Mathematical Foundations of Machine Learning(CS 4783/5783)

Lecture 4: Rademacher Complexity, Binary Classification, Growth Function and VC dimension

1 Recap

In the previous lecture notes, we showed the following corollary.

Corollary 1. *For any class $\mathcal{F}$ and any loss bounded by* 1*, for any $\delta > 0$, with probability at least $1 - \delta$,*

$$L_{\mathbf{D}}(\hat{f}_{\mathrm{ERM}}) - \min_{f \in \mathcal{F}} L_{\mathbf{D}}(f) \leq O\left(\sqrt{\frac{\log |\mathcal{F}_{|x_1,\dots,x_n}|}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$$

Hence, it is clear that bounding $|\mathcal{F}_{|x_1,\dots,x_n}|$ yields a bound on the performance of ERM. In this lecture we will restrict ourself to binary classification problem where $\mathcal{F} \subset \{\pm 1\}^{\mathcal{X}}$ and loss $\ell(y', y) = \mathbb{1}_{\{y' \neq y\}}$. Under this setting we consider the the worst case $|\mathcal{F}_{|x_1,\dots,x_n}|$ which we next introduce as growth function.

2 Growth Function and VC dimension

Growth function is defined as,

$$\Pi(\mathcal{F}, n) = \max_{x_1,\dots,x_n} \left|\mathcal{F}_{|x_1,\dots,x_n}\right|$$

That is, for the worst $x_1, \dots, x_n$, how many possible labelings does $\mathcal{F}$ make on $x_1, \dots, x_n$. From the result in the previous lecture, we have already seen that the Rademacher Complexity is bounded by $O\left(\sqrt{\frac{\log |\mathcal{F}_{|x_1,\dots,x_n}|}{n}}\right)$ and so overall, for any $\delta > 0$, with probability at least $1 - \delta$,

$$L_{\mathbf{D}}(\hat{f}_{\mathrm{ERM}}) - \min_{f \in \mathcal{F}} L_{\mathbf{D}}(f) \leq O\left(\sqrt{\frac{2 \log \Pi(\mathcal{F}, n)}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$$

Note that $\Pi(\mathcal{F}, n)$ is at most 2^n but it could be much smaller. In general how do we get a handle on growth function for a hypothesis class $\mathcal{F}$? Is there a generic characterization of growth function of a hypothesis class ?

Definition 1. *VC dimension of a binary function class $\mathcal{F}$ is the largest number of points $d = \mathrm{VC}(\mathcal{F})$, such that*

$$\Pi_{\mathcal{F}}(d) = 2^d$$

If no such d exists then $\mathrm{VC}(\mathcal{F}) = \infty$

If for any set $\{x_1, \dots, x_n\}$ we have that $|\mathcal{F}_{|x_1,\dots,x_n}| = 2^n$ then we say that such a set is shattered. Alternatively VC dimension is the size of the largest set that can be shattered by $\mathcal{F}$.

Eg. Thresholds One point can be shattered, but two points cannot be shattered. Hence VC dimension is 1. (If we allow both threshold to right and left, VC dimension is 2).

Eg. Spheres Centered at Origin in d dimensions one point can be shattered. But even two can't be shattered. VC dimension is 1!

Eg. Half-spaces Consider the hypothesis class where all points to the left (or right) of a hyperplane in $\mathbb{R}^d$ are marked positive and the rest negative. In 1 dimension this is threshold both to left and right. VC dimension is 2. In d dimensions, think of why $d+1$ points can be shattered. $d+2$ points can't be shattered. Hence VC dimension is $d+1$.

Claim 2. *If a class of models $\mathcal{F}$ has* $\mathrm{VC}(\mathcal{F}) = \infty$*, then for any n, and any learning algorithm, there is a distribution over instances such that with probability at least* $1/2$*:*

$$L_{\mathbf{D}}(\hat{f}_S) - \inf_{f \in \mathcal{F}} L_{\mathbf{D}}(f) \geq \frac{1}{2}$$

where $\hat{f}_S$ is the model returned by any learning algorithm that uses training sample S of size n.

