq \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left| \underbrace{\mathbb{E}\left[\ell(f(x), y)\right]}_{\text{test error}} - \underbrace{\frac{1}{n} \sum_{t=1}^{n} \ell\left(f\left(x_t\right), y_t\right)}_{\text{training error}} \right| \right]$$

Empirical Risk Minimization

Proof (part II):

$$\mathcal{V}_n^{stat}(\mathcal{F}) = \inf_{\hat{f}} \sup_{D} \mathbb{E}_S \left[L_D(\hat{f}) - \inf_{f\in\mathcal{F}} L_D(f) \right]$$

$$\leq \sup_{D} \mathbb{E}_S \left[L_D(\hat{f}_{erm}) - \inf_{f\in\mathcal{F}} L_D(f) \right]$$

from last slide

$$\leq \sup_{D} \mathbb{E}_S \left[\sup_{f\in\mathcal{F}} \left| \underbrace{\mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n}\sum_{t=1}^{n} \ell\left(f\left(x_t\right), y_t\right)}_{Z} \right| \right]$$

$$\leq \mathbb{E}_S \left[\sup_{f\in\mathcal{F}} \left| \underbrace{\mathbb{E}\left[\ell(f(x), y)\right]}_{\text{test error}} - \underbrace{\frac{1}{n}\sum_{t=1}^{n} \ell\left(f\left(x_t\right), y_t\right)}_{\text{training error}} \right| \right]$$

generalization gap!

Jensen's inequality

$\sup \mathbb{E}[(X)] \leq \mathbb{E}[\sup(X)]$

X is a random variable.

- It remains to bound this generalization gap!

Empirical Risk Minimization

Proof (part III): Idea: using the Hoeffding's Inequality! (from last lecture)

$$\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right|\right]$$

$$=\mathbb{E}_S\left[\mathbb{1}_{\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right|\le\epsilon}\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right|\right]$$

$$+\mathbb{E}_S\left[\mathbb{1}_{\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right|>\epsilon}\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right|\right]$$

What are we doing?

Why do we need this?

Empirical Risk Minimization

Proof (part III): Idea: using the Hoeffding's Inequality! (from last lecture)

$$\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right|\right]$$

$$=\mathbb{E}_S\left[\mathbb{1}_{\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right|\le\epsilon}\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right|\right]$$

$$+\mathbb{E}_S\left[\mathbb{1}_{\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right|>\epsilon}\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right|\right]$$

- Truncate the value of the random variable

$$\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{i=1}^{n}\ell(f(x_t),y_t)\right| =: \sup_{f\in\mathcal{F}} Z = Z'$$

into two parts:

① $Z'\le\varepsilon$

② $Z'>\varepsilon$

- $\varepsilon>0$ is a free parameter that's tunable
- This is a standard "truncation technique" in probability theory, which turns a concentration inequality into expectation bounds
- May not be TIGHT! (depending on the upper bounds of Z)

Empirical Risk Minimization

Proof (part III): Let's consider the two parts one-by-one :

$$\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right|\right]$$

$$=\mathbb{E}_S\left[\mathbb{1}_{\underbrace{\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right|\leq\epsilon}_{①}}\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right|\right]$$

$Z' \leq \varepsilon$

$$+\mathbb{E}_S\left[\mathbb{1}_{\underbrace{\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right|>\epsilon}_{②}}\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right|\right]$$

$Z' > \varepsilon$

$$\leq\epsilon+2P\left(\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right|>\epsilon\right)$$

→ Can we bound this probability ?

$$\boxed{\leq\epsilon+2|\mathcal{F}|P\left(\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right|>\epsilon\right)}$$

→ We use the union bound. It's only bounded when

How to bound this ?

$$\leq\epsilon+4|\mathcal{F}|\exp\left(-\frac{\epsilon^2 n}{2}\right)$$

$\exp\left(\frac{-\log n|\mathcal{F}|^2}{2}\right)\cdot 4|\mathcal{F}|.$

- $|\mathcal{F}|$ is finite
  ↓
  function class

Let $\epsilon=\sqrt{\log(n|\mathcal{F}|^2)/n}$, we have $\mathcal{V}_n^{stat}(\mathcal{F})\leq 8\sqrt{\frac{\log(n|\mathcal{F}|^2)}{n}}$. This finished the proof.

