

Machine Learning Theory (CS 6783)

Lecture 4 : MDL principle, Uniform Rate and Infinite classes

1 Recap

1. ERM Algorithm: $\hat{f}_{\text{ERM}} = \operatorname{argmin}_{f \in \mathcal{F}} \sum_{(x,y) \in S} \ell(f(x), y)$
2. For the ERM algorithm, we have

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

3. For finite class,

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

What about infinite \mathcal{F} ?

2 MDL bound (Occam's Razor Principle)

We saw how one can get bounds for the case when \mathcal{F} has finite cardinality. How about the case when \mathcal{F} has infinite cardinality? To start with, let us consider the case when \mathcal{F} is a countable set. One thing we can do is to try to be smarter with the application of union bound and Hoeffding bound applied in the analysis of the finite case.

Claim 1. For any countable set \mathcal{F} , any fixed distribution π on \mathcal{F} ,

$$\mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left\{ \left| L_D(f) - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right| - \sqrt{\frac{\log(n/\pi^2(f))}{n}} \right\} \right] \leq \frac{4}{\sqrt{n}}$$

Proof. The basic idea is to use Hoeffding bound along with union bound as before, but instead of using same ϵ for every $f \in \mathcal{F}$ in Hoeffding bound, we use f specific $\epsilon(f)$. We shall specify the exact form of $\epsilon(f)$ later. For now note that, since the losses are bounded by 1,

$$\sup_{f \in \mathcal{F}} \left\{ \left| L_D(f) - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right| - \epsilon(f) \right\} \leq 0 + 2 \mathbf{1}_{\{\sup_{f \in \mathcal{F}} \{|L_D(f) - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t)| - \epsilon(f) > 0\}\}}$$

Hence, taking expectation w.r.t. sample we have that

$$\mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left\{ \left| L_D(f) - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right| - \epsilon(f) \right\} \right] \leq 2P \left(\sup_{f \in \mathcal{F}} \left\{ \left| L_D(f) - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right| - \epsilon(f) > 0 \right\} \right)$$

By Hoeffding inequality, for any fixed $f \in \mathcal{F}$

$$P \left(\left| \mathbb{E}[\ell(f(x), y)] - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right| - \epsilon(f) > 0 \right) \leq 2 \exp \left(-\frac{\epsilon^2(f)n}{2} \right)$$

Taking union bound we have,

$$P \left(\sup_{f \in \mathcal{F}} \left| \mathbb{E}[\ell(f(x), y)] - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right| - \epsilon(f) > 0 \right) \leq \sum_{f \in \mathcal{F}} 2 \exp \left(-\frac{\epsilon^2(f)n}{2} \right)$$

Hence we conclude that

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

For the prior choice of π of distribution over set \mathcal{F} , let us use

$$\epsilon(f) = \sqrt{\frac{\log(n/\pi^2(f))}{n}}$$

Hence we can conclude that,

$$\begin{aligned} \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left\{ \left| L_D(f) - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right| - \sqrt{\frac{\log(n/\pi^2(f))}{n}} \right\} \right] &\leq 4 \sum_{f \in \mathcal{F}} \exp \left(-\frac{\epsilon^2(f)n}{2} \right) \\ &\leq \frac{4 \sum_f \pi(f)}{\sqrt{n}} = \frac{4}{\sqrt{n}} \end{aligned}$$

□

The above claim provides us an intuition for MDL principle, the MDL learning rule picks the hypothesis in \mathcal{F} as follows :

$$\hat{y}_{\text{mdl}} = \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) + 3 \sqrt{\frac{\log(n/\pi^2(f))}{n}}$$

Interpretation : minimize empirical error while staying close to prior π . Why is this learning rule appealing ?

