Lecture 5: VC Theory - Generalization

Lecturer: Liwei Wang Scribe: Xinyu Deng, Yusen Wu(1806), Shiyuan Feng, Jingxuan Zheng, Yang He, Jiaxin Ge

5.1 Review

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

$$\begin{aligned}
&P(\exists f \in \mathcal{F} \;\; P_D(Y \neq f(x)) - \frac{1}{n}\Sigma_{i=1}^{n} 1[Y_i \neq f(x_i)] \geq \epsilon) \\
\leq &2P(\exists f \in \mathcal{F} \;\; \frac{1}{n}\Sigma_{i=1}^{n} 1[Y_i \neq f(x_i)] - \frac{1}{n}\Sigma_{i=n+1}^{2n} 1[Y_i \neq f(x_i)] \geq \frac{\epsilon}{2}) \quad \text{(Double Sample Trick)} \\
= &2E_{x_1 y_1, \cdots, x_n y_n}\{P_{\sigma \in S_in}(\exists f \in \mathcal{F} \;\; \frac{1}{n}\Sigma_{i=1}^{n} 1[Y_{\sigma(i)} \neq f(X_{\sigma(i)}] - \cdots \geq \frac{\epsilon}{2})\} \leq N(2n)c_1 e^{-c_2 n \epsilon^2} \quad \text{(Symmetrization)}
\end{aligned} \tag{5.1}$$

We define:

$$N^{\mathcal{F}}(x_1, y_1, \cdots, x_n, y_n) := |\{(f(x_1), f(x_2), \cdots, f(x_n)) : f \in \mathcal{F}\}|, \quad f(x_i) \in \{0, 1\} \tag{5.2}$$

$$N^{\Phi}(\delta_1, \cdots, \delta_n) := |\{(\phi_f(\delta_1), \cdots, \phi(\delta_n)) : \phi_f \in \Phi\}|, \text{ where } \delta_i = (x_i, y_i), \phi_f(\delta_i) = I[f(x_i) \neq y_i] \tag{5.3}$$

$$N^{\mathcal{F}}(n) := \max_{x_1, y_1, \ldots, x_n, y_n} N^{\mathcal{F}}(x_1, y_1, \ldots, x_n, y_n) \tag{5.4}$$

$$N^{\Phi}(n) = \max_{\delta_1, \cdots, \delta_n} N^{\Phi}(\delta_1, \ldots, \delta_n) \tag{5.5}$$

where Φ is a set of indicator functions. (Note that these two sets are same in size)
From the last lecture we have known that When n grows past some threshold (say d), the expressiveness of Φ will fall short. So we speculate that

$$N^{\Phi}(n) \begin{cases} = 2^n, & n \leq d \\ \leq \sum_{k=0}^{d} \binom{n}{k} = O(n^d), & n > d \end{cases} \tag{5.6}$$

$$|\{(\phi(y_1) \to n_1, \ldots, \phi(y_n) \to n_n) : \phi \in \Phi\}| \leq \sum_{k=0}^{d} \binom{n}{k} \tag{5.7}$$

We have to proof the following inequality:

$$N^{\Phi}(n) \leq \sum_{k=0}^{d} \binom{n}{k} \quad for \quad n > d \tag{5.8}$$

5.2 Proof for Inequality(5.8)

Let's consider the special case first— From our assumption, we know that when d+1, there is a case of

$$(\phi(x_1), \phi(x_2), ..., \phi(x_n)) \tag{5.9}$$

that cannot be obtained. In this special case, we assume that we cannot obtain d+1-zero cases. Which means that we can only have at most d zeros in this equation.
There are totally

$$\sum_{k=0}^{d} \binom{n}{k} \tag{5.10}$$

possible value assignments that have less than d+1 zeros.
Special cases are limited, so consider turning general cases into special cases. We will give the complete proof next.

Proof: First, we list all the situations that cannot be obtained and consider what happen at 1-st component. There are three possibility,

0 as 1-st component,eg:

$$\begin{cases} 0, *, 1\cdots & nbits \\ 0, 1, *\cdots \\ \cdots \\ 0, 0, *\cdots \end{cases}$$

1 as 1-st component, eg:

$$\begin{cases} 1, *, 1\cdots \\ 1, 1, *\cdots \\ \cdots \\ 1, 0, *\cdots \end{cases}$$

no restriction as 1-st component, eg:

$$\begin{cases} *, *, 1\cdots \\ *, 1, *\cdots \\ \cdots \\ *, 0, *\cdots \end{cases}$$

If we turn the 1 at 1-st component into 0, we will find that all possibilities are reduced.

Similarly, if we turn 1 into 0 at any component, we will find that all possibilities are reduced.

