Machine Learning Theory (CS 6783)

Lecture 10 : Fat-shattering dimension, Supervised Learnability

1 Recap

1. Covering : V is an ℓ_p-cover of $\mathcal{F}$ on $x_1, \ldots, x_n$ at scale β if

$$\forall f \in \mathcal{F}, \exists \mathbf{v} \in V \text{ s.t. } \left(\frac{1}{n}\sum_{t=1}^{n} |f(x_t) - \mathbf{v}[t]|^p\right)^{1/p} \leq \beta$$

$\mathcal{N}_p(\mathcal{F}, \beta; x_1, \ldots, x_n) = \min\{|V| : V \text{ is an } \ell_p\text{-cover of } \mathcal{F} \text{ on } x_1, \ldots, x_n \text{ at scale } \beta\}$

2.

$$\mathbb{E}_S\left[L_D(\hat{y}_{\text{erm}}) - \inf_{f\in\mathcal{F}} L_D(f)\right] \leq 2\mathbb{E}_S\left[\hat{\mathcal{R}}_S(\mathcal{F})\right] \leq 2\inf_{\beta>0}\left\{\beta + \sqrt{\frac{\log \mathcal{N}_1(\mathcal{F}, \beta; x_1, \ldots, x_n)}{n}}\right\}$$

3.

$$\hat{\mathcal{R}}_S(\mathcal{F}) \leq \hat{D}_S(\mathcal{F}) := \inf_{\alpha>0}\left\{4\alpha + 12\int_{\alpha}^{1}\sqrt{\frac{\log \mathcal{N}_2(\mathcal{F}, \beta; x_1, \ldots, x_n)}{n}}d\beta\right\}$$

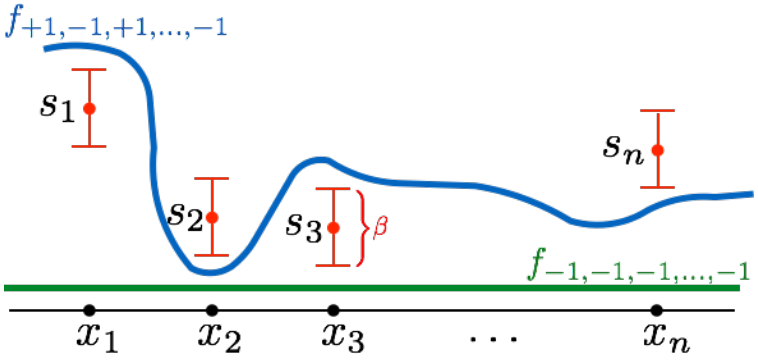
Also, $\hat{\mathcal{R}}_S(\mathcal{F}) \geq \tilde{\Omega}\left(\hat{D}_S(\mathcal{F})\right)$

2 Fat Shattering Dimension

Definition 1. *We say that $\mathcal{F}$ shatters $x_1, \ldots, x_n$ at scale γ, if there exists witness $s_1, \ldots, s_n$ such that, for every $\epsilon \in \{\pm 1\}^n$, there exists $f_\epsilon \in \mathcal{F}$ such that*

$$\forall t \in [n], \quad \epsilon_t \cdot (f_\epsilon(x_t) - s_t) \geq \gamma/2$$

Further $\quad \text{fat}_\gamma(\mathcal{F}) = \max\{n : \exists x_1, \ldots, x_n \in \mathcal{X} \text{ s.t. } \mathcal{F} \text{ } \gamma\text{-shatters } x_1, \ldots, x_n\}$



Theorem 1. *For any $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ and any $\gamma \in (0,1)$*

$$\mathcal{N}_2(\mathcal{F}, \gamma, n) \leq \left(\frac{2}{\gamma}\right)^{K \ \mathrm{fat}_{c\gamma}(\mathcal{F})}$$

where in the above c and K are universal constants.

Using the above with the dudley chaining bounds we get,

$$\mathcal{D}_S(\mathcal{F}) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{K \ \mathrm{fat}_{c\delta}(\mathcal{F}) \log(2/\delta)} d\delta \right\}$$

Thus bound on fat-shattering dimension leads to bound on Rademacher complexity.

Binary function class For any $\delta \in [0,1)$, and any $c \leq 1$, $\mathrm{fat}_{c\delta}(\mathcal{F}) = \mathrm{fat}_0(\mathcal{F}) = \mathrm{VC}(\mathcal{F})$ we can conclude that $\mathcal{V}_n^{\mathrm{stat}}(\mathcal{F}) \leq \mathcal{R}_n(\mathcal{F}) \leq \sqrt{\frac{\mathrm{VC}(\mathcal{F})}{n}}$.

Linear Predictors Let $\mathcal{X} = \{x : \|x\|_2 \leq 1\}$ and let $\mathcal{F} = \{x \mapsto f^\top x : \|f\|_2 \leq 1\}$.
1. $\mathrm{fat}_\gamma(\mathcal{F}) \geq \lfloor 4\gamma^{-2} \rfloor$ **:**
For all $i \in [d]$, let $x_i = e_i$ and let $s_i = 0$. Given $\epsilon \in \{\pm 1\}^d$, consider the vector f such that $f[i] = \epsilon_i \gamma/2$. Clearly f, γ-shatters these set of d points. Now for $\|f\|_2 \leq 1$, we need that $\sum_{i=1}^{d} f^2[i] = d\gamma^2/4 \leq 1$. This implies that $d \leq 4\gamma^{-2}$. Thus we can provide $4/\gamma^2$ points that can be γ-shattered.

2. $\mathrm{fat}_\gamma(\mathcal{F}) \leq 4\gamma^{-2}$ **:**
Typically uses Maurey's theorem but we will take a different route in just a bit.

2.1 Back to Rademacher

Claim 2.

$$\mathcal{R}_n(\mathcal{F}) \geq \sup\{\gamma/2 : \mathrm{fat}_\gamma(\mathcal{F}) > n\}$$

Proof. Think about Rademacher complexity on shattered points. □

The claim above is the same as saying (converse) $\mathrm{fat}_\gamma \leq \min\{n : \mathcal{R}_n(\mathcal{F}) \leq \gamma/2\}$. Using this for linear class example, since we know that $\mathcal{R}_n(\mathcal{F}) \leq \frac{1}{\sqrt{n}}$, we can conclude that for the linear class, $\mathrm{fat}_\gamma \leq \min\{n : \mathcal{R}_n(\mathcal{F}) \leq \gamma/2\} \leq \min\{n : \frac{1}{\sqrt{n}} \leq \gamma/2\} \leq \lceil \frac{4}{\gamma^2} \rceil$.

Using a more refined argument, the claim above can be improved, it can be shown that for any $\gamma > \mathcal{R}_n(\mathcal{F})$,

$$\mathrm{fat}_\gamma(\mathcal{F}) \leq \frac{8n\mathcal{R}_n^2(\mathcal{F})}{\gamma^2}$$

from this we can conclude that

$$\hat{\mathcal{R}}_S(\mathcal{F}) \geq \tilde{\Omega}\left(\inf_{\alpha \geq 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{K \ \mathrm{fat}_{c\delta}(\mathcal{F}) \log(2/\delta)} d\delta \right\} \right)$$