Let us use the claim above to analyze the learning rule. Note that from the above claim, we have that,

$$\mathbb{E}_S \left[L_D(\hat{y}_{\text{mdl}}) - \frac{1}{n} \sum_{t=1}^n \ell(\hat{y}_{\text{mdl}}(x_t), y_t) - \sqrt{\frac{\log(n/\pi^2(\hat{y}_{\text{mdl}}))}{n}} \right] \leq \frac{4}{\sqrt{n}}$$

By definition of \hat{y}_{mdl} we can conclude that

$$\mathbb{E}_S \left[L_D(\hat{y}_{\text{mdl}}) - \inf_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{t=1}^n \ell(\hat{y}_{\text{mdl}}(x_t), y_t) + \sqrt{\frac{\log(n/\pi^2(\hat{y}_{\text{mdl}}))}{n}} \right\} \right] \leq \frac{4}{\sqrt{n}}$$

In other words,

$$\mathbb{E}_S [L_D(\hat{y}_{\text{mdl}})] \leq \mathbb{E}_S \left[\inf_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) + \sqrt{\frac{\log(n/\pi^2(f))}{n}} \right\} \right] + \frac{4}{\sqrt{n}}$$

Let $f_D = \underset{f \in \mathcal{F}}{\text{argmin}} L_D(f)$, replacing the infimum above we conclude that

$$\begin{aligned} \mathbb{E}_S [L_D(\hat{y}_{\text{mdl}})] &\leq \mathbb{E}_S \left[\frac{1}{n} \sum_{t=1}^n \ell(f_D(x_t), y_t) + \sqrt{\frac{\log(n/\pi^2(f_D))}{n}} \right] + \frac{4}{\sqrt{n}} \\ &= L_D(f_D) + \sqrt{\frac{\log(n/\pi^2(f_D))}{n}} + \frac{4}{\sqrt{n}} \\ &= \inf_{f \in \mathcal{F}} L_D(f) + \sqrt{\frac{\log(n/\pi^2(f_D))}{n}} + \frac{4}{\sqrt{n}} \end{aligned} \tag{1}$$

(2)

Thus with the above bound, even for countably infinite \mathcal{F} we can get bounds on $\mathbb{E}_S [L_D(\hat{y})] - \inf_{f \in \mathcal{F}} L_D(f)$ that decreases with n .

3 Infinite Hypothesis Class : first attempt

As a first attempt, one can think of approximating the function class to desired accuracy by a finite number of representative elements. We call this a point-wise cover.

Definition 1. We say that set $\mathcal{F}_\delta = \{\tilde{f}_1, \dots, \tilde{f}_N\}$ is an δ point-wise cover for function class \mathcal{F} if $\forall f \in \mathcal{F}$ there exists $i \in [N]$ s.t.

$$\sup_{x,y} |\ell(f(x), y) - \ell(\tilde{f}_i(x), y)| \leq \delta$$

Further define $N(\delta)$ to be the smallest N such that there exists an δ cover of \mathcal{F} of cardinality at most N .

Claim 2. For any function class \mathcal{F} , we have that

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \leq \inf_{\delta > 0} \left\{ 4\delta + \sqrt{\frac{\log N(\delta)}{n}} \right\}$$

Proof. Let $\mathcal{F}_\delta = \{\tilde{f}_1, \dots, \tilde{f}_{N(\delta)}\}$ be an δ cover for the function class \mathcal{F} . Further for every $f \in \mathcal{F}$, let $i(f)$ correspond to the index of the element in \mathcal{F}_δ that is δ close to that f . Now note that,

$$\begin{aligned} & \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left\{ \mathbb{E} [\ell(f(x), y)] - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right\} \right] \\ & \leq \mathbb{E}_S \left[\max_{i \in [N_\delta]} \left\{ \mathbb{E} [\ell(\tilde{f}_i(x), y)] - \frac{1}{n} \sum_{t=1}^n \ell(\tilde{f}_i(x_t), y_t) \right\} \right] \\ & \quad + \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E} [\ell(f(x), y)] - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) - \mathbb{E} [\ell(\tilde{f}_{i(f)}(x), y)] + \frac{1}{n} \sum_{t=1}^n \ell(\tilde{f}_{i(f)}(x_t), y_t) \right| \right] \\ & \leq \sqrt{\frac{\log N(\delta)}{n}} + 4\delta \end{aligned}$$

where the first term in the last inequality is by using the finite class bound and the second term is by using the definition of δ cover as $\tilde{f}_{i(f)}$ is δ close to f . Since choice of δ was arbitrary we can take the infimum over choices of δ to conclude the proof. \square

