Generalization theory

Daniel Hsu

Columbia TRIPODS Bootcamp

Motivation

Support vector machines

$\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{-1, +1\}$.

- Return solution $\hat{w} \in \mathbb{R}^d$ to following optimization problem:

$$\min_{w \in \mathbb{R}^d} \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}[1 - y_i w^{\mathsf{T}} x_i]_+.$$

- Loss function is *hinge loss*

$$\ell(\hat{y}, y) = [1 - y\hat{y}]_+ = \max\{1 - y\hat{y}, 0\}.$$

  (Here, we are okay with a real-valued prediction.)
- The $\frac{\lambda}{2}\|w\|_2^2$ term is called *Tikhonov regularization*, which we'll discuss later.

Basic statistical model for data

IID model of data

- Training data and test example are *independent and identically distributed* $(\mathcal{X} \times \mathcal{Y})$-valued random variables:

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y) \sim_{\text{iid}} P.$$

Basic statistical model for data

IID model of data

- Training data and test example are *independent and identically distributed* $(\mathcal{X} \times \mathcal{Y})$-valued random variables:

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y) \sim_{\text{iid}} P.$$

SVM in the iid model

- Return solution $\hat{w}$ to following optimization problem:

$$\min_{w \in \mathbb{R}^d} \frac{\lambda}{2} \|w\|_2^2 + \frac{1}{n} \sum_{i=1}^{n} [1 - Y_i w^\mathsf{T} X_i]_+ .$$

- Therefore, $\hat{w}$ is a random variable, depending on $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Convergence of empirical risk

For w that does not depend on training data:

Empirical risk

$$\mathcal{R}_n(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w^\mathsf{T} X_i, Y_i)$$

is a sum of iid random variables.

Convergence of empirical risk

For w that does not depend on training data:

Empirical risk

$$\mathcal{R}_n(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w^\mathsf{T} X_i, Y_i)$$

is a sum of iid random variables.

Law of Large Numbers gives an asymptotic result:

$$\mathcal{R}_n(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(w^\mathsf{T} X_i, Y_i) \xrightarrow{p} \mathbb{E}[\ell(w^\mathsf{T} X, Y)] = \mathcal{R}(w).$$

(This can be made non-asymptotic.)

Uniform convergence of empirical risk

However, $\hat{w}$ does depend on training data.

Empirical risk of $\hat{w}$ is *not* a sum of iid random variables:

$$\mathcal{R}_n(\hat{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{w}^{\mathsf{T}} X_i, Y_i).$$

Uniform convergence of empirical risk

However, $\hat{w}$ does depend on training data.

Empirical risk of $\hat{w}$ is *not* a sum of iid random variables:

$$\mathcal{R}_n(\hat{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(\hat{w}^\intercal X_i, Y_i).$$

Idea: $\hat{w}$ could conceivably take any value w, but if

$$\sup_w |\mathcal{R}_n(w) - \mathcal{R}(w)| \xrightarrow{p} 0, \tag{1}$$

then $\mathcal{R}_n(\hat{w}) \xrightarrow{p} \mathcal{R}(\hat{w})$ as well.

(1) is called *uniform convergence*.

Detour: Concentration inequalities

Symmetric random walk

Rademacher random variables

$\varepsilon_1, \ldots, \varepsilon_n$ iid with $\mathbb{P}(\varepsilon_i = -1) = \mathbb{P}(\varepsilon_i = 1) = 1/2$.

Symmetric random walk: position after n steps is

$$S_n = \sum_{i=1}^{n} \varepsilon_i.$$

Symmetric random walk

Rademacher random variables

$\varepsilon_1, \ldots, \varepsilon_n$ iid with $\mathbb{P}(\varepsilon_i = -1) = \mathbb{P}(\varepsilon_i = 1) = 1/2$.

Symmetric random walk: position after n steps is

$$S_n = \sum_{i=1}^{n} \varepsilon_i.$$

How far from origin?

- By independence, $\mathrm{var}(S_n) = \sum_{i=1}^{n} \mathrm{var}(\varepsilon_i) = n$.
- So expected distance from origin is

$$\mathbb{E}|S_n| \leq \sqrt{\mathrm{var}(S_n)} \leq \sqrt{n}.$$

Symmetric random walk

Rademacher random variables

$\varepsilon_1, \ldots, \varepsilon_n$ iid with $\mathbb{P}(\varepsilon_i = -1) = \mathbb{P}(\varepsilon_i = 1) = 1/2$.

Symmetric random walk: position after n steps is

$$S_n = \sum_{i=1}^{n} \varepsilon_i.$$

How far from origin?

- By independence, $\mathrm{var}(S_n) = \sum_{i=1}^{n} \mathrm{var}(\varepsilon_i) = n$.
- So expected distance from origin is

$$\mathbb{E}|S_n| \leq \sqrt{\mathrm{var}(S_n)} \leq \sqrt{n}.$$

How many realizations are $\gg \sqrt{n}$ from origin?

Markov's inequality

For any random variable X and any $t \geq 0$,

$$\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}|X|}{t}.$$

- Proof:

$$t \cdot \mathbb{1}\{|X| \geq t\} \leq |X|.$$

Markov's inequality

For any random variable X and any $t \geq 0$,

$$\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}|X|}{t}.$$

- Proof:

$$t \cdot \mathbb{1}\{|X| \geq t\} \leq |X|.$$

Application to symmetric random walk:

$$\mathbb{P}(|S_n| \geq c\sqrt{n}) \leq \frac{\mathbb{E}|S_n|}{c\sqrt{n}} \leq \frac{1}{c}.$$

Hoeffding's inequality

If $X_1, \dots, X_n$ are independent random variables, with X_i taking values in $[a_i, b_i]$, then for any $t \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}(X_i)) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Hoeffding's inequality

If $X_1, \ldots, X_n$ are independent random variables, with X_i taking values in $[a_i, b_i]$, then for any $t \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}(X_i)) \geq t\right) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