Proof. Assume that a hypothesis class $\mathcal{F}$ has infinite VC dimension. This means that for any n, we can find $2n$ points $x_1, \ldots, x_{2n}$ that are shattered by $\mathcal{F}$. Now draw $y_1, \ldots, y_{2n} \in \{\pm 1\}$ as Rademacher random variables. Let D be the uniform distribution over the $2n$ instance pairs $(x_1, y_1), \ldots, (x_{2n}, y_{2n})$. Notice that since $x_1, \ldots, x_{2n}$ are shattered by $\mathcal{F}$, we are in the realizable PAC setting for any choice of y's. Now assume we get n input instances drawn iid from this distribution. Clearly in this sample of size n, we can at most witness n unique instances. Let us denote $J \subset [2n]$ as the indices of the $2n$ instances witnessed in the draw of n samples S. Clearly $|J| \leq n$. Now given any sample S, let $\hat{f}_S$ be the model returned by a learning algorithm. Notice first that for any model g, under this above distribution:

$$L_{\mathbf{D}}(g) = \frac{1}{2n} \sum_{i=1}^{2n} \mathbb{1}_{\{g(x_i) \neq y_i\}}$$

Also, as mentioned earlier, since $x_1, \ldots, x_{2n}$ is shattered, $\min_{f \in \mathcal{F}} L_{\mathbf{D}}(f) = 0$. Hence,

$$\begin{aligned} L_{\mathbf{D}}(\hat{f}_S) - \min_{f \in \mathcal{F}} L_{\mathbf{D}}(f) &= \frac{1}{2n} \sum_{j=1}^{2n} \mathbb{1}_{\{\hat{f}_S(x_i) \neq y_i\}} - 0 \\ &= \frac{1}{2n} \left(\sum_{i \in J} \mathbb{1}_{\{\hat{f}_S(x_i) \neq y_i\}} + \sum_{i \in [2n] \setminus J} \mathbb{1}_{\{\hat{f}_S(x_i) \neq y_i\}} \right) \\ &\geq \frac{1}{2n} \sum_{i \in [2n] \setminus J} \mathbb{1}_{\{\hat{f}_S(x_i) \neq y_i\}} \end{aligned}$$

Now notice that $y_1, \ldots, y_{2n}$ are chosen as random coin flips. Further, sample S only reveals labels of y_i's for any $i \in J$. On indices $[2n] \setminus J$, the random labels cannot be predicted by any algorithm with accuracy better than 1/2. That is, for any $i \in [2n] \setminus J$, $\mathbb{1}_{\{\hat{f}_S(x_i) \neq y_i\}}$ is 1 with probability 1/2 and 0 with probability 1/2. Hence, we can interpret $\sum_{i\in[2n]\setminus J} \mathbb{1}_{\{\hat{f}_S(x_i)\neq y_i\}}$ as the number of 1's we get when we flip $|[2n]\setminus J|$ fair coins. This follows the binomial distribution and we can conclude easily that with probability 1/2, $\sum_{i\in[2n]\setminus J} \mathbb{1}_{\{\hat{f}_S(x_i)\neq y_i\}} \geq \frac{|[2n]\setminus J|}{2}$. However, since $|J|$ is at most, n, we can conclude that with probability at least 1/2 over draws of y's:

$$L_{\mathbf{D}}(\hat{f}_S) - \min_{f\in\mathcal{F}} L_{\mathbf{D}}(f) \geq \frac{n}{2n} = \frac{1}{2}$$

In fact, from the above its also clear that for any constant $c \in (0,1)$, we can conclude a lower bound on excess risk that holds with probability at least c where the lower bound only depends on c. □

The above claim shows that VC dimension being finite is at least necessary if we would like to obtain bounds for binary classification problem for which excess risk goes to 0 with number of samples n. On the other hand, the following celebrated lemma due to Vapnik and Chervonenkis (English version published in 1971 and Russian version in 1968) and also independently shown by Sauer and also by Shelah shows that when VC dimension is finite, then growth function only grows as $n^{\mathrm{VC}(\mathcal{F})}$. This result along with the bound on performance of ERM implies that if VC dimension is indeed finite than ERM learns at a rate of $O\left(\sqrt{\frac{\mathrm{VC}(\mathcal{F})}{n}}\right)$, thus showing that finite VC dimension is also sufficient to get error bounds.

Lemma 3 (VC'71/Sauer'72/Shelah'72)**.** *For any class $\mathcal{F} \subset \{\pm1\}^{\mathcal{X}}$ with $\mathrm{VC}(\mathcal{F}) = d$, we have that,*

$$\Pi(\mathcal{F}, n) \leq \sum_{i=0}^{d} \binom{n}{i}$$

Proof of the above lemma is done via induction on $n + d$. Also note that $\sum_{i=0}^{d} \binom{n}{i} \leq n^d$

Corollary 4. *For any class of models $\mathcal{F} \subset \{\pm1\}^{\mathcal{X}}$, for any $\delta > 0$, with probability at least $1-\delta$,*

$$L_{\mathbf{D}}(\hat{f}_{\mathrm{ERM}}) - \min_{f\in\mathcal{F}} L_{\mathbf{D}}(f) \leq O\left(\sqrt{\frac{\mathrm{VC}(\mathcal{F})\log n}{n}} + \sqrt{\frac{\log(1/\delta)}{n}}\right)$$