Remarks: We have used the following fact (check yourself!)

$$\forall f \in \mathcal{F}, \ \Pr\left(\Big| \underbrace{\mathbb{E}[\ell(f(x), y)] - \frac{1}{n}\sum_{t=1}^{n} \ell(f(x_t), y_t)}_{Z} \Big| > \varepsilon \right)$$

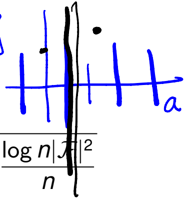
$$\leq 2\exp\left(-\frac{\varepsilon^2 n}{2}\right)$$ The Hoeffding's Inequality

- Standard Form:
  - $\Pr\left(|S_n - \mathbb{E}[S_n]| \geq \varepsilon \right) \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2} \right)$
  - $X_1 \dots X_n$ iid
  - $S_n = \sum_{i=1}^{n} X_i$
  - $a_i \leq X_i \leq b_i, \ \forall \ i = 1 \dots n.$
- Our Form: Define $X_i = \frac{1}{n}\ell(f(x_i), y_i)$

Empirical Risk Minimization

(x, y)
$\in \quad \in$
$\mathbb{R} \quad \{0,+1\}$

$$\mathcal{F} := \{ y = \mathbb{1}(x > a) \mid a \in \mathbb{R} \}$$



$$\mathcal{V}_n^{stat}(\mathcal{F}) \leq \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n} \sum_{t=1}^{n} \ell\left(f\left(x_t\right), y_t\right) \right| \right] \leq 8 \sqrt{\frac{\log n |\mathcal{F}|^2}{n}}$$

- It shows the connection to

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n} \sum_{t=1}^{n} \ell\left(f\left(x_t\right), y_t\right) \right|$$

- It requires that $\mathcal{F}$ is finite, i.e., $|\mathcal{F}| < \infty$
- How about $|\mathcal{F}| = \infty$?

A deeper question:

How to measure the complexity of $\mathcal{F}$?

- When $\mathcal{F}$ is finite: $|\mathcal{F}|$
- When $\mathcal{F}$ is an infinite set: ?

Next slide

1 Introduction

2 Minimax rate

goal:
bound
$\bigcup$ stat
$\cap$

3 Empirical Risk Minimization

4 Growth Function and VC dimension

5 Rademacher Complexity

- Growth function (also known as shattering coefficient)
  Given $\{(x_i, y_i)\}_{1 \le i \le n}$ and define $S = \{x_1, x_2, \ldots, x_n\}$. Let $\mathcal{F}_S = \mathcal{F}_{x_1,\ldots,x_n} = \{f(x_1), \ldots, f(x_n) : f \in \mathcal{F}\}$ and suppose $f(x) \in \{0, 1\}$. The growth function is the maximum number of ways into which n points can be classified by the function class:

$$G(\mathcal{F}, n) = \sup_{x_1,\ldots,x_n} |\mathcal{F}_S|$$

$(x_1\ x_2)$ $(0, 1)$

$(x_1'\ x_2')$ $(1, 0)$

  - When $\mathcal{F}$ is finite, $G(\mathcal{F}, n) \le |\mathcal{F}|$.
  - It always holds that $G(\mathcal{F}, n) \le 2^n$.
  - We say $\mathcal{F}$ shatters S if $|\mathcal{F}_S| = 2^{|S|}$.

- In other words, "shattering is the ability of a model to classify a set of points perfectly".
- Idea: $\mathcal{F}$ is infinite, but can be evaluated by finite samples!

