Lecture 3 VC Theory for Generalization Error

He Li Machine Learning Course

2017/10/10

1. Simple Classifiers

$(x_i, y_i)_{i=1}^n$, $x_i \in \mathbb{R}^d, x = (x^{(1)}, ..., x^{(d)})$ instance, $y \in \{\pm 1\}$ label

- Decision Tree: hypothesis space $\mathcal{F} = \{\text{all decision tree}, \text{depth} \leq \alpha\}$
- Linear Classifier: hypothesis space $\mathcal{F} = \left\{(\omega, b) : \omega \in \mathbb{R}^d, b \in \mathbb{R}, \|\omega\| = 1\right\}$, then classifier is:

$$f(x) = \text{sgn}(\omega^T x + b)$$

Empirical Risk Minimiaztion

Given $\mathcal{F}$, training data $(x_i, y_i)_{i=1}^n$. **Empirical Risk Minimiaztion(ERM)** is a algorithm, which finds $f \in \mathcal{F}$, s.t.

$$\hat{f} = \underset{f \in \mathcal{F}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^n I\left[y_i \neq f(x_i)\right]$$

- Empirical Error: $\hat{f}$'s performance on training data. Denote

$$\mathbb{P}_S(y_i \neq f(x_i)) = \frac{1}{n} \sum_{i=1}^n I\left[y_i \neq \hat{f}(x_i)\right]$$

  where$S = (x_i, y_i)_{i=1}^n$

- Generalization Error: $\hat{f}$'s performance on test data. Denote

$$\mathbb{P}_D(y_i \neq f(x_i)) = \underset{(x,y)\sim D}{\mathbb{E}} \{I(y \neq \hat{f}(x))\}$$

A small empirical error cannot indicate a small generalization error(cannot using Chernoff bound), since on training data, $z_i = I\left[y_i \neq \hat{f}(x_i)\right]$ are not independent.

Finite hypothesis space

Consider $\mathcal{F}$ is finite, $|\mathcal{F}| < \infty$. ERM learns $\hat{f} \in \mathcal{F}$, then

$$P\left\{\mathbb{P}_D(y_i \neq f(x_i)) - \mathbb{P}_S(y_i \neq f(x_i)) \geq \epsilon\right\} \leq |\mathcal{F}| e^{-2n\epsilon^2}$$

called **union bound**. If we fix $f \in \mathcal{F}$, $P\left\{\mathbb{P}_D(y_i \neq f(x_i)) - \mathbb{P}_S(y_i \neq f(x_i)) \geq \epsilon\right\}$ satisties Chernoff bound.

$$P\left\{\mathbb{P}_D(y_i \neq f(x_i)) - \mathbb{P}_S(y_i \neq f(x_i)) \geq \epsilon\right\} \leq e^{-2n\epsilon^2}$$

Then, by union bound

$$P\left\{\exists f \in \mathcal{F},\ \mathbb{P}_D(y_i \neq f(x_i)) - \mathbb{P}_S(y_i \neq f(x_i)) \geq \epsilon\right\} \leq |\mathcal{F}| e^{-2n\epsilon^2}$$

Then we have our conclusion.

2. VC bound

Uniform Law of Large Numbers

Denote $z_i = (x_i, y_i)$, $\phi_f(z_i) = I(y_i \neq f(x_i))$. Consider:

$$\mathbb{P}\left(\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_i \phi(z_i) - \mathbb{E}[\phi(z)]\right| \geq \epsilon\right)$$

Step I (Double Sample Trick)

