Lecture 4: VC Theory

Lecturer: Liwei Wang

Scribe: Shihui Li, Zhu Xu, Juling Fan, Huixuan Zhang, Kelin Fu, Weiyuan Ding, Xiangyang Li, Xingyu Xiang

4.1 Recall

A more common way to use the **concentration inequality** is that:

let $\delta = 2e^{-2n\epsilon^2}$, as $P(|\frac{1}{n}\sum_{i=1}^{n} x_i - p| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$, we can declare that

$$|\frac{1}{n}\sum_{i=1}^{n} x_i - p| \leq \sqrt{\frac{\ln\frac{2}{\delta}}{2n}} = O(\sqrt{\ln\frac{1}{\delta}n})$$

with probability at least $1-\delta$.

4.2 Finite Hypothesis Space

Suppose $\hat{f} \in \mathcal{F}$ is learned from training data $(x_1, y_1), \ldots, (x_n, y_n)$, $|\mathcal{F}| < \infty$.

Define the training error as:

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

Define the test error as:

$$E\{I[Y \neq \hat{f}(X)]\} = \Pr(Y \neq \hat{f}(X))$$

We hope to conduct a **Worst-Case Analysis** to find an upper bound on the difference between the two errors, no matter how the model $\hat{f}$ has been learned. So we have:

$$\begin{aligned}
\Pr[\Pr_D(Y \neq \hat{f}(X)) - \frac{1}{n}\sum_{i=1}^{n} I[Y_i \neq \hat{f}(X_i)] \geq \epsilon] \leq \\
\Pr[\exists f \in \mathcal{F}, \Pr_D(Y \neq f(X)) - \frac{1}{n}\sum_{i=1}^{n} I[Y_i \neq f(X_i)] \geq \epsilon] \leq \\
\sum_{f\in\mathcal{F}} \Pr[\Pr_D(Y \neq f(X)) - \frac{1}{n}\sum_{i=1}^{n} I[Y_i \neq f(X_i)] \geq \epsilon] \leq \\
|\mathcal{F}|e^{-2n\epsilon^2}
\end{aligned} \tag{4.1}$$

From this, we can draw an intuitive conclusion that **the size of hypothesis space affects the upper bound of over-fitting probability**.

4.3 Infinite Hypothesis Space

As for the case where $|\mathcal{F}| = \infty$, note that we still have:

$$\begin{aligned}
\Pr[\Pr_D(Y \neq \hat{f}(X)) - \frac{1}{n}\sum_{i=1}^{n} I[Y_i \neq \hat{f}(X_i)] \geq \epsilon] \leq \\
\Pr[\exists f \in \mathcal{F}, \Pr_D(Y \neq f(X)) - \frac{1}{n}\sum_{i=1}^{n} I[Y_i \neq f(X_i)] \geq \epsilon]
\end{aligned} \tag{4.2}$$

4.3.1 Step I: Double Sample Trick

Lemma 4.1 *Consider $2n$ iid random variables $X_1, ..., X_n, X_{n+1}, ..., X_{2n}$ with $EX_i = p$. Let $\nu_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \nu_2 = \frac{1}{n}\sum_{i=n+1}^{2n} X_i$. For $n \geq \frac{\ln 2}{\epsilon^2}$, we have:*

$$\frac{1}{2}\Pr(|\nu_1 - p| \geq 2\epsilon) \leq \Pr(|\nu_1 - \nu_2| \geq \epsilon) \leq 2\Pr(|\nu_1 - p| \geq \frac{1}{2}\epsilon)$$

Proof: For the second part, note that

$$\Pr(|\nu_1 - \nu_2| \geq \epsilon) \leq \Pr(|\nu_1 - p| \geq \frac{\epsilon}{2} \vee |\nu_2 - p| \geq \frac{\epsilon}{2})$$

For the first part, if $|\nu_1 - p| \geq 2\epsilon, |\nu_2 - p| \leq \epsilon$, we will always have $|\nu_1 - \nu_2| \geq \epsilon$. Therefore,

$$\Pr(|\nu_1 - \nu_2| \geq \epsilon) \geq \Pr(|\nu_1 - p| \geq 2\epsilon)\Pr(|\nu_2 - p| \leq \epsilon)$$

. ■

Therefore, according to this lemma, we have:

$$\begin{aligned}
\Pr[\exists f \in \mathcal{F}, \Pr_D(Y \neq f(X)) - \frac{1}{n}\sum_{i=1}^{n} I[Y_i \neq f(X_i)] \geq \epsilon] \leq \\
2\Pr[\exists f \in \mathcal{F}, \frac{1}{n}\sum_{i=1}^{n} I[Y_i \neq f(X_i)] - \frac{1}{n}\sum_{i=n+1}^{2n} I[Y_i \neq f(X_i)] \geq \frac{\epsilon}{2}]
\end{aligned} \tag{4.3}$$

4.3.2 Step II: Sample and Permute

When drawing (x_i, y_i), we can follow these two steps: first draw an unordered set $z_1, ..., z_{2n}(z_i = (x_i, y_i))$ and second generate a random permutation $\sigma \in S_{2n}$ as the order. With this method, we have:

$$\begin{aligned} 2\Pr[\exists f \in \mathcal{F}, \frac{1}{n}\sum_{i=1}^{n} I[Y_i \neq f(X_i)] - \frac{1}{n}\sum_{i=n+1}^{2n} I[Y_i \neq f(X_i)] \geq \frac{\epsilon}{2}] = \\ 2\mathbb{E}_{(z_1,\ldots,z_{2n})}\{\Pr_{\sigma\in S_{2n}}[\exists f \in \mathcal{F}, \frac{1}{n}\sum_{i=1}^{n} I[Y_{\sigma(i)} \neq f(X_{\sigma(i)})] - \frac{1}{n}\sum_{i=n+1}^{2n} I[Y_{\sigma(i)} \neq f(X_{\sigma(i)})] \geq \frac{\epsilon}{2}]\} \end{aligned} \tag{4.4}$$