Example : linear predictor, absolute loss, 1 dimension

$$f(x) = f \cdot x, \quad \mathcal{F} = \mathcal{X} = [-1, 1], \quad \mathcal{Y} = [-1, 1], \quad \ell(y', y) = |y - y'|$$

$$N_\delta = \frac{2}{\delta}, \text{ Cover given by } f_1 = -1, f_2 = -1 + \delta, \dots, f_{N_\delta-1} = 1 - \delta, f_{N_\delta} = 1.$$

$$V_n^{\text{stat}}(\mathcal{F}) \leq \sqrt{\frac{\log n}{n}}$$

Example : linear predictor/loss, d dimensions

$$f(x) = \mathbf{f}^\top \mathbf{x}. \quad \mathcal{F} = \mathcal{X} = \{\mathbf{v} \in \mathbb{R}^d : \|\mathbf{v}\|_2 \leq 1\}. \quad \mathcal{Y} = [-1, 1]. \quad \ell(y', y) = y \cdot y'$$

$$N_\delta = \Theta\left(\frac{2}{\delta}\right)^d$$

$$V_n^{\text{stat}}(\mathcal{F}) \leq \sqrt{\frac{d \log n}{n}}$$

Example : thresholds

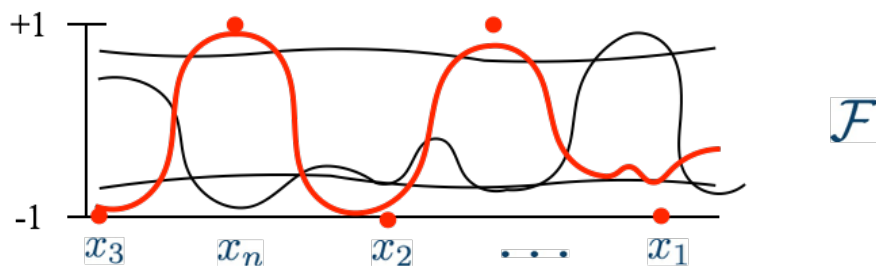
$$f(x) = \text{sign}(f - x), \quad \mathcal{F} = \mathcal{X} = [-1, 1], \quad \mathcal{Y} = \{-1, 1\}, \quad \ell(y', y) = \mathbf{1}_{\{y \neq y'\}}, \quad N_\delta = \infty \text{ for any } \delta < 1.$$

4 Symmetrization and Rademacher Complexity

$$\begin{aligned}
 \mathbb{E}_S [L_D(\hat{y}_{\text{erm}})] - \inf_{f \in \mathcal{F}} L_D(f) &\leq \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left\{ \mathbb{E} [\ell(f(x), y)] - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right\} \right] \\
 &\leq \mathbb{E}_{S, S'} \left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{t=1}^n \ell(f(x'_t), y'_t) - \frac{1}{n} \sum_{t=1}^n \ell(f(x_t), y_t) \right\} \right] \\
 &= \mathbb{E}_{S, S'} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{t=1}^n \epsilon_t (\ell(f(x'_t), y'_t) - \ell(f(x_t), y_t)) \right\} \right] \\
 &\leq 2 \mathbb{E}_S \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t) \right\} \right] \\
 &=: \mathcal{R}_n(\ell \circ \mathcal{F})
 \end{aligned}$$

Where in the above each ϵ_t is a Rademacher random variable that is +1 with probability 1/2 and -1 with probability 1/2. The above is called Rademacher complexity of the loss class $\ell \circ \mathcal{F}$. In general Rademacher complexity of a function class measures how well the function class correlates with random signs. The more it can correlate with random signs the more complex the class is.

Example : $\mathcal{X} = [0, 1]$, $\mathcal{Y} = [-1, 1]$



How does this help?