Growth Function

- Growth function (also known as shattering coefficient) Given $\{(x_i, y_i)\}_{1\le i\le n}$ and define $S = \{x_1, x_2, \ldots, x_n\}$. Let $\mathcal{F}_S = \mathcal{F}_{x_1,\ldots,x_n} = \{f(x_1), \ldots, f(x_n) : f \in \mathcal{F}\}$ and suppose $f(x) \in \{0, 1\}$. The growth function is the maximum number of ways into which n points can be classified by the function class:

$$G(\mathcal{F}, n) = \sup_{x_1,\ldots,x_n} |\mathcal{F}_S|$$

  - When $\mathcal{F}$ is finite, $G(\mathcal{F}, n) \le |\mathcal{F}|$.
  - It always holds that $G(\mathcal{F}, n) \le 2^n$.
  - We say $\mathcal{F}$ shatters S if $|\mathcal{F}_S| = 2^{|S|}$.
- Uniform convergence bound

all possible realizations by $\mathcal{F}$

Claim: $$P\left(\sup_{f\in\mathcal{F}}\left|\underbrace{\mathbb{E}[\ell(f(x), y)] - \frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right), y_t\right)}_{Z}\right| \ge \epsilon\right) \le 2G(\mathcal{F}, 2n)\exp\left(-\frac{\epsilon^2 n}{4}\right) \quad (5)$$

* Connection with bound of $\mathcal{V}_n^{stat}$?

Derivation of

$$\Pr\left(\sup_{f\in\mathcal{F}}\left|\underbrace{\mathbb{E}[\ell(f(x),y)]}_{L(f)}-\underbrace{\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)}_{L_{emp}(f)}\right|\geq\varepsilon\right)$$

$$\leq 2G(\mathcal{F},2n)\cdot\exp\left(-\frac{\varepsilon^2 n}{4}\right). \quad \text{— (1)}$$

proof: by Vapnik & Chervonenkis

Lemma (Symmetrization Lemma)

If $n\varepsilon^2\geq 2$, we have

$$\Pr\left(\sup_{f\in\mathcal{F}}|L(f)-L_{emp}(f)|\geq\varepsilon\right)$$

$$\leq 2\Pr\left(\sup_{f\in\mathcal{F}}|L_{emp}(f)-L'_{emp}(f)|>\varepsilon/2\right)$$

an iid dummy copy of the original samples $((x_1,y_1)\cdots(x_n,y_n))$

[Similar to the dummy dataset we created in the Rademacher complexity proof]

Proof sketch of the Symmetrization Lemma

Use the truncation technique!

Therefore, if $n\varepsilon^2 \geq 2$

$$\Pr\left(\sup_{f\in\mathcal{F}} |L(f) - L_{emp}(f)| > \varepsilon\right)$$

$$\leq 2\Pr\left(\sup_{f\in\mathcal{F}} |L_{emp}(f) - L'_{emp}(f)| > \frac{\varepsilon}{2}\right)$$

[Apply the symmetrization Lemma]

$$= 2\Pr\left(\sup_{f\in\mathcal{F}_{Z_{2n}}} |L_{emp}(f) - L'_{emp}(f)| > \frac{\varepsilon}{2}\right)$$

[only functions in $\mathcal{F}_{Z_{2n}}$ are important]

$$\leq 2\,G(\mathcal{F}, 2n)\cdot \exp\left(-\frac{n\varepsilon^2}{4}\right)$$

since we have $\frac{\varepsilon}{2}$

Since $\mathcal{F}_{2n}$ contains at most $G(\mathcal{F}, 2n)$ functions, i.e. there are at most $G(\mathcal{F}, 2n)$ possible realizations!

#

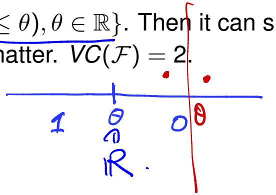
VC dimension

- VC (Vapnik-Chervonenkis) dimension

  The VC dimension of a class $\mathcal{F}$ is the largest n such that $G(\mathcal{F}, n) = 2^n$. In other words, VC dimension of a function class F is the cardinality of the largest set that it can shatters. It is a measure of the capacity (complexity, expressive power, richness, or flexibility) of a set of functions.

- Examples
  - $\mathcal{F} = \{f(x) = I(x \leq \theta), \theta \in \mathbb{R}\}$. Then it can shatter 2 points but for any three points it cannot shatter. $VC(\mathcal{F}) = 2$.

