Generalization Bounds

CMPUT 296: Basics of Machine Learning

Textbook Ch.12

Logistics

- Thought questions #3 marked *(finally...)*
- **Thought questions #4** due **TODAY at 11:59pm** (Mountain time)
    - Anything between logistic regression and generalization bounds is fair game
- **Last class** will be on **Tuesday (Dec 1)**
    - It will be a review class for the final exam
- **Assignment #3** due **Thursday Dec 3**
        - But there is no class on Dec 3
- **FINAL EXAM** will be **Friday Dec 18**

Outline

1. Recap & Logistics
2. McDiarmid's Inequality
3. Proof of Generalization Bound for Binary Classification

Recap: Rademacher Complexity

The **empirical Rademacher complexity** of $\mathcal{F}$ with respect to $\mathcal{D}$ is

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

where

- $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid 1 \leq i \leq n\}$ is a dataset
- $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ is a vector of n random variables, with $\sigma_i \overset{i.i.d}{\sim} \text{Uniform}\{-1, +1\}$ **("Rademacher variables")**
- $\mathcal{F}$ is a hypothesis class

The **Rademacher complexity** of a hypothesis class $\mathcal{F}$ is the expected **empirical Rademacher complexity** over all datasets $\mathcal{D}$ (of size n):

$$R_n(\mathcal{F}) = \mathbb{E}\left[\hat{R}_{\mathcal{D}}(\mathcal{F})\right]$$

Recap: Uniform Generalization Bound for Binary Classification

Theorem:

Let $\mathscr{F}$ be a family of binary classification functions taking values in $\{-1, +1\}$, and let $\mathbf{cost}$ be the 0-1 classification cost.

Then for every $f \in \mathscr{F}$, and every $\delta > 0$,

$$C(f) \leq \hat{C}(f) + R_n(\mathscr{F}) + \sqrt{\frac{\log 1/\delta}{2n}}$$

with probability $1 - \delta$.

Implication: Training Error vs. Test Error

$$C(f) \leq \hat{C}(f) + R_n(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2n}}, \quad \forall f \in \mathcal{F}$$

1. The **more data** you have, the closer your training error will be to the true generalization error
2. The **simpler** your hypothesis class, the closer your training error will be to the true error

- Notice that this bound applies to **all** $f \in \mathcal{F}$

Question: Why does this bound not go to zero for infinite data?

McDiarmid's Inequality

Let $S = (X_1, \ldots, X_n) \in \mathcal{X}^n$ be a vector of $n \geq 1$ independent random variables.

If there exist constants $c_1, \ldots, c_n$ such that $f : \mathcal{X}^n \to \mathbb{R}$ satisfies

$$\left| f(x_1, \ldots, x_k, \ldots, x_n) - f(x_1, \ldots, x'_k, \ldots, x_n) \right| \leq c_k$$

for every $1 \leq k \leq n$ and any values $x_1, \ldots, x_n, x'_k \in \mathcal{X}$, then for all $\epsilon > 0$,

1. $$\Pr\left[f(S) - \mathbb{E}[f(S)] \geq \epsilon\right] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right), \text{ and}$$

2. $$\Pr\left[f(S) - \mathbb{E}[f(S)] \leq -\epsilon\right] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Proof: Generalization Bound for Binary Classification

Update our notation for empirical cost to indicate the dataset:

$$\hat{C}_{\mathscr{D}}(f) = \frac{1}{n} \sum_{i=1}^{n} \text{cost}(f(\mathbf{x}_i), y_i)$$

Define $\Phi(\mathscr{D})$ as the maximum difference between the true cost of a predictor and the empirical cost on dataset $\mathscr{D}$:

$$\Phi(\mathscr{D}) \doteq \max_{f \in \mathscr{F}} C(f) - \hat{C}_{\mathscr{D}}(f)$$

(A bit sloppy; should really use $\sup_{f \in \mathscr{F}}$)

Proof (2): Bounding $\left|\Phi(\mathscr{D}) - \Phi(\mathscr{D}')\right|$

Let $\mathscr{D}$ and $\mathscr{D}'$ be two datasets of n observations that differ in **exactly one** datapoint: $(x_k, y_k) \neq (x'_k, y'_k)$, and $(x_i, y_i) = (x'_i, y'_i)$ for all $i \neq k$

$$\Phi(\mathscr{D}) - \Phi(\mathscr{D}') = \left(\max_{f \in \mathscr{F}} C(f) - \hat{C}_{\mathscr{D}}(f)\right) - \left(\max_{g \in \mathscr{F}} C(g) - \hat{C}_{\mathscr{D}'}(g)\right)$$

$$\leq \max_{f \in \mathscr{F}} \left(C(f) - \hat{C}_{\mathscr{D}}(f)\right) - \left(C(f) - \hat{C}_{\mathscr{D}'}(f)\right)$$

$$= \max_{f \in \mathscr{F}} C(f) - \hat{C}_{\mathscr{D}}(f) - C(f) + \hat{C}_{\mathscr{D}'}(f)$$

$$= \max_{f \in \mathscr{F}} \hat{C}_{\mathscr{D}'}(f) - \hat{C}_{\mathscr{D}}(f)$$

