DDA4210 Advanced Machine Learning
Lecture 03 Learning Theory

Jicong Fan

School of Data Science, CUHK-Shenzhen

February 08

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?")

Basic notation

- 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.: $\mathcal{F} = \left\{x \mapsto f^\top x : \|f\|_2 \leq 1\right\}$, $\mathcal{F} = \left\{x \mapsto \operatorname{sign}\left(f^\top x\right)\right\}$

Probably approximately correct (PAC) learning

- Learner only observes training samples

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

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

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

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

- 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 $1 - \delta$,

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

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

Statistical Learning (agnostic PAC)

- Learner only observes training samples

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

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

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

$$\mathbb{E}_{(x,y)\sim D}[\ell(\hat{f}(x), y)] - \inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim D}[\ell(f(x), y)]$$

- 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 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.

Online learning

- 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

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

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

Minimax Rate

- How well does the best learning algorithm do in the worst case scenario?

Minimax Rate = “Best Possible Guarantee"

- PAC framework

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

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

- Statistical learning

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

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.,

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

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

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

Empirical Risk Minimization

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

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

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

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

- Generalization error bound

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

* Connection with Statistical Learning?

Empirical Risk Minimization

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

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

Empirical Risk Minimization

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

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

- Hoeffding inequality for ERM
  - Suppose $\sup_{y,y' \in \mathcal{Y}} |\ell(y, y')| \leq 1$

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

* What's the drawback of this bound?

Empirical Risk Minimization

- 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

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

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

Empirical Risk Minimization

Proof (part I):

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

Empirical Risk Minimization

Proof (part I):

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

Empirical Risk Minimization

Proof (part II):

$$
\begin{aligned}
\mathcal{V}_n^{stat}(\mathcal{F}) &= \inf_{\hat{f}} \sup_{D} \mathbb{E}_S \left[L_D(\hat{f}) - \inf_{f\in\mathcal{F}} L_D(f) \right] \\
&\leq \sup_{D} \mathbb{E}_S \left[L_D(\hat{f}_{erm}) - \inf_{f\in\mathcal{F}} L_D(f) \right] \\
&\leq \sup_{D} \mathbb{E}_S \left[\sup_{f\in\mathcal{F}} \left| \mathbb{E}\left[\ell(f(x), y)\right] - \frac{1}{n}\sum_{t=1}^{n} \ell\left(f\left(x_t\right), y_t\right) \right| \right] \\
&\leq \mathbb{E}_S \left[\sup_{f\in\mathcal{F}} \left| \underbrace{\mathbb{E}\left[\ell(f(x), y)\right]}_{\text{test error}} - \underbrace{\frac{1}{n}\sum_{t=1}^{n} \ell\left(f\left(x_t\right), y_t\right)}_{\text{training error}} \right| \right]
\end{aligned}
$$

Empirical Risk Minimization

Proof (part III):

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

Empirical Risk Minimization

Proof (part III):

$$
\begin{aligned}
&\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right|\right]\\
&=\mathbb{E}_S\left[\mathbb{1}_{\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right|\le\epsilon}\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right|\right]\\
&\quad+\mathbb{E}_S\left[\mathbb{1}_{\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell(f(x_t),y_t)\right|>\epsilon}\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right|\right]\\
&\le\epsilon+2P\left(\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right|>\epsilon\right)\\
&\le\epsilon+2|\mathcal{F}|P\left(\left|\mathbb{E}\left[\ell(f(x),y)\right]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right|>\epsilon\right)\\
&\le\epsilon+4|\mathcal{F}|\exp\left(-\frac{\epsilon^2 n}{2}\right)
\end{aligned}
$$

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

Empirical Risk Minimization

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

- 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 $|\mathcal{F}| = \infty$?

Growth Function

- Growth function (also known as shattering coefficient) Given $\{(x_i, y_i)\}_{1\leq i\leq n}$ 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, $G(\mathcal{F}, n) \leq |\mathcal{F}|$.
  - It always holds that $G(\mathcal{F}, n) \leq 2^n$.
  - We say $\mathcal{F}$ shatters S if $|\mathcal{F}_S| = 2^{|S|}$.

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

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

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

VC dimension

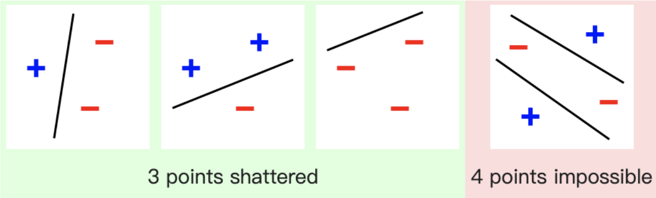
- VC (Vapnik-Chervonenkis) dimension

  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. $VC(\mathcal{F}) = 2$.

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. $VC(\mathcal{F}) = 2$.
  - $\mathcal{F}$ is a set of lines in 2-D space: $VC(\mathcal{F}) = 3$.



  - Linear function in $\mathbb{R}^d$: $VC(\mathcal{F}) =?$
  - 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

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

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

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

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

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

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

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

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

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

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

Rademacher complexity

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

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

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

Rademacher complexity

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

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

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

Lemma 2

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

Rademacher complexity

Proof (part I):

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

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

Rademacher complexity

Proof (part II):

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

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

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

Rademacher complexity

Proof (part II):

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

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

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

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

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

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

Proof (part III):

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

This finished the proof.

Rademacher complexity bound

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

Lemma 3 (McDiarmid 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 $\mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\left\{\mathbb{E}[\ell(f(x),y)]-\frac{1}{n}\sum_{t=1}^{n}\ell\left(f\left(x_t\right),y_t\right)\right\}\right]\leq\ 2\mathbb{E}_S\mathcal{R}_S(\mathcal{F})$ with

Lemma 3 (McDiarmid 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

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

Rademacher complexity bound

We have got

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

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

We have got

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

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

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

*Note that $\mathbb{E}_S\mathcal{R}_S(\mathcal{F})\leq\sqrt{\frac{2\log G(\mathcal{F},n)}{n}}$.

Rademacher complexity of linear function class

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):

$$\begin{aligned}
\mathcal{R}\left(\mathcal{F}_2 \circ S\right) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{a}\in\mathcal{F}_2\circ S}\frac{1}{n}\sum_{i=1}^{n}\sigma_i a_i\right] \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}:\|\mathbf{w}\|\leq\omega}\sum_{i=1}^{n}\sigma_i\left\langle\mathbf{w},\mathbf{x}_i\right\rangle\right] \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}:\|\mathbf{w}\|\leq\omega}\left\langle\mathbf{w},\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\right\rangle\right] \\
&= \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{w}:\|\mathbf{w}\|\leq\omega}\|\mathbf{w}\|\left\|\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\right\|\right] \qquad \text{(Chauchy-Schwatz inequality)} \\
&\leq \frac{\omega}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\right\|\right] = \frac{\omega}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left(\left\|\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\right\|^2\right)^{1/2}\right] \\
&\leq \frac{\omega}{n}\left(\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{n}\sigma_i\mathbf{x}_i\right\|^2\right]\right)^{1/2} \qquad \text{(Jensen's inequality)}
\end{aligned}$$

Rademacher complexity of linear function class

Proof (part II):

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

This finished the proof.

Generalization bound of linear models

Lemma 5

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

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

Generalization bound of linear models

Lemma 5

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

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

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 PCA, 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.