Lecture 9 Rademacher complexity

IEMS 402 Statistical Learning

Northwestern

Ref

https://www.cs.cmu.edu/~ninamf/ML11/lect1117.pdf

Chaining

Hidden Assumption

Dudley's Theorem

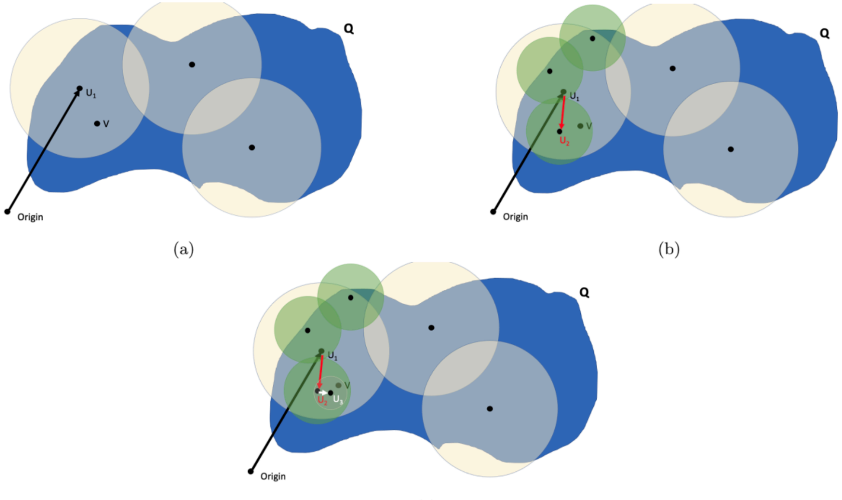
Theorem 3.1. Dudley:

$$\hat{R}(F) \leq 12 \int_0^\infty \frac{logN(\epsilon, F, L_2(P_n))}{n} d\epsilon$$

Chaining

The Chaining idea is to rewrite f as follows:

$$f = f + \sum_{j=1}^{N} (\hat{f}_j - \hat{f}_{j-1}) + \cancel{\hat{f}_0} - \hat{f}_N.$$



(a) (b) (c)

Example

Example. F = the non-decreasing function from $\mathbb{R}$ to $[0, 1]$.

We can actually cover such a function uniformly. We only need to approximate it at n points, marked in the figure. If it is within α at each of these points then the L_2 distance will be no more than α. From the approximating points one can produce a non-decreasing function: for each of the α-levels (of which there will be $1/\alpha$), just specify one of the n points at which it increases above that level. From this we can (loosely, but to the right order of magnitude) upper bound the size of the class of estimate functions: $|\hat{F}| \leq n^{1/\alpha}$.

We see that we can cover F in L_2:

$$N(\alpha, F, L_2(P_n)) \leq Cn^{1/\alpha}.$$

1. The Discretization Theorem gives

$$\hat{R}_n(F) \leq c \left(\frac{\log n}{n} \right)^{1/3}$$

2. The Chaining Theorem gives

$$\hat{R}_n(F) \leq 12 \int_0^1 \sqrt{\frac{\log n}{\alpha n}} d\alpha = 12\sqrt{\frac{\log n}{n}} \int_0^1 \sqrt{\frac{1}{\alpha}} d\alpha = 24\sqrt{\frac{\log n}{n}}$$

Chaining

Rademacher Complexity

Rademacher Complexity

Definition. The *empirical Rademacher complexity* of $\mathcal{F}$ is defined to be

$$\hat{R}_m(\mathcal{F}) = \mathsf{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\right)\right]$$

where $\sigma_1, \ldots, \sigma_m$ are independent random variables uniformly chosen from $\{-1, 1\}$. We will refer to such random variables as *Rademacher variables*.

Rademacher Complexity

Theorem 2. *Fix distribution $D|_Z$ and parameter $\delta \in (0,1)$. If $\mathcal{F} \subseteq \{f : Z \to [a, a+1]\}$ and $S = \{z_1, \ldots, z_n\}$ is drawn i.i.d. from $D|_Z$ then with probability $\geq 1 - \delta$ over the draw of S, for every function $f \in \mathcal{F}$,*

$$\mathsf{E}_D[f(z)] \leq \hat{\mathsf{E}}_S[f(z)] + 2R_m(\mathcal{F}) + \sqrt{\frac{\ln(1/\delta)}{m}}. \tag{1}$$

In addition, with probability $\geq 1 - \delta$, for every function $f \in \mathcal{F}$,

$$\mathsf{E}_D[f(z)] \leq \hat{\mathsf{E}}_S[f(z)] + 2\hat{R}_m(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{m}}. \tag{2}$$

Tighter than Covering number

Theorem 4. *For any* $A \subseteq \mathbb{R}^m$, *let* $R = \sup_{a \in A} \left(\sum_{i=1}^m a_i^2\right)^{1/2}$. *Then*

$$\hat{R}_m(A) = \mathsf{E}_\sigma \left[\sup_{a \in A} \left(\frac{1}{m} \sum_{i=1}^m \sigma_i a_i \right) \right] \leq \frac{R\sqrt{2 \ln |A|}}{m}$$

Massart's Finite Lemma

Property

For a class of functions F, let $co(F)$ represents its convex hull,

$$co(F) := \left\{ \sum_{i=1}^{k} \alpha_i f_i : k \geq 1,\ \alpha_i \geq 0,\ \|\alpha\|_1 = 1,\ f_i \in F \right\}.$$

Then we have: $R_n(F) = R_n(co(F))$. Based on the definition:

Property

$R_n(F+g) = R_n(F)$, where $F+g$ is defined as $\{x \mapsto f(x)+g(x) \,:\, f \in F\}$.

Property

Ledoux-Talagrand contraction inequality: If $\phi_i : \mathbb{R} \to \mathbb{R}$ satisfies $|\phi_i(a) - \phi_i(b)| \le L|a-b|$), then

$$\mathbb{E} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \phi_i(f(x_i)) \le L \cdot \mathbb{E} \sup_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i)$$