E.g., Rademacher random variables have $[a_i, b_i] = [-1, +1]$, so

$$\mathbb{P}(S_n \geq t) \leq \exp(-2t^2/(4n)).$$

Applying Hoeffding's inequality to symmetric random walk

Union bound: For any events A and B,

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B).$$

Applying Hoeffding's inequality to symmetric random walk

Union bound: For any events A and B,

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B).$$

---

1. Apply Hoeffding to $\varepsilon_1, \ldots, \varepsilon_n$:

$$\mathbb{P}(S_n \geq c\sqrt{n}) \leq \exp(-c^2/2).$$

Applying Hoeffding's inequality to symmetric random walk

Union bound: For any events A and B,

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B).$$

---

1. Apply Hoeffding to $\varepsilon_1, \ldots, \varepsilon_n$:

$$\mathbb{P}(S_n \geq c\sqrt{n}) \leq \exp(-c^2/2).$$

2. Apply Hoeffding to $-\varepsilon_1, \ldots, -\varepsilon_n$:

$$\mathbb{P}(-S_n \geq c\sqrt{n}) \leq \exp(-c^2/2).$$

Applying Hoeffding's inequality to symmetric random walk

Union bound: For any events A and B,

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B).$$

---

1. Apply Hoeffding to $\varepsilon_1, \ldots, \varepsilon_n$:

$$\mathbb{P}(S_n \geq c\sqrt{n}) \leq \exp(-c^2/2).$$

2. Apply Hoeffding to $-\varepsilon_1, \ldots, -\varepsilon_n$:

$$\mathbb{P}(-S_n \geq c\sqrt{n}) \leq \exp(-c^2/2).$$

3. Therefore, by union bound,

$$\mathbb{P}(|S_n| \geq c\sqrt{n}) \leq 2\exp(-c^2/2).$$

(Compare to bound from Markov's inequality: $1/c$.)

Equivalent form of Hoeffding's inequality

Let $X_1, \ldots, X_n$ be independent random variables, with X_i taking values in $[a_i, b_i]$, and let $S_n = \sum_{i=1}^n X_i$. For any $\delta \in (0, 1)$,

$$\mathbb{P}\left(S_n - \mathbb{E}[S_n] < \sqrt{\frac{1}{2}\sum_{i=1}^n (b_i - a_i)^2 \ln(1/\delta)}\right) \geq 1 - \delta.$$

This is a "high probability" upper-bound on $S_n - \mathbb{E}[S_n]$.

Uniform convergence: Finite classes

Back to statistical learning

Cast of characters:

- feature and outcome spaces: $\mathcal{X}, \mathcal{Y}$
- function class: $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$
- loss function: $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ (assume bounded by 1)
- training and test data: $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y) \sim_{\text{iid}} P$

Back to statistical learning

Cast of characters:

- feature and outcome spaces: $\mathcal{X}, \mathcal{Y}$
- function class: $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$
- loss function: $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ (assume bounded by 1)
- training and test data: $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y) \sim_{\text{iid}} P$

We let $\hat{f} \in \arg\min_{f \in \mathcal{F}} \mathcal{R}_n(f)$ be minimizer of empirical risk

$$\mathcal{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i).$$

Back to statistical learning

Cast of characters:

- feature and outcome spaces: $\mathcal{X}, \mathcal{Y}$
- function class: $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$
- loss function: $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ (assume bounded by 1)
- training and test data: $(X_1, Y_1), \ldots, (X_n, Y_n), (X, Y) \sim_{\text{iid}} P$

We let $\hat{f} \in \arg\min_{f \in \mathcal{F}} \mathcal{R}_n(f)$ be minimizer of empirical risk

$$\mathcal{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i).$$

Our worry: over-fitting $\mathcal{R}(\hat{f}) \gg \mathcal{R}_n(\hat{f})$.

Convergence of empirical risk for fixed function

For any fixed function $f \in \mathcal{F}$,

$$\mathbb{E}\left[\mathcal{R}_n(f)\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(f(X_i), Y_i)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\ell(f(X_i), Y_i)\right] = \mathcal{R}(f).$$

Convergence of empirical risk for fixed function

For any fixed function $f \in \mathcal{F}$,

$$\mathbb{E}\left[\mathcal{R}_n(f)\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(f(X_i), Y_i)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\ell(f(X_i), Y_i)\right] = \mathcal{R}(f).$$

Since $\mathcal{R}_n(f)$ is sum of n independent $[0, \frac{1}{n}]$-valued random variables,

$$\mathbb{P}\left(|\mathcal{R}_n(f) - \mathcal{R}(f)| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(\frac{1}{n})^2}\right) = 2\exp(-2nt^2)$$

for any $t > 0$, by Hoeffding's inequality and union bound.

Convergence of empirical risk for fixed function

For any fixed function $f \in \mathcal{F}$,

$$\mathbb{E}\left[\mathcal{R}_n(f)\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\ell(f(X_i), Y_i)\right] = \mathcal{R}(f).$$

Since $\mathcal{R}_n(f)$ is sum of n independent $[0, \frac{1}{n}]$-valued random variables,

$$\mathbb{P}\left(|\mathcal{R}_n(f) - \mathcal{R}(f)| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(\frac{1}{n})^2}\right) = 2\exp(-2nt^2)$$

for any $t > 0$, by Hoeffding's inequality and union bound.

This argument does not apply to $\hat{f}$, because $\hat{f}$ depends on $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Uniform convergence

We cannot directly apply Hoeffding's inequality to $\hat{f}$, since its empirical risk $\mathcal{R}_n(\hat{f})$ is not average of iid random variables.

Uniform convergence

We cannot directly apply Hoeffding's inequality to $\hat{f}$, since its empirical risk $\mathcal{R}_n(\hat{f})$ is not average of iid random variables.

One possible solution: ensure empirical risk of every $f \in \mathcal{F}$ is close to its expected value.

This is called *uniform convergence*.

Uniform convergence

We cannot directly apply Hoeffding's inequality to $\hat{f}$, since its empirical risk $\mathcal{R}_n(\hat{f})$ is not average of iid random variables.

One possible solution: ensure empirical risk of every $f \in \mathcal{F}$ is close to its expected value.

This is called *uniform convergence*.

- How much data is needed to ensure this?