$2^1 = 2.$



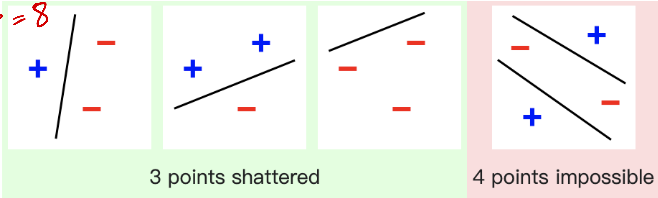
VC dimension

- VC (Vapnik-Chervonenkis) dimension

  The VC dimension of a class $\mathcal{F}$ is the largest n such that $G(\mathcal{F}, n) = 2^n$. In other words, VC dimension of a function class F is the cardinality of the largest set that it can shatters. It is a measure of the capacity (complexity, expressive power, richness, or flexibility) of a set of functions.

- Examples $|\mathcal{F}| = \infty$
  - $\mathcal{F} = \{f(x) = I(x \leq \theta), \theta \in \mathbb{R}\}$. Then it can shatter 2 points but for any three points it cannot shatter. $VC(\mathcal{F}) = 2$.
  - $\mathcal{F}$ is a set of lines in 2-D space: $VC(\mathcal{F}) = 3$. $f(x) = W_0 + W_1 X_1 + W_2 X_2$

$2^3 = 8$



  - Linear function in $\mathbb{R}^d$: $VC(\mathcal{F}) = ?$ $d+1$ (Maybe one question in your assignments)
  - How about rectangles and circles in 2-D space?

- Sauer's lemma

Lemma 1 (Vapnik,Chervonenkis,Sauer,Shelah)

Let $\mathcal{F}$ be a function class with finite VC dimension d. Then

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

$G(\mathcal{F}, d) = 2^d$

for all $n \in \mathbb{N}$. In particular, for all $n \geq d$, we have

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

- This bound is tight!
- This lemma can be used to derive lower bounds on VC dimensions

VC generalization bound

- Recall that

Claim. $P\left(\sup_{f\in\mathcal{F}}\left|\mathbb{E}[\ell(f(x),y)]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right|\geq\epsilon\right)\leq 2G(\mathcal{F},2n)\exp\left(-\frac{\epsilon^2 n}{4}\right)$

$|\mathcal{F}|$

Let the RHS be some $\delta>0$ and then solve it for ϵ. We have

$$\mathbb{E}[\ell(f(x),y)]\leq\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)+\sqrt{\frac{4\left(\left(\log(2G(\mathcal{F},2n)\right)-\log\delta\right))}{n}}$$

VC generalization bound

- Recall that

$$P\left(\sup_{f\in\mathcal{F}}\left|\mathbb{E}[\ell(f(x),y)]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right|\geq\epsilon\right)\leq 2G(\mathcal{F},2n)\exp\left(-\frac{\epsilon^2 n}{4}\right)$$

Let the RHS be some $\delta>0$ and then solve it for ϵ. We have

$$\boxed{\mathbb{E}[\ell(f(x),y)]\leq\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)+\sqrt{\frac{4\left(\left(\log(2G(\mathcal{F},2n)\right)-\log\delta\right))}{n}}}$$

VC Inequality

- Using Lemma 1 (suppose $n\geq d$), we have

$$\boxed{\mathbb{E}[\ell(f(x),y)]\leq\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)+\sqrt{\frac{4\left(d_{VC}\log(\frac{2en}{d_{VC}})-\log\delta\right)}{n}}}$$

The bound is very general (loose) since VC dimension only depends function space but not the dataset.

Can we tighten the bound?

- Rademacher variable σ_i: $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$
- Empirical Rademacher complexity

$$\mathcal{R}(\mathcal{F}) := \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right]$$

$\sigma_1 = 1 \quad 1$

$\sigma_2 = -1 \quad 1$

  - It is a measure of the capacity of a function space and depends on both dataset and $\mathcal{F}$

Rademacher complexity

$f(x_i) \in \{-1, 1\}$

- Rademacher variable σ_i: $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$
- Empirical Rademacher complexity

$$\mathcal{R}(\mathcal{F}) := \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(x_i) \right]$$

$$\mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \ell(f(x_i), y_i) \right]$$

  - It is a measure of the capacity of a function space and depends on both dataset and $\mathcal{F}$

- Uniform convergence bound

This measures the biggest difference of the losses measured over the whole domain and the sample set.

