ERM

1) Choose hypothesis space F

2) Learn $\hat{f} = \underset{f\in F}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} I[y_i \neq f(x_i)]$

If $|F| < \infty$ then $P\left\{\sup_{f\in F}\big(P_D(y \neq f(x)) - P_S(y \neq f(x)) \geq \varepsilon\big)\right\} \leq |F|e^{-2n\varepsilon^2}$

For $|F| = \infty$ we have three steps:

Step 1 Double Sample Trick

$$P\left(\sup_{\varphi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\varphi(z_i) - \frac{1}{n}\sum_{i=n+1}^{2n}\varphi(z_i)\right| \geq \varepsilon\right) \leq 2P\left(\sup_{\varphi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\varphi(z_i) - E[\varphi(z)]\right| \geq \frac{\varepsilon}{2}\right)$$

Step 2 Symmetrization

Fix $\{z^{(1)} \dots z^{(2n)}\}$

$$P_D\left(\sup_{\varphi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\varphi(z_i) - \frac{1}{n}\sum_{i=n+1}^{2n}\varphi(z_i)\right| \geq \varepsilon\right)$$
$$\leq N^{\Phi}(z_1 \dots z_n)P_D\left(\left|\frac{1}{n}\sum_{i=1}^{n}\varphi(z_i) - \frac{1}{n}\sum_{i=n+1}^{2n}\varphi(z_i)\right| \geq \varepsilon\right)$$

Where $N^{\Phi}(z_1 \dots z_n) \coloneqq |\{\varphi(z_1) \dots \varphi(z_n), \varphi \in \Phi\}| \leq 2^n$

$$E_{(z^{(1)} \dots z^{(n)})}P_{z_1 \dots z_n}\left(\sup_{\varphi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\varphi(z_i) - \frac{1}{n}\sum_{i=n+1}^{2n}\varphi(z_i)\right| \geq \varepsilon\right) \leq EN^{\Phi}(z_1 \dots z_n)e^{-O(n\varepsilon^2)}$$
$$\leq N^{\Phi}(2n)e^{-O(n\varepsilon^2)}$$

Step 3 VC-dimension

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

① $N^{\Phi}(n) \leq 2^n$ But we don't know how it grows actually, maybe exponential or polynomial

② when n is small $N^{\Phi}(n)$ can be exponential but when n is large it is relatively smaller

Thm

$$N^{\Phi}(n)\begin{cases} = 2^n & \text{if } n \leq d \\ \leq \sum_{k=0}^{d}\binom{n}{k} \leq \left(\frac{en}{d}\right)^d & \text{if } n > d \end{cases}$$

Phase change

(To prove it as homework)

So just to focus on what cases cannot be reached

Start with special case: n=d+1, if (0,0…) cannot be realized then $N^{\Phi}(n) = \sum_{k=0}^{d} \binom{n}{k}$

And intuitively it is the worst case so for others it is better

Proof Fix $z_1 \ldots z_n$

Note unrealizable patterns as $(*,*,0,1 \ldots \ldots)$ but they maybe intersected. So if I change the first bit from 1 to 0 otherwise keeps the same then the union of them are smaller. So when I replace 1 with 0 then the union is that they cannot have more than d+1 zeros, then it is the special case.

Then d is called the VC-dimension of the set. And d means that there exists z_1 to z_d s.t. Φ can reach every possible results but cannot find z_1 to z_{d+1}

So:

$$P\left(\sup_{\varphi\in\Phi}\left|\frac{1}{n}\sum_{i=1}^{n}\varphi(z_i) - E[\varphi(z)]\right| \geq \varepsilon\right) \leq e^{-O(n\varepsilon^2)}\left(\frac{en}{d}\right)^d$$

Thm' $\forall\delta > 0$ with prob $1-\delta$ over the random draw of training data (x_i, y_i)

$$P_D\big(y \neq f(x)\big) \leq P_S\big(y \neq f(x)\big) + O(\sqrt{\frac{d\ln(n) + \ln\left(\frac{1}{\delta}\right)}{n}}$$

Linear classifier: VC-dim = r+1 where $x \in R^r$

For ERM f^* is the best one and $\hat{f} = \underset{f\in F}{\operatorname{argmin}} P_D(y \neq f(x))$, then with probability $1-\delta$

$$P_D\left(y \neq \hat{f}(x)\right) \leq P_D\big(y \neq f^*(x)\big) + O(\sqrt{\frac{d\ln(n) + \ln\left(\frac{1}{\delta}\right)}{n}}$$

Lecture 4 Practical Algorithms

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

But indicator function is hard to minimize

So we change classification error from 0-1 loss to other functions.

1. Linear classification

$$x \in R^d, y \in \{\pm 1\}, f(x) = \operatorname{sign}(w^T x + b)$$

① $\max_{w,b,t} t$ s.t. $y_i\big(w^T x_i + b\big) \geq t$ and $\big||w|\big| = 1$

For time limits, if it is a convex optimization and constrain is linear then it is easy to be solved.