Uniform convergence for all functions in a finite class

If $|\mathcal{F}| < \infty$, then by Hoeffding's inequality and union bound,

$$\begin{aligned}
\mathbb{P}\left(\exists f \in \mathcal{F} \text{ s.t. } |\mathcal{R}_n(f) - \mathcal{R}(f)| \geq t\right) &= \mathbb{P}\left(\bigcup_{f \in \mathcal{F}} \{|\mathcal{R}_n(f) - \mathcal{R}(f)| \geq t\}\right) \\
&\leq \sum_{f \in \mathcal{F}} \mathbb{P}\left(|\mathcal{R}_n(f) - \mathcal{R}(f)| \geq t\right) \\
&\leq |\mathcal{F}| \cdot 2\exp(-2nt^2).
\end{aligned}$$

Uniform convergence for all functions in a finite class

If $|\mathcal{F}| < \infty$, then by Hoeffding's inequality and union bound,

$$\begin{aligned}
\mathbb{P}\left(\exists f \in \mathcal{F} \text{ s.t. } |\mathcal{R}_n(f) - \mathcal{R}(f)| \geq t\right) &= \mathbb{P}\left(\bigcup_{f \in \mathcal{F}} \{|\mathcal{R}_n(f) - \mathcal{R}(f)| \geq t\}\right) \\
&\leq \sum_{f \in \mathcal{F}} \mathbb{P}\left(|\mathcal{R}_n(f) - \mathcal{R}(f)| \geq t\right) \\
&\leq |\mathcal{F}| \cdot 2\exp(-2nt^2).
\end{aligned}$$

Choose t so that RHS is δ, and "invert".

Theorem. For any $\delta \in (0, 1)$,

$$\mathbb{P}\left(\forall f \in \mathcal{F} \;:\; |\mathcal{R}_n(f) - \mathcal{R}(f)| < \sqrt{\frac{\ln(2|\mathcal{F}|/\delta)}{2n}}\right) \geq 1 - \delta.$$

What we get from uniform convergence

If $n \gg \log|\mathcal{F}|$, then with high probability, no function $f \in \mathcal{F}$ will over-fit the training data.

What we get from uniform convergence

If $n \gg \log |\mathcal{F}|$, then with high probability, no function $f \in \mathcal{F}$ will over-fit the training data.

Also: An *empirical risk minimizer* (*ERM*), like $\hat{f}$, is near optimal!

Theorem. With probability at least $1 - \delta$,

$$\begin{aligned}
\mathcal{R}(\hat{f}) - \mathcal{R}(f^*) &= \mathcal{R}(\hat{f}) - \mathcal{R}_n(\hat{f}) && (\leq \epsilon) \\
&\quad + \mathcal{R}_n(\hat{f}) - \mathcal{R}_n(f^*) && (\leq 0) \\
&\quad + \mathcal{R}_n(f^*) - \mathcal{R}(f^*) && (\leq \epsilon) \\
&\leq 2\epsilon
\end{aligned}$$

where $f^* \in \arg\min_{f \in \mathcal{F}} \mathcal{R}(f)$ and $\epsilon = \sqrt{\frac{\ln(2|\mathcal{F}|/\delta)}{2n}}$.

Uniform convergence: General case

Uniform convergence: General case

Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ be a class of real-valued functions, and let P be a probability distribution on $\mathcal{X}$.

Uniform convergence: General case

Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ be a class of real-valued functions, and let P be a probability distribution on $\mathcal{X}$.

Notation:

- Let $Pf = \mathbb{E}[f(X)]$ for $X \sim P$.
- Let P_n be the *empirical distribution* on $X_1, \ldots, X_n \sim_{\text{iid}} P$, which assigns probability mass $1/n$ to each X_i.
- So $P_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$.

We are interested in the maximum (or supremum) deviation:

$$\sup_{f \in \mathcal{F}} |P_n f - Pf|.$$

Uniform convergence: General case

Let $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ be a class of real-valued functions, and let P be a probability distribution on $\mathcal{X}$.

Notation:

- Let $Pf = \mathbb{E}[f(X)]$ for $X \sim P$.
- Let P_n be the *empirical distribution* on $X_1, \ldots, X_n \sim_{\text{iid}} P$, which assigns probability mass $1/n$ to each X_i.
- So $P_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$.

We are interested in the maximum (or supremum) deviation:

$$\sup_{f \in \mathcal{F}} |P_n f - Pf|.$$

The arguments from before show that for any finite class of bounded functions $\mathcal{F}$,

$$\sup_{f \in \mathcal{F}} |P_n f - Pf| \xrightarrow{p} 0,$$

and also give a non-asymptotic rate of convergence.

Infinite classes

For which classes $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ does uniform convergence hold?

Infinite classes

For which classes $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ does uniform convergence hold?

Example:

$$\mathcal{F} = \{f_S(x) = \mathbb{1}\{x \in S\} : S \subset \mathbb{R}, |S| < \infty\},$$

i.e., $\{0, 1\}$-valued functions that take value 1 on a finite set.

- If P is continuous, then $Pf = 0$ for all $f \in \mathcal{F}$.
- But $\sup_{f \in \mathcal{F}} P_n f = 1$ for all n.
- So $\sup_{f \in \mathcal{F}} |P_n f - Pf| = 1$ for all n.

Infinite classes

For which classes $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ does uniform convergence hold?

Example:

$$\mathcal{F} = \{f_S(x) = \mathbb{1}\{x \in S\} : S \subset \mathbb{R}, |S| < \infty\},$$

i.e., $\{0, 1\}$-valued functions that take value 1 on a finite set.

- If P is continuous, then $Pf = 0$ for all $f \in \mathcal{F}$.
- But $\sup_{f \in \mathcal{F}} P_n f = 1$ for all n.
- So $\sup_{f \in \mathcal{F}} |P_n f - Pf| = 1$ for all n.

What is the appropriate “complexity” measure of a function class?

Rademacher complexity

Let $\varepsilon_1, \dots, \varepsilon_n$ be independent Rademacher random variables.

Uniform convergence with $\mathcal{F}$ holds iff

$$\lim_{n\to\infty} \underbrace{\mathbb{E}\mathbb{E}_\varepsilon \left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right]}_{\mathrm{Rad}_n(\mathcal{F})} = 0$$

(where $\mathbb{E}_\varepsilon$ is expectation with respect to $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$).

Rademacher complexity

Let $\varepsilon_1, \ldots, \varepsilon_n$ be independent Rademacher random variables.

Uniform convergence with $\mathcal{F}$ holds iff

$$\lim_{n\to\infty} \underbrace{\mathbb{E}\mathbb{E}_\varepsilon \left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right]}_{\mathrm{Rad}_n(\mathcal{F})} = 0$$

(where $\mathbb{E}_\varepsilon$ is expectation with respect to $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$).