Lemma 2

$$\mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left\{ \underbrace{\mathbb{E}[\ell(f(x), y)] - \frac{1}{n} \sum_{t=1}^{n} \ell(f(x_t), y_t)}_{Z} \right\} \right] \leq 2\mathbb{E}_S \mathcal{R}(\underbrace{\ell \circ \mathcal{F}}_{\text{(abuse notation)}})$$

Rademacher complexity

Proof (part I):

$$
\begin{aligned}
&\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left\{\overbrace{\mathbb{E}[\ell(f(x),y)]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)}^{Z}\right\}\right]\\
=\;&\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left\{\mathbb{E}_{S'}\left[\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t'\right),y_t'\right)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right\}\right]\\
\leq\;&\mathbb{E}_S\left[\mathbb{E}_{S'}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t'\right),y_t'\right)-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right\}\right]\right]\\
=\;&\mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left\{\underbrace{\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t'\right),y_t'\right)}_{(x_t',y_t')_t\text{ are samples from }S'}-\underbrace{\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)}_{(x_t,y_t)_t\text{ are samples from }S}\right\}\right]
\end{aligned}
$$

We have introduced a dummy dataset S'.
What does this inequality mean?

Proof (part II):

$$\mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t'\right),y_t'\right)-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right\}\right]$$

$$=\mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\left(\ell\left(f(x_j'),y_j'\right)-\ell(f(x_j),y_j)+\sum_{i\neq j}\left(\ell\left(f(x_j'),y_j'\right)-\ell(f(x_j),y_j)\right)\right)\right\}\right]$$

$$=\mathop{\mathbb{E}}_{S,S',\sigma_j}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\left(\sigma_j\left(\overbrace{\ell\left(f(x_j'),y_j'\right)}^{①}-\overbrace{\ell(f(x_j),y_j)}^{②}\right)+\sum_{i\neq j}\left(\ell\left(f(x_j'),y_j'\right)-\ell(f(x_j),y_j)\right)\right)\right\}\right]$$

changes of signs will only switch ① and ② .

but ① and ② are iid

Rademacher complexity

Proof (part II):

$$\mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t'\right),y_t'\right)-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right\}\right]$$

$$=\mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\left(\ell\left(f(x_i'),y_i'\right)-\ell(f(x_i),y_i)+\sum_{i\neq j}\left(\ell\left(f(x_j'),y_j'\right)-\ell(f(x_j),y_j)\right)\right)\right\}\right]$$

$$=\mathop{\mathbb{E}}_{S,S',\sigma_j}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\left(\sigma_i\Big(\ell\left(f(x_i'),y_i'\right)-\ell(f(x_i),y_i)\Big)+\sum_{i\neq j}\Big(\ell\left(f(x_j'),y_j'\right)-\ell(f(x_j),y_j)\Big)\right)\right\}\right]$$

$$=\mathop{\mathbb{E}}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{j=1}^{n}\sigma_j\Big(\ell\left(f(x_j'),y_j'\right)-\ell(f(x_j),y_j)\Big)\right\}\right]$$

$\sigma = \{\sigma_1, \sigma_2, \dots, \sigma_n\}$

$$\leq\mathop{\mathbb{E}}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{j=1}^{n}\sigma_j\ell\left(f(x_j'),y_j'\right)\right\}+\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{j=1}^{n}(-\sigma_j)\ell\left(f(x_j),y_j\right)\right\}\right]$$

$$=\mathop{\mathbb{E}}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{j=1}^{n}\sigma_j\ell\left(f(x_j'),y_j'\right)\right\}+\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{j=1}^{n}\sigma_j\ell\left(f(x_j),y_j\right)\right\}\right]$$

Rademacher complexity

Proof (part III):

$$
\begin{aligned}
& \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left\{ \mathbb{E}[\ell(f(x), y)] - \frac{1}{n} \sum_{t=1}^{n} \ell\left(f\left(x_t\right), y_t\right) \right\} \right] \\
\leq & \ \mathbb{E}_{S,S'} \left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{t=1}^{n} \ell\left(f\left(x_t'\right), y_t'\right) - \frac{1}{n} \sum_{t=1}^{n} \ell\left(f\left(x_t\right), y_t\right) \right\} \right] \\
\leq & \ \mathop{\mathbb{E}}_{S,S',\sigma} \left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{j=1}^{n} \sigma_j \ell\left(f(x_j'), y_j'\right) \right\} + \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{j=1}^{n} \sigma_j \ell\left(f(x_j), y_j\right) \right\} \right] \\
= & \ \mathbb{E}_{S'} \mathcal{R}_{S'}(\ell \circ \mathcal{F}) + \mathbb{E}_S \mathcal{R}_S(\ell \circ \mathcal{F}) \\
= & \ 2 \mathbb{E}_S \mathcal{R}_S(\ell \circ \mathcal{F})
\end{aligned}
$$

This finished the proof.

