Lecturer: Prof. McDonald Scribe: Adithya Vadapalli

Introduction

In this lecture, first we will wrap up our discussion about VC-Dimension. Then we will give a brief introduction to Minimax theory.

22.1 VC - Wrap up

Suppose $f \in \mathcal{F}$. $f : X \to \{0, 1\}$. $\widehat{f}_{erm} = \text{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} I(f(x_i) \neq y_i)$ If VC dimension is finite , then with probability at least $1 - \delta$,

$$R(\widehat{f}_{erm}) \leq R(f^*) + c(\sqrt{V/n} + \sqrt{\log(1/\delta)/n})$$

For any estimator,

$$R(f) \leq R_n(\widehat{f}) + c(\sqrt{V/n} + \sqrt{\log(1/\delta)/n})$$

We have the following theorem,

Theorem 22.1. $\mathbb{P}(\sup_f |\widehat{R}(f) - R(f)| \geq \epsilon) \leq 8(n+1)^V e^{-n\epsilon^2/32}$

Note that, $(n+1)^V$ is called the shattering coefficient of f.

Corollary 22.2. $\mathbb{P}(|R(\widehat{f}_{erm}) - R(f^*| > \epsilon)) \leq 8(n+1)^V e^{-n\epsilon^2/128}$

Corollary 22.3. $\mathbb{E}[\sup_{f\in\mathcal{F}} |\widehat{R}(f) - R(f)|] = O(\sqrt{V \log n/n})$

Corollary 22.4. $\mathbb{E}[|R(\widehat{f}_{erm}) - R(f^*|] = O(\sqrt{V \log n/n})$

Proof. Write $Z_n = \sup_f |\widehat{R}_n(f) - R(f)|$.

And, $\mathbb{P}(Z_n > \epsilon) \leq c(n+1)^V e^{-n\epsilon^2 c} \implies \mathbb{P}(Z_n^2 > \epsilon^2) \leq c(n+1)^V e^{-n\epsilon^2 c}$. Substitute ϵ^2 with t. We get, $\mathbb{P}(Z_n^2 > t) \leq c(n+1)^V e^{-n\epsilon^2 c}$.

We have, $\mathbb{E}[Z_n^2] = \int_0^\infty \mathbb{P}(Z_n^2 > t)dt = \int_0^S \mathbb{P}(Z_n^2 > t)dt + \int_S^\infty \mathbb{P}(Z_n^2 > t)dt \leq S + \int_S^\infty \mathbb{P}(Z_n^2 > t) \leq S + c(n+1)^V \int_S^\infty \mathbb{P}(Z^2 > t)dt = S + c(n+1)^V(\frac{e^{-cnS}}{cn}) = V\log(n+1)/cn + c/n$ ■

Thus we have the following $\mathbb{E}[Z_n] \leq \sqrt{\mathbb{E}[Z_n^2]} \leq c\sqrt{V \log(n+1)/n}$

22.1.1 Tightness of the bounds

How tight are the above bound? Suppose $Y = f^*(x)$ (deterministic learning). $f^* \in \mathcal{F}$. It can be shown that (Vapnik, 1998) $\mathbb{E}[R\widehat{f}_{erm} - R(f^*)] = O(V/n)$.

The "noisiness" of $\mathbb{E}[Y|X=x] = \eta(X)$ is important.

Recall that, $f^*(x) = \begin{cases} 1 & \eta(X) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$

Definition 22.5. $\mathcal{P}(h) = \{p \in \mathcal{P} : |2\eta(x) - 1| \geq h\}$.

If $h = 0$, it is the noisy case. No constraints on $\mathcal{P}$. We will have, $O(\sqrt{V/n})$. $h = 1$ us noiseless. We will have $O(V/n)$.

Theorem 22.6. *Suppose* $f^* \in \mathcal{F}$, $VC(f) = V < \infty$. $\inf_{f \in \mathcal{F}} R(f) \geq c \min(\sqrt{V/n}, V/n)$

We have the following implications.

1. If $h = 0$, $\widehat{f}_{erm}$ is optimal.
2. If $h = 1$, $\widehat{f}_{erm}$ is optimal.

22.2 Statistical Minimax

We have $\theta \in \Theta$, estimate θ using data $(X_1, \cdots, X_n)$. Let $\widehat{\theta}$ be estimator. $\widehat{\theta} = \widehat{\theta}(X_n)$. We specify a loss function $L(\theta, \widehat{\theta}) \to \mathbb{R}^+$.

1. $||\theta - \widehat{\theta}||_2^2$ is squared loss.
2. $|\theta - \widehat{\theta}|$ is absolute loss.
3. $||\theta - \widehat{\theta}||_p^p$ is ℓ_p loss.
4. $I(\theta \neq \widehat{\theta})$ is 0-1 loss.

Definition 22.7. $R(\theta, \widehat{\theta}) = \mathbb{E}[L(\theta, \widehat{\theta})] = \int \cdots \int L(\theta, \widehat{\theta}(X^n)) dF(X^n)$

22.2.1 Example

Let $X \sim \mathcal{N}(0, 1)$, $\widehat{\theta}_1 = X$, $\widehat{\theta}_2 = 3$. Then, $R(\theta, \widehat{\theta}_1) = 1$ and $R(\theta, \widehat{\theta}_2) = (\theta - 3)^2$

Definition 22.8. *(Bayes' Risk)* $B_n(\widehat{\theta}) = \int R(\theta, \widehat{\theta}) \Pi(\theta) d\theta$

Definition 22.9. *The minimax estimator of* θ *is the one that satisfies* $\sup_\theta R(\theta, \widehat{\theta}) = \mathcal{R}_n(\theta)$.