$\mathrm{Rad}_n(\mathcal{F})$ is the *Rademacher complexity* of $\mathcal{F}$, which measures how well vectors in (random) set

$$\mathcal{F}(X_{1:n}) = \{(f(X_1), \ldots, f(X_n)) : f \in \mathcal{F}\}$$

can correlate with uniformly random signs $\varepsilon_1, \ldots, \varepsilon_n$.

Extreme cases of Rademacher complexity

For simplicity, assume $X_1, \ldots, X_n$ are distinct (e.g., P continuous).

Extreme cases of Rademacher complexity

For simplicity, assume $X_1, \ldots, X_n$ are distinct (e.g., P continuous).

- $\mathcal{F}$ contains *a single function* $f_0 \colon \mathcal{X} \to \{-1, +1\}$:

$$\mathrm{Rad}_n(\mathcal{F}) = \mathbb{E}\mathbb{E}_\varepsilon \left[\left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_0(X_i) \right| \right] \leq \frac{1}{\sqrt{n}}.$$

Extreme cases of Rademacher complexity

For simplicity, assume $X_1, \dots, X_n$ are distinct (e.g., P continuous).

- $\mathcal{F}$ contains *a single function* $f_0 \colon \mathcal{X} \to \{-1, +1\}$:

$$\mathrm{Rad}_n(\mathcal{F}) = \mathbb{E}\mathbb{E}_\varepsilon \left[\left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_0(X_i) \right| \right] \leq \frac{1}{\sqrt{n}}.$$

- $\mathcal{F}$ contains *all functions* $\mathcal{X} \to \{-1, +1\}$:

$$\mathrm{Rad}_n(\mathcal{F}) = \mathbb{E}\mathbb{E}_\varepsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right| \right] = 1.$$

Uniform convergence via Rademacher complexity

Theorem.

1. *Uniform convergence in expectation*:
   For any $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$,
   $$\mathbb{E}\left[\sup_{f \in \mathcal{F}} |P_n f - P f|\right] \leq 2 \operatorname{Rad}_n(\mathcal{F}).$$
2. *Uniform convergence with high probability*:
   For any $\mathcal{F} \subset [-1, +1]^{\mathcal{X}}$ and $\delta \in (0, 1)$, with probability $\geq 1 - \delta$,
   $$\sup_{f \in \mathcal{F}} |P_n f - P f| \leq 2 \operatorname{Rad}_n(\mathcal{F}) + \sqrt{\frac{2 \ln(1/\delta)}{n}}.$$

Step 1: Symmetrization by "ghost sample"

Let P_n' be empirical distribution on independent copies $X_1', \ldots, X_n'$ of $X_1, \ldots, X_n$. Write $\mathbb{E}'$ for expectation with respect to $X_{1:n}'$.

Step 1: Symmetrization by "ghost sample"

Let P'_n be empirical distribution on independent copies $X'_1, \dots, X'_n$ of $X_1, \dots, X_n$. Write $\mathbb{E}'$ for expectation with respect to $X'_{1:n}$.

Then

$$
\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}} |P_n f - Pf|\right] &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \mathbb{E}'\left\{ \frac{1}{n}\sum_{i=1}^n f(X_i) - f(X'_i) \right\} \right|\right] \\
&\le \mathbb{E}\left[\mathbb{E}'\left\{ \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - f(X'_i) \right| \right\}\right] \\
&= \mathbb{E}\mathbb{E}' \sup_{f\in\mathcal{F}} |P_n f - P'_n f|.
\end{aligned}
$$

Step 1: Symmetrization by "ghost sample"

Let P_n' be empirical distribution on independent copies $X_1', \ldots, X_n'$ of $X_1, \ldots, X_n$. Write $\mathbb{E}'$ for expectation with respect to $X_{1:n}'$.

Then

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}} |P_n f - Pf|\right] &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}'\left\{\frac{1}{n}\sum_{i=1}^n f(X_i) - f(X_i')\right\}\right|\right] \\
&\le \mathbb{E}\left[\mathbb{E}'\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - f(X_i')\right|\right\}\right] \\
&= \mathbb{E}\mathbb{E}' \sup_{f\in\mathcal{F}} |P_n f - P_n' f|.
\end{aligned}$$

The random variable $P_n f - P_n' f$ is arguably nicer than $P_n f - Pf$ because it is symmetric.

Step 2: Symmetrization by random signs

Consider any $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \{-1, +1\}^n$. Distribution of

$$P_n f - P'_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i) - f(X'_i)$$

is the same distribution of

$$P_n f - P'_n f = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(f(X_i) - f(X'_i) \right).$$

Step 2: Symmetrization by random signs

Consider any $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \{-1, +1\}^n$. Distribution of

$$P_n f - P'_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i) - f(X'_i)$$

is the same distribution of

$$P_n f - P'_n f = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(f(X_i) - f(X'_i) \right).$$

Thus, this is also true for uniform average over all $\varepsilon \in \{-1, +1\}^n$ (i.e., expectation over Rademacher ε):

$$\mathbb{E}\mathbb{E}' \sup_{f \in \mathcal{F}} |P_n f - P'_n f| = \mathbb{E}\mathbb{E}'\mathbb{E}_\varepsilon \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \left(f(X_i) - f(X'_i) \right) \right|.$$

Step 3: Back to a single sample

By triangle inequality,

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left(f(X_i)-f(X_i')\right)\right|$$
$$\leq \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right| + \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i')\right|$$

The two terms on the RHS have the same distribution.

Step 3: Back to a single sample

By triangle inequality,

$$\begin{aligned}
&\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left(f(X_i)-f(X_i')\right)\right| \\
&\le \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right| + \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i')\right|
\end{aligned}$$

The two terms on the RHS have the same distribution.