Rademacher complexity bound

Combining $\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left\{\underbrace{\mathbb{E}[\ell(f(x),y)]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)}_{Z(f)}\right\}\right]\leq\ 2\mathbb{E}_S\mathcal{R}_S(\mathcal{F})$ with

Lemma 3 (McDiarmid's Inequality)

Let $x_1,\ldots,x_n$ be independent random variables taking on values in a set A and let $c_1,\ldots,c_n$ be positive real constants. If $\varphi : A^n\to\mathbb{R}$ satisfies

$$\sup_{x_1,\ldots,x_n,x_i'\in A}\left|\varphi\left(x_1,\ldots,x_i,\ldots,x_n\right)-\varphi\left(x_1,\ldots,x_i',\ldots,x_n\right)\right|\leq c_i,$$

for $1\leq i\leq n$, then

$$\mathrm{P}\left(\varphi\left(x_1,\ldots,x_n\right)-\mathbb{E}\left[\varphi\left(x_1,\ldots,x_n\right)\right]\geq\epsilon\right)\leq e^{-2\epsilon^2/\sum_{i=1}^{n}c_i^2}$$

Rademacher complexity bound

Combining $\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left\{\underbrace{\mathbb{E}[\ell(f(x),y)]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)}_{Z(f)}\right\}\right]\le 2\mathbb{E}_S\mathcal{R}_S(\ell\circ\mathcal{F})$ with

Lemma 3 (McDiarmid's Inequality)

Let $x_1,\ldots,x_n$ be independent random variables taking on values in a set A and let $c_1,\ldots,c_n$ be positive real constants. If $\varphi : A^n \to \mathbb{R}$ satisfies

$$\sup_{x_1,\ldots,x_n,x_i'\in A}\left|\varphi\left(x_1,\ldots,x_i,\ldots,x_n\right)-\varphi\left(x_1,\ldots,x_i',\ldots,x_n\right)\right|\le c_i,$$

for $1\le i\le n$, then

$$\mathrm{P}\left(\varphi\left(x_1,\ldots,x_n\right)-\mathbb{E}\left[\varphi\left(x_1,\ldots,x_n\right)\right]\ge\epsilon\right)\le e^{-2\epsilon^2/\sum_{i=1}^{n}c_i^2}$$

?

- $Z(f)$ is a random variable that depends on the dataset S and f.
- Idea: $\varphi(x_1 \cdots x_n) \longleftrightarrow \sup_{f\in\mathcal{F}}\left(\mathbb{E}[\ell(f(x),y)]-\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right)$

Hence. $\sup\left|\varphi(x_1,\cdots x_i,\cdots x_n)-\varphi(x_1\cdots,x_i',\cdots,x_n)\right|\le\frac{1}{n}$

since $\ell(\cdot,\cdot)\in[0,1]$.

$\Longrightarrow \Pr\left(\varphi-\mathbb{E}[\varphi]\ge\varepsilon\right)\le\exp\left(-2\varepsilon^2 n\right)$

So, w.p at least $1-\exp(-2\varepsilon^2 n)$, $\varphi-\mathbb{E}[\varphi]\le\varepsilon$.