Proposition: $X_1, ..., X_n, X_{n+1}, ..., X_{2n}$ are i.i.d Bernoulli $R.V.$. $\mathbb{E}(X) = p$. Denote $\nu_1 = \frac{1}{n}\sum_{i=1}^n X_i$, $\nu_2 = \frac{1}{n}\sum_{i=n+1}^{2n} X_i$. If $n \geq \frac{\ln 2}{\epsilon^2}, \epsilon > 0$, then

$$\frac{1}{2}\mathbb{P}\left(|\nu_1 - p| \geq 2\epsilon\right) \leq \mathbb{P}\left(|\nu_1 - \nu_2| \geq \epsilon\right) \leq 2\mathbb{P}\left(|\nu_1 - p| \geq \frac{\epsilon}{2}\right)$$

proof

right inequality:

$$|\nu_1 - \nu_2| \geq \epsilon \Longrightarrow |\nu_1 - p| \geq \frac{\epsilon}{2} \text{ or } |\nu_2 - p| \geq \frac{\epsilon}{2}$$

Therefore,

$$\begin{aligned}
\{|\nu_1 - \nu_2| \geq \epsilon\} &\subset \{|\nu_1 - p| \geq \frac{\epsilon}{2}\} \cup \{|\nu_2 - p| \geq \frac{\epsilon}{2}\} \\
\mathbb{P}\left(|\nu_1 - \nu_2| \geq \epsilon\right) &\leq \mathbb{P}\left(|\nu_1 - p| \geq \frac{\epsilon}{2} \cup |\nu 2 - p| \geq \frac{\epsilon}{2}\right) \\
&\leq \mathbb{P}\left(|\nu_1 - p| \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(|\nu_2 - p| \geq \frac{\epsilon}{2}\right) \\
&= 2\mathbb{P}\left(|\nu_1 - p| \geq \frac{\epsilon}{2}\right)
\end{aligned}$$

Similarly, left inequality:

$$|\nu_1 - p| \geq 2\epsilon, |\nu_2 - p| \leq \epsilon \Longrightarrow |\nu_1 - \nu_2| \geq \epsilon$$

□

Lemma (Homework)

$$\begin{aligned}
&\frac{1}{2}\mathbb{P}\left(\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^n \phi(z_i) - \mathbb{E}[\phi(z)]\right| \geq 2\epsilon\right) \\
\leq &\mathbb{P}\left(\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^n \phi(z_i) - \frac{1}{n}\sum_{i=n+1}^{2n} \phi(z_i)\right| \geq \epsilon\right) \\
\leq &2\mathbb{P}\left(\sup_{\phi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^n \phi(z_i) - \mathbb{E}[\phi(z)]\right| \geq \frac{\epsilon}{2}\right)
\end{aligned}$$

Step II (Symmetrization)

Denote

$$N^{\Phi}(z_1, ..., z_n) = |\{(\phi(z_1), ..., \phi(z_n)), \phi \in \Phi\}|$$
$$N^{\Phi}(n) = \max_{z_1 \dots z_n} N^{\Phi}(z_1 ... z_n)$$

$N^{\Phi}(n)$ is growth function.

Lemma

$$\mathbb{P}\left(\sup_{\phi\in\Phi} |\nu_1(z) - \nu_2(z)| \geq \epsilon\right) \leq \mathbb{E}\left[N^{\Phi}(z_1, ..., z_n)\right]\ 2e^{-2n\epsilon^2} \leq N^{\Phi}(2n)\, 2e^{-2n\epsilon^2}$$

where $\nu_1(z) = \frac{1}{n}\sum_{i=1}^{n} \phi(z_i)$, $\nu_2(z) = \frac{1}{n}\sum_{i=n+1}^{2n} \phi(z_i)$.

Draw a set, permutation, fix set (draw without replacement)

$$\mathbb{P}_{\text{permutation}}\left(\sup_{\phi\in\Phi} |\nu_1(z) - \nu_2(z)| \geq \epsilon\right) \leq N^{\Phi}(z_1, ..., z_n)\, 2e^{-2n\epsilon^2}$$

Take expectation,

$$\mathbb{E}_{\text{set}}\left\{\mathbb{P}_{\text{permutation}}\left(\sup_{\phi\in\Phi} |\nu_1(z) - \nu_2(z)| \geq \epsilon\right)\right\} \leq \mathbb{E}\left[N^{\Phi}(z_1, ..., z_n)\right]\ 2e^{-2n\epsilon^2}$$