So

$$\begin{aligned}
\mathbb{E}\mathbb{E}'\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left(f(X_i)-f(X_i')\right)\right| &\le 2\mathbb{E}\mathbb{E}_\varepsilon \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right| \\
&= 2\,\mathrm{Rad}_n(\mathcal{F}).
\end{aligned}$$

Recap

For any $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$,

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} |P_n f - Pf|\right] \leq 2\,\mathrm{Rad}_n(\mathcal{F}).$$

For any $\mathcal{F} \subset [-1, +1]^{\mathcal{X}}$ and $\delta \in (0, 1)$, with probability $\geq 1 - \delta$,

$$\sup_{f \in \mathcal{F}} |P_n f - Pf| \leq 2\,\mathrm{Rad}_n(\mathcal{F}) + \sqrt{\frac{2\ln(1/\delta)}{n}}.$$

Recap

For any $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$,

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} |P_n f - Pf|\right] \leq 2\,\mathrm{Rad}_n(\mathcal{F}).$$

For any $\mathcal{F} \subset [-1, +1]^{\mathcal{X}}$ and $\delta \in (0, 1)$, with probability $\geq 1 - \delta$,

$$\sup_{f \in \mathcal{F}} |P_n f - Pf| \leq 2\,\mathrm{Rad}_n(\mathcal{F}) + \sqrt{\frac{2\ln(1/\delta)}{n}}.$$

Conclusion

If $\mathrm{Rad}_n(\mathcal{F}) \to 0$, then uniform convergence holds.

Recap

For any $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$,

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} |P_n f - P f|\right] \leq 2\,\mathrm{Rad}_n(\mathcal{F}).$$

For any $\mathcal{F} \subset [-1, +1]^{\mathcal{X}}$ and $\delta \in (0, 1)$, with probability $\geq 1 - \delta$,

$$\sup_{f \in \mathcal{F}} |P_n f - P f| \leq 2\,\mathrm{Rad}_n(\mathcal{F}) + \sqrt{\frac{2\ln(1/\delta)}{n}}.$$

Conclusion

If $\mathrm{Rad}_n(\mathcal{F}) \to 0$, then uniform convergence holds.

(Can also show: If uniform convergence holds, then $\mathrm{Rad}_n(\mathcal{F}) \to 0$.)

Analysis of SVM

Loss class

Back to classes of prediction functions $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$.

Loss class

Back to classes of prediction functions $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$.

Consider a loss function $\ell \colon \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ that satisfies $\ell(0, y) \leq 1$ for all $y \in \mathcal{Y}$, and is 1-Lipschitz in first argument: for all $\hat{y}, \hat{y}' \in \mathbb{R}$,

$$|\ell(\hat{y}, y) - \ell(\hat{y}', y)| \leq |\hat{y} - \hat{y}'|.$$

(Example: hinge loss $\ell(\hat{y}, y) = [1 - \hat{y}y]_+$.)

Loss class

Back to classes of prediction functions $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$.

Consider a loss function $\ell \colon \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ that satisfies $\ell(0, y) \leq 1$ for all $y \in \mathcal{Y}$, and is 1-Lipschitz in first argument: for all $\hat{y}, \hat{y}' \in \mathbb{R}$,

$$|\ell(\hat{y}, y) - \ell(\hat{y}', y)| \leq |\hat{y} - \hat{y}'|.$$

(Example: hinge loss $\ell(\hat{y}, y) = [1 - \hat{y}y]_+$.)

Define the associated *loss class* by

$$\ell_{\mathcal{F}} = \{(x, y) \mapsto \ell(f(x), y) : f \in \mathcal{F}\}.$$

Then

$$\mathrm{Rad}_n(\ell_{\mathcal{F}}) \leq 2\,\mathrm{Rad}_n(\mathcal{F}) + \sqrt{\frac{2 \ln 2}{n}}.$$

So uniform convergence holds for $\ell_{\mathcal{F}}$ if it holds for $\mathcal{F}$.

Rademacher complexity of linear predictors

Linear functions $\mathcal{F}_{\text{lin}} = \{w \in \mathbb{R}^d\}$.

What is the Rademacher complexity of $\mathcal{F}_{\text{lin}}$?

$$\text{Rad}_n(\mathcal{F}_{\text{lin}}) = \mathbb{E}\mathbb{E}_\varepsilon \left[\sup_{w \in \mathbb{R}^d} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i w^\top X_i \right| \right].$$

Rademacher complexity of linear predictors

Linear functions $\mathcal{F}_{\text{lin}} = \{w \in \mathbb{R}^d\}$.

What is the Rademacher complexity of $\mathcal{F}_{\text{lin}}$?

$$\text{Rad}_n(\mathcal{F}_{\text{lin}}) = \mathbb{E}\mathbb{E}_\varepsilon \left[\sup_{w \in \mathbb{R}^d} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i w^\mathsf{T} X_i \right| \right].$$

Inside the $\mathbb{E}\mathbb{E}_\varepsilon$:

$$\sup_{w \in \mathbb{R}^d} \left| w^\mathsf{T} \left(\frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \right) \right| = \sup_{w \in \mathbb{R}^d} \|w\|_2 \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \right\|_2.$$

As long as $\sum_{i=1}^n \varepsilon_i X_i \neq 0$, this is unbounded! **:-(**

Regularization

Recall SVM optimization problem:

$$\min_{w \in \mathbb{R}^d} \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}[1 - y_i w^\mathsf{T} x_i]_+.$$

Regularization

Recall SVM optimization problem:

$$\min_{w \in \mathbb{R}^d} \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}[1 - y_i w^\mathsf{T} x_i]_+.$$

Objective value at $w = 0$ is 1, so objective value at minimizer $\hat{w}$ is no worse than this:

$$\frac{\lambda}{2}\|\hat{w}\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}[1 - y_i \hat{w}^\mathsf{T} x_i]_+ \leq 1.$$

Regularization

Recall SVM optimization problem:

$$\min_{w \in \mathbb{R}^d} \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}[1 - y_i w^\mathsf{T} x_i]_+.$$

Objective value at $w = 0$ is 1, so objective value at minimizer $\hat{w}$ is no worse than this:

$$\frac{\lambda}{2}\|\hat{w}\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}[1 - y_i \hat{w}^\mathsf{T} x_i]_+ \leq 1.$$

Therefore

$$\|\hat{w}\|_2^2 \leq \frac{2}{\lambda}.$$

Rademacher complexity of bounded linear predictors

Bounded linear functions $\mathcal{F}_{\ell_2,B} = \{w \in \mathbb{R}^d : \|w\|_2 \leq B\}$.

Rademacher complexity of bounded linear predictors

Bounded linear functions $\mathcal{F}_{\ell_2,B} = \{w \in \mathbb{R}^d : \|w\|_2 \le B\}$.