Therefore, if we turn 1 into 0 at all components, all possibilities will be reduced to a special case where only $d+1$ zeros cannot be obtained. So the possibilities of being able to obtain is more than that of the special case.

In this special case, the number of 0 can be 0, 1, $\cdots$, d. So we have $N^{\Phi}(n) \leq \sum_{k=0}^{d} \binom{n}{k}$ ■

$$\sum_{k=0}^{d} \binom{n}{k} \leq (\frac{en}{d})^d \tag{5.11}$$

Apply Chernoff bound and assume $d < \frac{n}{2}$

5.3 Step 3: VC Dimension

Definition 5.1 (VC Dimension) *The VC Dim of a set* Φ *of indicator function is the maximum* n*, so that* $N^{\Phi}(n) = 2^n$

Then, for any indicator function set Φ,if $VCD(\Phi) = d < \infty$ then

$$N^{\Phi}(n) \begin{cases} = 2^n, & n \le d \\ \le \sum_{k=0}^{d} \binom{n}{k} \le (\frac{en}{d})^d, & n > d \end{cases} \tag{5.12}$$

5.4 Step 1,2,3

$$\begin{aligned} P(\exists f \in \mathcal{F}: \; P_D(Y \neq f(x)) - \frac{1}{n} \sum I[Y_i \neq f(x_i)] \ge \epsilon) \\ \le N^{\Phi}(2n) \cdot c_1 e^{-c_2 n \epsilon^2} \\ \le (\frac{2en}{d})^d c_1 \cdot e^{-c_2 n \epsilon^2} \end{aligned} \tag{5.13}$$

Theorem 5.2 *Let* $\delta = (\frac{2en}{d})^d * 4e^{-\frac{1}{2}ne^2}$ *then we have with prob. at* $1-\delta$ *(over the random draw of the training dataset S).*

$$P_D(Y \neq f(X)) \le P_S[Y \neq f(X)] + O(\sqrt{\frac{d \ln n + \ln \frac{1}{\delta}}{n}}) \tag{5.14}$$

holds true for all $f \in \mathcal{F}$ *simultanously, where* d *is the VC dimension of the hypothesis space* $\mathcal{F}$*, where* $P_S[Y \neq f(X)] := \frac{1}{n} \sum I[Y_i \neq f(x_i)]$.

Linear classifiers in $\mathcal{R}^d$

$$\mathcal{F} := \{sgn(w^T x + b), w \in \mathcal{R}^d, b \in \mathcal{R}\} \tag{5.15}$$

then,

$$VCD(\mathcal{F}) = d + 1 \tag{5.16}$$

Proof: Let $x_1 = (1, 0, ..., 0), x_2 = (0, 1, ..., 0), x_d = (0, 0, ..., 1), x_{d+1} = (0, 0, ..., 0) \in \mathcal{R}^d$. They can represent any set in which d points are independent, cause $(x_1, ..., x_d)$ is a set of base in $\mathcal{R}^d$. Or we can discuss this Linear classification problem in $\mathcal{R}^{d-1}$. Then we have:

$$\begin{aligned} N^{\mathcal{F}}(x_1, ..., x_{d+1}) &= |\{(f(x_1), ..., f(x_{d+1}) | f \in \mathcal{F}\}| \\ &= |\{(sgn(w_1 + b), ..., sgn(w_d + b), sgn(b)) | w \in \mathcal{R}^d, b \in \mathcal{R}\}| \\ &= 2^{d+1} \end{aligned} \tag{5.17}$$

Thus $VCD(\mathcal{F}) \ge d + 1$. Next we need to prove $VCD(\mathcal{F}) < d + 2$:
$\forall x_1, ..., x_{d+2} \in \mathcal{R}^d, (x_1, 1), ..., (x_{d+2}, -1), \exists c_1, ..., c_{d+1}, s.t.$

$$w^T x_{d+2} + b = \sum_{i=1}^{d+1} c_i (w^T x_i + b), \forall w \in \mathcal{R}^d \tag{5.18}$$

Assuming that $(sgn(c_1), ..., sgn(c_{d+1}), -1) \in \{(f(x_1), ..., f(x_{d+2}) | f \in \mathcal{F}\}$. Then $\exists f \in \mathcal{F}$, that is $\exists w \in \mathcal{R}^d, b \in \mathcal{R}$, s.t. $sgn(c_i) = sgn(w^T x_i + b)$ and $sgn(w^T x_i + b) = -1$, which is contradict to 5.18.

Thus $(sgn(c_1), ..., sgn(c_{d+1}), -1) \notin \{(f(x_1), ..., f(x_{d+2}) | f \in \mathcal{F}\}, \forall f \in \mathcal{F}$, so $VCD(\mathcal{F}) < d + 2$. So we have proved $VCD(\mathcal{F}) = d + 1$

■

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.