Rademacher complexity bound

Combining $\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left\{\mathbb{E}[\ell(f(x),y)]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right\}\right]\leq 2\mathbb{E}_S\mathcal{R}_S(\ell\circ\mathcal{F})$ with

Lemma 3 (McDiarmid's Inequality)

Let $x_1,\ldots,x_n$ be independent random variables taking on values in a set A and let $c_1,\ldots,c_n$ be positive real constants. If $\varphi : A^n\to\mathbb{R}$ satisfies

$$\sup_{x_1,\ldots,x_n,x_i'\in A}\left|\varphi\left(x_1,\ldots,x_i,\ldots,x_n\right)-\varphi\left(x_1,\ldots,x_i',\ldots,x_n\right)\right|\leq c_i,$$

for $1\leq i\leq n$, then

$$\mathrm{P}\left(\varphi\left(x_1,\ldots,x_n\right)-\mathbb{E}\left[\varphi\left(x_1,\ldots,x_n\right)\right]\geq\epsilon\right)\leq e^{-2\epsilon^2/\sum_{i=1}^{n}c_i^2}$$

Assume $0\leq\ell\leq 1$, thus with probability at least $1-\delta$, we have

Using the standard truncation technique,

$$\sup_{f\in\mathcal{F}}\left\{\mathbb{E}[\ell(f(x),y)]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right\}$$

$$\leq\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left\{\mathbb{E}[\ell(f(x),y)]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right\}\right]+\sqrt{\frac{\log(1/\delta)}{2n}}$$

(why?) $$\leq 2\mathbb{E}_S\mathcal{R}_S(\mathcal{F})+\sqrt{\frac{\log(1/\delta)}{2n}}$$

Rademacher complexity bound

We have got

$$\sup_{f\in\mathcal{F}}\left\{\mathbb{E}[\ell(f(x),y)]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right\}$$
$$\leq 2\mathbb{E}_S\mathcal{R}_S(\ell\circ\mathcal{F})+\sqrt{\frac{\log(1/\delta)}{2n}}$$

Apply McDiarmid's inequality again on Rademacher complexity itself. The bounded difference of $\mathcal{R}_S(\ell\circ\mathcal{F}) := \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f\left(x_i\right)$ is still $1/n$. Then with probability of at least $1-\delta$, we have

Rademacher complexity bound

We have got

$$\sup_{f\in\mathcal{F}}\left\{\mathbb{E}[\ell(f(x),y)]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right\}$$

$$\leq 2\mathbb{E}_S\mathcal{R}_S(\ell\circ\mathcal{F})+\sqrt{\frac{\log(1/\delta)}{2n}}$$

Apply McDiarmid's inequality again on Rademacher complexity itself. The bounded difference of $\mathcal{R}_S(\ell\circ\mathcal{F}) := \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i f\left(x_i\right)$ is still $1/n$. Then with probability of at least $1-\delta$, we have

$$\sup_{f\in\mathcal{F}}\left\{\mathbb{E}[\ell(f(x),y)]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right\}$$

$$\leq 2\mathcal{R}_S(\ell\circ\mathcal{F})+3\sqrt{\frac{\log(2/\delta)}{2n}}$$ ← Apply the McDiarmid's inequality twice

*Note that $\mathbb{E}_S\mathcal{R}_S(\ell\circ\mathcal{F})\leq\sqrt{\frac{2\log G(\mathcal{F},n)}{n}}$. (why ?) The Massart's lemma

Rademacher complexity of linear function class

Examples.

Linear function space: $\mathcal{F}_2 = \{x \to \langle w, x \rangle : \|w\|_2 \leq 1\}$

Lemma 4

Let $S = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ *be vectors in a Hilbert space. Suppose* $\|\mathbf{x}_i\| \leq B$, $i = 1, 2, \ldots, n$. *Define:*

$$\mathcal{F}_2 \circ S = \{(\langle \mathbf{w}, \mathbf{x}_1 \rangle, \ldots, \langle \mathbf{w}, \mathbf{x}_n \rangle) : \|\mathbf{w}\|_2 \leq \omega\}.$$

Then $\mathcal{R}(\mathcal{F}_2 \circ S) \leq \frac{\omega B}{\sqrt{n}}$.