What is the Rademacher complexity of $\mathcal{F}_{\ell_2,B}$?

$$\begin{aligned}
\mathrm{Rad}_n(\mathcal{F}_{\ell_2,B}) &= \mathbb{E}\mathbb{E}_\varepsilon \left[\sup_{\|w\|_2 \le B} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i w^\intercal X_i \right| \right] \\
&= B\mathbb{E}\mathbb{E}_\varepsilon \left[\sup_{\|u\|_2 \le 1} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i u^\intercal X_i \right| \right] \\
&= B\mathbb{E}\mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \right\|_2 \le B \sqrt{\mathbb{E}\mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \right\|_2^2}.
\end{aligned}$$

Rademacher complexity of bounded linear predictors

Bounded linear functions $\mathcal{F}_{\ell_2,B} = \{w \in \mathbb{R}^d : \|w\|_2 \leq B\}$.

What is the Rademacher complexity of $\mathcal{F}_{\ell_2,B}$?

$$
\begin{aligned}
\mathrm{Rad}_n(\mathcal{F}_{\ell_2,B}) &= \mathbb{E}\mathbb{E}_\varepsilon \left[\sup_{\|w\|_2 \leq B} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i w^\mathsf{T} X_i \right| \right] \\
&= B \mathbb{E}\mathbb{E}_\varepsilon \left[\sup_{\|u\|_2 \leq 1} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i u^\mathsf{T} X_i \right| \right] \\
&= B \mathbb{E}\mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \right\|_2 \leq B \sqrt{\mathbb{E}\mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \right\|_2^2}.
\end{aligned}
$$

This is d-dimensional random walk, where i-th step is $\pm X_i$.

Rademacher complexity of bounded linear predictors (2)

$$\begin{aligned}
\mathbb{E}\mathbb{E}_{\varepsilon}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i X_i\right\|_2^2 &= \frac{1}{n^2}\mathbb{E}\mathbb{E}_{\varepsilon}\left[\sum_{i=1}^{n}\|\varepsilon_i X_i\|_2^2 + \sum_{i\neq j}\varepsilon_i\varepsilon_j X_i^{\mathsf{T}} X_j\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}\|X_i\|_2^2\right] \\
&= \frac{1}{n}\mathbb{E}\|X\|_2^2.
\end{aligned}$$

Rademacher complexity of bounded linear predictors (2)

$$\begin{aligned}
\mathbb{E}\mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \right\|_2^2 &= \frac{1}{n^2} \mathbb{E}\mathbb{E}_\varepsilon \left[\sum_{i=1}^n \|\varepsilon_i X_i\|_2^2 + \sum_{i \neq j} \varepsilon_i \varepsilon_j X_i^\intercal X_j \right] \\
&= \frac{1}{n^2} \mathbb{E} \left[\sum_{i=1}^n \|X_i\|_2^2 \right] \\
&= \frac{1}{n} \mathbb{E}\|X\|_2^2.
\end{aligned}$$

Conclusion

Rademacher complexity of $\mathcal{F}_{\ell_2,B} = \{w \in \mathbb{R}^d : \|w\|_2 \leq B\}$:

$$\mathrm{Rad}_n(\mathcal{F}_{\ell_2,B}) \leq B \sqrt{\frac{\mathbb{E}\|X\|_2^2}{n}}.$$

Risk bound for SVM

$$
\begin{aligned}
&\mathbb{E}\left[\mathcal{R}(\hat{w}) - \mathcal{R}(w^*)\right] \\
&= \mathbb{E}\left[\mathcal{R}(\hat{w}) - \mathcal{R}_n(\hat{w})\right] && (\leq \epsilon) \\
&\quad + \mathbb{E}\left[\tfrac{\lambda}{2}\|\hat{w}\|_2^2 + \mathcal{R}_n(\hat{w}) - \tfrac{\lambda}{2}\|w^*\|_2^2 - \mathcal{R}_n(w^*)\right] && (\leq 0) \\
&\quad + \mathbb{E}\left[\mathcal{R}_n(w^*) - \mathcal{R}(w^*)\right] && (= 0) \\
&\quad + \mathbb{E}\left[\tfrac{\lambda}{2}\|w^*\|_2^2 - \tfrac{\lambda}{2}\|\hat{w}\|_2^2\right] \\
&\leq \epsilon + \tfrac{\lambda}{2}\|w^*\|_2^2
\end{aligned}
$$

where

$$
w^* \in \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \tfrac{\lambda}{2}\|w\|_2^2 + \mathcal{R}(w), \qquad \epsilon = O\left(\sqrt{\frac{\mathbb{E}\|X\|_2^2}{\lambda n}} + \frac{1}{\sqrt{n}}\right).
$$

Risk bound for SVM

$$
\begin{aligned}
&\mathbb{E}\left[\mathcal{R}(\hat{w}) - \mathcal{R}(w^*)\right] \\
&= \mathbb{E}\left[\mathcal{R}(\hat{w}) - \mathcal{R}_n(\hat{w})\right] && (\leq \epsilon) \\
&\quad + \mathbb{E}\left[\tfrac{\lambda}{2}\|\hat{w}\|_2^2 + \mathcal{R}_n(\hat{w}) - \tfrac{\lambda}{2}\|w^*\|_2^2 - \mathcal{R}_n(w^*)\right] && (\leq 0) \\
&\quad + \mathbb{E}\left[\mathcal{R}_n(w^*) - \mathcal{R}(w^*)\right] && (= 0) \\
&\quad + \mathbb{E}\left[\tfrac{\lambda}{2}\|w^*\|_2^2 - \tfrac{\lambda}{2}\|\hat{w}\|_2^2\right] \\
&\leq \epsilon + \tfrac{\lambda}{2}\|w^*\|_2^2
\end{aligned}
$$

where

$$
w^* \in \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \tfrac{\lambda}{2}\|w\|_2^2 + \mathcal{R}(w), \quad \epsilon = O\left(\sqrt{\frac{\mathbb{E}\|X\|_2^2}{\lambda n}} + \frac{1}{\sqrt{n}}\right).
$$

This suggests we should use $\lambda \to 0$ such that $\lambda n \to \infty$ as $n \to \infty$.

Kernels

Excess risk bound has no *explicit* dependence on the dimension d. In particular, it holds in infinite dimensional inner product spaces.

- SVM can be applied in such spaces as long as there is an algorithm for computing inner products.
- This is the *kernel trick*, and these corresponding spaces are called *Reproducing Kernel Hilbert Spaces* (*RKHS*).

Kernels

Excess risk bound has no *explicit* dependence on the dimension d. In particular, it holds in infinite dimensional inner product spaces.

