Lecture Notes 10
36-705

Let $\mathcal{F}$ be a set of functions and recall that

$$\Delta_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}[f] \right|.$$

Let us also recall the Rademacher complexity measures

$$\mathcal{R}(x_1, \ldots, x_n) = \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(x_i) \right| \right]$$

and

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_\epsilon \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right| \right]$$

where $\epsilon_1, \ldots, \epsilon_n$ are Rademacher random variables. We have already shown that

$$\mathbb{E}[\Delta_n(\mathcal{F})] \leq 2\mathcal{R}_n(\mathcal{F}).$$

1 Rademacher Complexity of a Finite Class

Suppose that we have a finite collection of functions $\mathcal{F} = \{f_1, \ldots, f_N\}$, which are bounded i.e. $\|f_i\|_\infty \leq b$ then we have the following bound on the Rademacher complexity.

Finite Class Bound: The Rademacher complexity for a finite class,

$$\mathcal{R}(\mathcal{F}) \leq 2b\sqrt{\frac{\log(2N)}{n}}.$$

Proof: Define,

$$\Theta := \mathbb{E}_{X,\epsilon} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right| \right].$$

For convenience let us augment the class $\mathcal{F}$ with the negative of every function, i.e. we take $\widetilde{\mathcal{F}} = \mathcal{F} \cup (-\mathcal{F})$, so that there are now $2N$ functions. Then,

$$\Theta \leq \mathbb{E}_{X,\epsilon} \left[\sup_{f \in \widetilde{\mathcal{F}}} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(X_i) \right].$$

Note that,

$$\begin{aligned}\exp(t\Theta) &\leq \exp\left(t\mathbb{E}_{X,\epsilon}\left[\sup_{f\in\widetilde{\mathcal{F}}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right]\right) \\ &\leq \mathbb{E}_{X,\epsilon}\exp\left(t\left[\sup_{f\in\widetilde{\mathcal{F}}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right]\right) \\ &\leq \mathbb{E}_{X,\epsilon}\sum_{j=1}^{2N}\prod_{i=1}^{n}\exp\left(\frac{t\epsilon_i f_j(X_i)}{n}\right) \\ &= \sum_{j=1}^{2N}\prod_{i=1}^{n}\mathbb{E}_{X,\epsilon}\exp\left(\frac{t\epsilon_i f_j(X_i)}{n}\right).\end{aligned}$$

Since $\|f_j\|_\infty \leq b$ we can use the argument we used in the proof of Hoeffding's inequality to obtain that,

$$\exp(t\Theta) \leq 2N\exp\left(\frac{4t^2b^2}{n}\right),$$

so we obtain that,

$$\Theta \leq \frac{\log(2N)}{t} + \frac{4tb^2}{n},$$

where t is a free parameter that is > 0. Choosing, $t = \sqrt{n\log(2N)/(4b^2)}$ we obtain,

$$\Theta \leq 2b\sqrt{\frac{\log(2N)}{n}}.$$

2 Using the Rademacher Theorem to obtain the VC theorem

The Rademacher theorem in a very straightforward way implies the VC theorem. We'll sketch the proof here. Our class of functions just corresponds to the indicators arising from the set system. These functions are upper bounded by $b = 1$. We can get a high-probability statement as in the initial section so we only need to deal with $\mathcal{R}(\mathcal{F})$.

We follow an identical argument to the one we did in the previous section,

$$\exp(t\Theta) \leq \mathbb{E}_{X,\epsilon}\exp\left(t\left[\sup_{f\in\widetilde{\mathcal{F}}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right]\right),$$

where the class $\widetilde{\mathcal{F}}$ just contains the set indicators and their negations.

The key point is to note here is the following: suppose we think of the vectors $(f(X_1), \ldots, f(X_n))$ for each function in $\widetilde{\mathcal{F}}$ and ask how many different such vectors are there? Each set in $\mathcal{A}$ picks out some subset of the points (and assigns them the value $+1$). Even though there are possibly infinitely many sets in $\mathcal{A}$ there are at most only twice (because we included the negations) the shattering number of different vectors.

The shattering number is precisely the (maximum) number of different vectors $(f(X_1), \ldots, f(X_n))$ we can induce using our collection of sets.

With this insight in hand we can just repeat the previous argument to conclude that,

$$\Theta \leq \sqrt{\frac{4\log(2s(\mathcal{A}, n))}{n}},$$

and putting this together with the high-probability bound from before we have that with probability at least $1 - \delta$,

$$\begin{aligned}\Theta &\leq \sqrt{\frac{4\log(2s(\mathcal{A}, n))}{n}} + \sqrt{\frac{2\ln(2/\delta)}{n}} \\ &\leq \sqrt{\frac{4\log(4s(\mathcal{A}, n)/\delta)}{n}},\end{aligned}$$

which is precisely the VC theorem (again always ignore constants).