Rademacher complexity of linear function class

Proof (part I):

$$\mathcal{R}\left(\mathcal{F}_2 \circ S\right) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{a}\in\mathcal{F}_2\circ S} \frac{1}{n}\sum_{i=1}^{n}\sigma_i a_i\right]$$

$$= \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}:\|\mathbf{w}\|\leq\omega} \sum_{i=1}^{n}\sigma_i\left\langle \mathbf{w},\mathbf{x}_i\right\rangle\right] \qquad a_i := \langle w, x_i \rangle$$

$$= \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}:\|\mathbf{w}\|\leq\omega} \left\langle \mathbf{w}, \sum_{i=1}^{n}\sigma_i\mathbf{x}_i\right\rangle\right] \qquad \text{linearity of inner products}$$

$$\leq \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}:\|\mathbf{w}\|\leq\omega} \|\mathbf{w}\|\left\|\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\right\|\right] \qquad \text{(Cauchy-Schwatz inequality)}$$

why ?

$$\leq \frac{\omega}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\right\|\right] = \frac{\omega}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left(\left\|\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\right\|^2\right)^{1/2}\right]$$

$$\leq \frac{\omega}{n}\left(\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\right\|^2\right]\right)^{1/2} \qquad \text{(Jensen's inequality)}$$

Proof (part II):

$$
\begin{aligned}
\mathcal{R}\left(\mathcal{F}_2 \circ S\right) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{a}\in\mathcal{F}_2\circ S} \frac{1}{n}\sum_{i=1}^{n}\sigma_i a_i\right] \\
&\leq \frac{\omega}{n}\left(\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{n}\sigma_i \mathbf{x}_i\right\|^2\right]\right)^{1/2} \\
&= \frac{\omega}{n}\sqrt{\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{i,j}\sigma_i\sigma_j\left\langle \mathbf{x}_i,\mathbf{x}_j\right\rangle\right]} \\
&= \frac{\omega}{n}\sqrt{\left(\sum_{i\neq j}\left\langle \mathbf{x}_i,\mathbf{x}_j\right\rangle \underbrace{\mathbb{E}_{\boldsymbol{\sigma}}\left[\sigma_i\sigma_j\right]}_{?} + \sum_{i=1}^{n}\left\langle \mathbf{x}_i,\mathbf{x}_i\right\rangle \underbrace{\mathbb{E}_{\boldsymbol{\sigma}}\left[\sigma_i^2\right]}_{=1}\right)} \\
&= \frac{\omega}{n}\sqrt{\sum_{i=1}^{n}\left\|\mathbf{x}_i\right\|^2} \leq \frac{\omega B}{\sqrt{n}}
\end{aligned}
$$

σ_i's are Rademacher variables

$\mathbb{E}[\sigma_i]\,\mathbb{E}[\sigma_j] = 0$, $i \neq j$

$\| \; 0$

This finished the proof.

Generalization bound of linear models

Lemma 5

If the loss function ℓ is η-Lipschitz, we have

$$\mathcal{R}(\ell \circ \mathcal{F}) \leq \eta \mathcal{R}(\mathcal{F})$$

$$|\ell(x') - \ell(x)| \leq \eta \|x - x'\|, \quad \forall x, x' \in \mathcal{X}$$

Generalization bound of linear models

Lemma 5

If the loss function ℓ is η-Lipschitz, we have

$$\mathcal{R}(\ell \circ \mathcal{F}) \leq \eta \mathcal{R}(\mathcal{F})$$

Linear function space: $\mathcal{F}_2 = \{x \to \langle w, x \rangle : \|w\| \leq \omega\}$. Suppose $\|x_i\| \leq B$, $i = 1, 2, \ldots, n$. Then with probability of at least $1 - \delta$, we have

$$\sup_{f \in \mathcal{F}_2} \left\{ \mathbb{E}[\ell(f(x), y)] - \frac{1}{n} \sum_{t=1}^{n} \ell\left(f\left(x_t\right), y_t\right) \right\}$$
$$\leq \frac{2\eta\omega B}{\sqrt{n}} + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

Or equivalently, suppose $f \in \mathcal{F}_2$, then with probability of at least $1 - \delta$,

$$\mathbb{E}[\ell(f(x), y)] \leq \frac{1}{n} \sum_{t=1}^{n} \ell\left(f\left(x_t\right), y_t\right) + \frac{2\eta\omega B}{\sqrt{n}} + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

Learning outcomes

- Understand the concepts of PAC, agnostic PAC, generalization bound, growth function, VC dimension, and Rademacher complexity.
- Understand the properties of the three generalization error bounds we have learned.
- Be able to compute the Rademacher complexities for some simple function classes.
- Be able to derive the generalization bounds for some simple machine learning models.