- SVM can be applied in such spaces as long as there is an algorithm for computing inner products.
- This is the *kernel trick*, and these corresponding spaces are called *Reproducing Kernel Hilbert Spaces* (*RKHS*).

Universal approximation

With some RKHS, can approximate any function arbitrarily well:

$$\lim_{\lambda \to 0} \left\{ \inf_{w \in \mathcal{F}} \frac{\lambda}{2} \|w\|^2 + \mathcal{R}(w) \right\} = \inf_{g \colon \mathcal{X} \to \mathbb{R}} \mathcal{R}(g).$$

Other regularizers

Instead of SVM, suppose $\hat{w}$ is solution to

$$\min_{w \in \mathbb{R}^d} \lambda \|w\|_1 + \mathcal{R}_n(w).$$

So $\hat{w} \in \mathcal{F}_{\ell_1,B} = \{w \in \mathbb{R}^d : \|w\|_1 \leq B\}$ for $B = 1/\lambda$.

Other regularizers

Instead of SVM, suppose $\hat{w}$ is solution to

$$\min_{w \in \mathbb{R}^d} \lambda \|w\|_1 + \mathcal{R}_n(w).$$

So $\hat{w} \in \mathcal{F}_{\ell_1,B} = \{w \in \mathbb{R}^d : \|w\|_1 \leq B\}$ for $B = 1/\lambda$.

What is Rademacher complexity of $\mathcal{F}_{\ell_1,B}$?

$$\begin{aligned} \mathrm{Rad}_n(\mathcal{F}_{\ell_1,B}) &= \mathbb{E}\mathbb{E}_\varepsilon \left[\sup_{\|w\|_1 \leq B} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i w^\mathsf{T} X_i \right| \right] \\ &= B \mathbb{E}\mathbb{E}_\varepsilon \left[\sup_{\|u\|_1 \leq 1} \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i u^\mathsf{T} X_i \right| \right] \\ &= B \mathbb{E}\mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \right\|_\infty . \end{aligned}$$

Rademacher complexity of ℓ_1-bounded linear predictors

Can show, using martingale argument,

$$\mathbb{E}\mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right\|_\infty \leq \sqrt{\frac{O(\log d) \cdot \mathbb{E}\|X\|_\infty^2}{n}}.$$

Rademacher complexity of ℓ_1-bounded linear predictors

Can show, using martingale argument,

$$\mathbb{E}\mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i X_i \right\|_\infty \leq \sqrt{\frac{O(\log d) \cdot \mathbb{E}\|X\|_\infty^2}{n}}.$$

Rademacher complexity of $\mathcal{F}_{\ell_1,B} = \{w \in \mathbb{R}^d : \|w\|_1 \leq B\}$:

$$\mathrm{Rad}_n(\mathcal{F}_{\ell_1,B}) \leq B\sqrt{\frac{O(\log d) \cdot \mathbb{E}\|X\|_\infty^2}{n}}.$$

Rademacher complexity of ℓ_1-bounded linear predictors

Can show, using martingale argument,

$$\mathbb{E}\mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right\|_\infty \leq \sqrt{\frac{O(\log d) \cdot \mathbb{E}\|X\|_\infty^2}{n}}.$$

Rademacher complexity of $\mathcal{F}_{\ell_1,B} = \{w \in \mathbb{R}^d : \|w\|_1 \leq B\}$:

$$\mathrm{Rad}_n(\mathcal{F}_{\ell_1,B}) \leq B\sqrt{\frac{O(\log d) \cdot \mathbb{E}\|X\|_\infty^2}{n}}.$$

Let $\mathcal{X} = \{-1,+1\}^d$. Then $\|x\|_2^2 = d$ but $\|x\|_\infty^2 = 1$ for all $x \in \mathcal{X}$.

Dependence on d much better than using bound for ℓ_2-bounded linear predictors, which would have looked like $B\sqrt{d/n}$.

Rademacher complexity of ℓ_1-bounded linear predictors

Can show, using martingale argument,

$$\mathbb{E}\mathbb{E}_\varepsilon \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right\|_\infty \leq \sqrt{\frac{O(\log d) \cdot \mathbb{E}\|X\|_\infty^2}{n}}.$$

Rademacher complexity of $\mathcal{F}_{\ell_1,B} = \{w \in \mathbb{R}^d : \|w\|_1 \leq B\}$:

$$\mathrm{Rad}_n(\mathcal{F}_{\ell_1,B}) \leq B\sqrt{\frac{O(\log d) \cdot \mathbb{E}\|X\|_\infty^2}{n}}.$$

Let $\mathcal{X} = \{-1,+1\}^d$. Then $\|x\|_2^2 = d$ but $\|x\|_\infty^2 = 1$ for all $x \in \mathcal{X}$.

Dependence on d much better than using bound for ℓ_2-bounded linear predictors, which would have looked like $B\sqrt{d/n}$.

This kind of bound is used to study generalization of *AdaBoost*.

Other examples of Rademacher complexity

- $\mathcal{F}$ = any class of $\{0,1\}$-valued functions with *VC dimension* V:

$$\mathrm{Rad}_n(\mathcal{F}) = O\left(\sqrt{\frac{V}{n}}\right).$$

- $\mathcal{F}$ = ReLU networks of depth D with parameter matrices of Frobenius norm ≤ 1:

$$\mathrm{Rad}_n(\mathcal{F}) = O\left(\sqrt{\frac{D \cdot \mathbb{E}\|X\|_2^2}{n}}\right).$$

- $\mathcal{F}$ = Lipschitz functions from $[0,1]^d$ to $\mathbb{R}$:

$$\mathrm{Rad}_n(\mathcal{F}) = O\left(n^{-1/(2+d)}\right).$$

- $\mathcal{F}$ = functions from $[0,1]^d$ to $\mathbb{R}$ with Lipschitz k-th derivatives:

$$\mathrm{Rad}_n(\mathcal{F}) = O\left(n^{-(k+1)/(2(k+1)+d)}\right).$$

Questions

Are these the “right” notions of complexity?

Questions

Are these the "right" notions of complexity?

- For SVM, the complexity of ℓ_2-bounded linear predictors is relevant because ℓ_2-regularization explicitly ensures the solution to SVM problem is ℓ_2-bounded.

Questions

Are these the "right" notions of complexity?