Proof (3):

Bounding $\left|\Phi(\mathscr{D}) - \Phi(\mathscr{D}')\right|$

$$\Phi(\mathscr{D}) - \Phi(\mathscr{D}') \leq \max_{f \in \mathscr{F}} \hat{C}_{\mathscr{D}'}(f) - \hat{C}_{\mathscr{D}}(f)$$

$$= \max_{f \in \mathscr{F}} \sum_{i=1}^{n} \frac{\text{cost}(f(x_i), y_i) - \text{cost}(f(x_i'), y_i')}{n}$$

$$= \max_{f \in \mathscr{F}} \frac{\text{cost}\left(f(x_k), y_k\right) - \text{cost}\left(f(x_k'), y_k'\right)}{n} \qquad = \frac{1}{n}$$

By an identical argument, $\Phi(\mathscr{D}') - \Phi(\mathscr{D}) \leq 1/n$. So

$$\left|\Phi(\mathscr{D}) - \Phi(\mathscr{D}')\right| \leq \frac{1}{n}$$

Proof (4): Bounding $\Phi(\mathscr{D})$

$$\left|\Phi(\mathscr{D}) - \Phi(\mathscr{D}')\right| \leq \frac{1}{n}$$

Apply McDiarmid's Inequality:

$$\Pr\left[\Phi(\mathscr{D}) - \mathbb{E}[\Phi(\mathscr{D})] \geq \epsilon\right] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n} 1/n^2}\right)$$

$$= \exp(-2n\epsilon^2) = \delta$$

$$\Pr\left[\Phi(\mathscr{D}) - \mathbb{E}[\Phi(\mathscr{D})] \geq \sqrt{\frac{\log 1/\delta}{2n}}\right] \leq \delta$$

$$\Pr\left[\Phi(\mathscr{D}) \leq \mathbb{E}[\Phi(\mathscr{D})] + \sqrt{\frac{\log 1/\delta}{2n}}\right] \geq 1 - \delta$$

$$\sum_{i=1}^{n} \frac{1}{n^2} = n\frac{1}{n^2} = \frac{1}{n}$$

$$\delta = \exp(-2n\epsilon^2)$$

$$\iff \log\delta = -2n\epsilon^2$$

$$\iff \frac{-\log\delta}{2n} = \epsilon^2$$

$$\iff \epsilon = \sqrt{\frac{\log 1/\delta}{2n}}$$

Proof (5): Bounding $\mathbb{E}[\Phi(\mathscr{D})]$

With probability $1-\delta$, we have $\Phi(\mathscr{D}) \leq \mathbb{E}[\Phi(\mathscr{D})] + \sqrt{\frac{\log 1/\delta}{2n}}$.

$$\mathbb{E}[\Phi(\mathscr{D})] = \mathbb{E}_{\mathscr{D}}\left[\max_{f\in\mathscr{F}} C(f) - \hat{C}_{\mathscr{D}}(f)\right]$$

$$= \mathbb{E}_{\mathscr{D}}\left[\max_{f\in\mathscr{F}} \mathbb{E}_{\mathscr{D}'}\left[\hat{C}_{\mathscr{D}'}(f) - \hat{C}_{\mathscr{D}}(f)\right]\right]$$

$$\leq \mathbb{E}_{\mathscr{D},\mathscr{D}'}\left[\max_{f\in\mathscr{F}} \hat{C}_{\mathscr{D}'}(f) - \hat{C}_{\mathscr{D}}(f)\right]$$

(max is a convex function)

$$= \mathbb{E}_{\mathscr{D},\mathscr{D}'}\left[\max_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n}\left(\text{cost}(f(x_i), y_i) - \text{cost}(f(x_i'), y_i')\right)\right]$$