With the union bound, we have

$$\begin{aligned} &\Pr_{\sigma\in S_{2n}}[\exists f \in \mathcal{F}, \frac{1}{n}\sum_{i=1}^{n} I[Y_{\sigma(i)} \neq f(X_{\sigma(i)})] - \frac{1}{n}\sum_{i=n+1}^{2n} I[Y_{\sigma(i)} \neq f(X_{\sigma(i)})] \geq \frac{\epsilon}{2}] \\ &\leq N^F(z_1, z_2, \cdots, z_{2n}) \cdot \Pr_{\sigma\in S_{2n}}[\frac{1}{n}\sum_{i=1}^{n} I[Y_{\sigma(i)} \neq f(X_{\sigma(i)})] - \frac{1}{n}\sum_{i=n+1}^{2n} I[Y_{\sigma(i)} \neq f(X_{\sigma(i)})] \geq \frac{\epsilon}{2}] \end{aligned} \tag{4.5}$$

where $N^F(z_1, z_2, \cdots, z_{2n})$ denotes the number of distinguishable classifiers on $z_1, z_2, \cdots, z_{2n}$.

Now, with the *draw without replacement* Chernoff bound, we have

$$\begin{aligned} &\Pr_{\sigma\in S_{2n}}[\frac{1}{n}\sum_{i=1}^{n} I[Y_{\sigma(i)} \neq f(X_{\sigma(i)})] - \frac{1}{n}\sum_{i=n+1}^{2n} I[Y_{\sigma(i)} \neq f(X_{\sigma(i)})] \geq \frac{\epsilon}{2}] \\ &= \Pr_{\sigma\in S_{2n}}[\frac{1}{n}\sum_{i=1}^{n} I[Y_{\sigma(i)} \neq f(X_{\sigma(i)})] - \frac{1}{2n}\sum_{i=1}^{2n} I[Y_{\sigma(i)} \neq f(X_{\sigma(i)})] \geq \frac{\epsilon}{4}] \\ &\leq e^{-2n(\frac{\epsilon}{4})^2} \\ &= e^{-\frac{n\epsilon^2}{8}} \end{aligned} \tag{4.6}$$

Therefore,

$$\begin{aligned} &2\Pr[\exists f \in \mathcal{F}, \frac{1}{n}\sum_{i=1}^{n} I[Y_i \neq f(X_i)] - \frac{1}{n}\sum_{i=n+1}^{2n} I[Y_i \neq f(X_i)] \geq \frac{\epsilon}{2}] \\ &= 2\mathbb{E}_{(z_1,\ldots,z_{2n})}\{\Pr_{\sigma\in S_{2n}}[\exists f \in \mathcal{F}, \frac{1}{n}\sum_{i=1}^{n} I[Y_{\sigma(i)} \neq f(X_{\sigma(i)})] - \frac{1}{n}\sum_{i=n+1}^{2n} I[Y_{\sigma(i)} \neq f(X_{\sigma(i)})] \geq \frac{\epsilon}{2}]\} \\ &\leq 2\mathbb{E}_{(z_1,\ldots,z_{2n})}\{N^F(z_1, z_2, \cdots, z_{2n}) \cdot \Pr_{\sigma\in S_{2n}}[\frac{1}{n}\sum_{i=1}^{n} I[Y_{\sigma(i)} \neq f(X_{\sigma(i)})] - \frac{1}{n}\sum_{i=n+1}^{2n} I[Y_{\sigma(i)} \neq f(X_{\sigma(i)})] \geq \frac{\epsilon}{2}]\} \\ &\leq 2\mathbb{E}_{(z_1,\ldots,z_{2n})}\{N^F(z_1, z_2, \cdots, z_{2n}) \cdot e^{-\frac{n\epsilon^2}{8}}\} \\ &= 2e^{-\frac{n\epsilon^2}{8}} \cdot \mathbb{E}_{(z_1,\ldots,z_{2n})}\{N^F(z_1, z_2, \cdots, z_{2n})\} \end{aligned} \tag{4.7}$$

Denote $N^F(n) := \max_{(z_1,z_2,\cdots,z_n)} N^F(z_1, z_2, \cdots, z_n)$. Obviously $\mathbb{E}_{(z_1,\ldots,z_{2n})}\{N^F(z_1, z_2, \cdots, z_{2n})\} \leq N^F(2n)$.

Note that $N^F(n)$ is monotonically non-decreasing with respect to n, and $N^F(n) \leq 2^n$. Intuitively, with small values of n, the functions within F will be able to do arbitrary classification on some n data points $(z'_1, z'_2, \cdots, z'_n)$. When n grows past some threshold (say d), the expressiveness of F will fall short. So we speculate that

$$\exists d \in \mathbb{N}, N^F(n) \begin{cases} = 2^n, & n \leq d \\ < 2^n, & n > d \end{cases} \tag{4.8}$$

Our next step is to figure out the asymptotic characteristics of $N^F(n)$.

References

[AGM97] N. Alon, Z. Galil and O. Margalit, On the Exponent of the All Pairs Shortest Path Problem, *Journal of Computer and System Sciences* **54** (1997), pp. 255–262.

[F76] M. L. Fredman, New Bounds on the Complexity of the Shortest Path Problem, *SIAM Journal on Computing* **5** (1976), pp. 83-89.