- For SVM, the complexity of ℓ_2-bounded linear predictors is relevant because ℓ_2-regularization explicitly ensures the solution to SVM problem is ℓ_2-bounded.
- Do training algorithms for neural nets lead to Frobenius norm-bounded parameter matrices?

Questions

Are these the "right" notions of complexity?

- For SVM, the complexity of ℓ_2-bounded linear predictors is relevant because ℓ_2-regularization explicitly ensures the solution to SVM problem is ℓ_2-bounded.

- Do training algorithms for neural nets lead to Frobenius norm-bounded parameter matrices?

Do complexity bounds suggest different algorithms?

Beyond uniform convergence

Deficiencies of uniform convergence analysis

- For certain loss functions, if $\mathcal{R}(f)$ is small, then variance of $\mathcal{R}_n(f)$ is also small, and bound should reflect this.
  - Instead of Hoeffding's inequality, use concentration inequality that involves variance information (e.g., *Bernstein's inequality*).
- Overkill to require *all* functions in $\mathcal{F}$ to not over-fit.
  - Just need to worry about the f, e.g., with small empirical risk.
  - Solution: *Local* Rademacher complexity.

Example: Occam's razor bound

Suppose $\mathcal{F}$ is countable and we fix (*a priori*) a probability distribution $\pi = (\pi_f : f \in \mathcal{F})$ on $\mathcal{F}$.

- Think of π as placing bets on which functions are likely to be the one to be picked by your learning algorithm.

Example: Occam's razor bound

Suppose $\mathcal{F}$ is countable and we fix (*a priori*) a probability distribution $\pi = (\pi_f : f \in \mathcal{F})$ on $\mathcal{F}$.

- Think of π as placing bets on which functions are likely to be the one to be picked by your learning algorithm.

For any fixed $f \in \mathcal{F}$,

$$\mathbb{P}\left(|\mathcal{R}_n(f) - \mathcal{R}(f)| \geq t_f\right) \leq 2\exp(-2nt_f^2)$$

for any $t_f > 0$, by Hoeffding's inequality and union bound.

Note: We can choose the t_f's non-uniformly.

Occam's razor bound (continued)

Let $t_f = \sqrt{\frac{\ln(1/\pi_f)+\ln(2/\delta)}{2n}}$.

By union bound,

$$
\begin{aligned}
&\mathbb{P}\left(\exists f \in \mathcal{F} \text{ s.t. } |\mathcal{R}_n(f) - \mathcal{R}(f)| \geq t_f\right) \\
&\quad \leq \sum_{f \in \mathcal{F}} \mathbb{P}\left(|\mathcal{R}_n(f) - \mathcal{R}(f)| \geq t_f\right) \\
&\quad \leq \sum_{f \in \mathcal{F}} 2\exp(-2nt_f^2) = \sum_{f \in \mathcal{F}} \pi_f \delta = \delta.
\end{aligned}
$$

Occam's razor bound (continued)

Let $t_f = \sqrt{\frac{\ln(1/\pi_f)+\ln(2/\delta)}{2n}}$.

By union bound,

$$\begin{aligned}
&\mathbb{P}\left(\exists f \in \mathcal{F} \text{ s.t. } |\mathcal{R}_n(f) - \mathcal{R}(f)| \geq t_f\right) \\
&\quad \leq \sum_{f \in \mathcal{F}} \mathbb{P}\left(|\mathcal{R}_n(f) - \mathcal{R}(f)| \geq t_f\right) \\
&\quad \leq \sum_{f \in \mathcal{F}} 2\exp(-2nt_f^2) = \sum_{f \in \mathcal{F}} \pi_f \delta = \delta.
\end{aligned}$$

Theorem. For any $\delta \in (0,1)$,

$$\mathbb{P}\left(\forall f \in \mathcal{F} \;:\; |\mathcal{R}_n(f) - \mathcal{R}(f)| < \sqrt{\frac{\ln(1/\pi_f) + \ln(2/\delta)}{2n}}\right) \geq 1 - \delta.$$

Occam's razor bound (continued)

Let $t_f = \sqrt{\frac{\ln(1/\pi_f)+\ln(2/\delta)}{2n}}$.

By union bound,

$$
\begin{aligned}
&\mathbb{P}\left(\exists f \in \mathcal{F} \text{ s.t. } |\mathcal{R}_n(f) - \mathcal{R}(f)| \geq t_f\right) \\
&\quad \leq \sum_{f\in\mathcal{F}} \mathbb{P}\left(|\mathcal{R}_n(f) - \mathcal{R}(f)| \geq t_f\right) \\
&\quad \leq \sum_{f\in\mathcal{F}} 2\exp(-2nt_f^2) = \sum_{f\in\mathcal{F}} \pi_f \delta = \delta.
\end{aligned}
$$

Theorem. For any $\delta \in (0,1)$,

$$\mathbb{P}\left(\forall f \in \mathcal{F} \ : \ |\mathcal{R}_n(f) - \mathcal{R}(f)| < \sqrt{\frac{\ln(1/\pi_f) + \ln(2/\delta)}{2n}}\right) \geq 1 - \delta.$$

Better bound for functions f with higher "prior probability" π_f!

Other forms of generalization analysis

- Stability
  - If a learning algorithm's output does not change much if a single data point is changed, then its output will generalize.
  - Connections to differential privacy and regularization.
- Compression bounds
  - If a learning algorithm's output is invariant to all but a small number $k \ll n$ of training data (e.g., $\#$ support vectors in SVM), then get bound of the form $\sqrt{k/(n-k)}$.
- Direct analyses
  - Some well-known learning algorithms do not fit the mold of typical (regularized) ERM algorithm, and seem to require a direct analysis.
  - E.g., nearest neighbor rule.
- Many others

Many active areas of research in learning theory

- Implicit bias of optimization algorithms
  - E.g., gradient descent for least squares linear regression converges to solution of smallest norm.
  - What about for other problems?
- Efficient algorithms for non-linear models
  - E.g., polynomials, neural networks, kernel machines.
  - Understand if/why existing algorithms work!
- Learning algorithms with robustness guarantees
  - Noisy labels, missing / malformed data, heavy-tail distributions, adversarial corruptions, etc.
- Interactive learning
  - Learning algorithms that interact with external environment (e.g., bandits, active learning, reinforcement learning).
- More: see proceedings of Conference on Learning Theory!