Jensen's inequality

For any **convex** function $f : \mathscr{X} \to \mathbb{R}$,

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

Proof (6): Bounding $\mathbb{E}[\Phi(\mathscr{D})]$

$$\mathbb{E}_{\mathscr{D}}[\Phi(\mathscr{D})] \leq \mathbb{E}_{\mathscr{D},\mathscr{D}'}\left[\max_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n}\left(\text{cost}(f(x_i),y_i) - \text{cost}(f(x_i'),y_i')\right)\right]$$

$$= \mathbb{E}_{\mathscr{D},\mathscr{D}',\sigma}\left[\max_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(\text{cost}(f(x_i),y_i) - \text{cost}(f(x_i'),y_i')\right)\right]$$

$$\leq \mathbb{E}_{\mathscr{D},\mathscr{D}',\sigma}\left[\left(\max_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\,\text{cost}(f(x_i),y_i)\right) + \left(\max_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n} -\sigma_i\,\text{cost}(f(x_i'),y_i')\right)\right]$$

$$= \mathbb{E}_{\mathscr{D},\sigma}\left[\max_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n}\sigma_i\,\text{cost}(f(x_i),y_i)\right] + \mathbb{E}_{\mathscr{D}',\sigma}\left[\max_{f\in\mathscr{F}} \frac{1}{n}\sum_{i=1}^{n} -\sigma_i\,\text{cost}(f(x_i'),y_i')\right]$$

Proof (7): Bounding $\mathbb{E}[\Phi(\mathscr{D})]$

$$\mathbb{E}_{\mathscr{D}}[\Phi(\mathscr{D})] \leq \mathbb{E}_{\mathscr{D},\sigma}\left[\max_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\,\mathrm{cost}(f(x_i),y_i)\right] + \mathbb{E}_{\mathscr{D}',\sigma}\left[\max_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}-\sigma_i\,\mathrm{cost}(f(x_i'),y_i')\right]$$

$$= 2\mathbb{E}_{\mathscr{D},\sigma}\left[\max_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\,\mathrm{cost}(f(x_i),y_i)\right]$$

$$= 2\mathbb{E}_{\mathscr{D},\sigma}\left[\max_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i(1-I_{f(x_i)=y_i})\right]$$

$$= 2\mathbb{E}_{\mathscr{D},\sigma}\left[\max_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\frac{1-y_if(x_i)}{2}\right]$$

$$= 2\frac{1}{2}\mathbb{E}_{\mathscr{D},\sigma}\left[\max_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\right] + \frac{1}{2}2\mathbb{E}_{\mathscr{D},\sigma}\left[\max_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i - y_if(x_i)\right]$$

$$= \mathbb{E}_{\mathscr{D},\sigma}\left[\max_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_if(x_i)\right] = \mathbb{E}_{\mathscr{D}}\left[\mathbb{E}_{\sigma}\left[\max_{f\in\mathscr{F}}\frac{1}{n}\sum_{i=1}^{n}\sigma_if(x_i)\right]\right] = \mathbb{E}_{\mathscr{D}}\left[\hat{R}_{\mathscr{D}}(\mathscr{F})\right] = \boxed{R_n(\mathscr{F})}$$

Proof: Putting it all together

1. With probability $1-\delta$, we have $\Phi(\mathscr{D}) \leq \mathbb{E}[\Phi(\mathscr{D})] + \sqrt{\frac{\log 1/\delta}{2n}}$
2. $\mathbb{E}_{\mathscr{D}}[\Phi(\mathscr{D})] \leq R_n(\mathscr{F})$
3. With probability $1-\delta$,

$$\Phi(\mathscr{D}) \leq R_n(\mathscr{F}) + \sqrt{\frac{\log 1/\delta}{2n}}$$

$$\max_{f \in \mathscr{F}} C(f) - \hat{C}_{\mathscr{D}}(f) \leq R_n(\mathscr{F}) + \sqrt{\frac{\log 1/\delta}{2n}}$$

$$C(f) \leq \hat{C}_{\mathscr{D}}(f) + R_n(\mathscr{F}) + \sqrt{\frac{\log 1/\delta}{2n}}, \quad \forall f \in \mathscr{F}. \blacksquare$$

Summary

- **McDiarmid's Inequality** is a generalization of Hoeffding's Inequality for "stable enough" functions
- We can use McDiarmid's Inequality to prove upper bounds on generalization performance such as (for binary classification)

$$C(f) \leq \hat{C}_{\mathcal{D}}(f) + R_n(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2n}}, \quad \forall f \in \mathcal{F}$$

- Bounds of this kind shed light on the relationship between **training error** and **generalization error**
  - The more flexible your hypothesis class (larger $R_n(\mathcal{F})$), the bigger the difference can be
  - This difference is the origin of **overfitting**