Machine Learning Theory (CS 6783)

Lecture 6 : Effective size, VC Dimension, Learnability and VC/Sauer/Shelah Lemma

1 Recap

1. For the ERM we have,

$$\mathbb{E}_S\left[L_D(\hat{f}_{ERM}) - \inf_{f\in\mathcal{F}} L_D(f)\right] \le \frac{2}{n}\mathbb{E}_S\left[\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t)\right]\right]$$

   RHS above is the so called Rademacher complexity of the loss composed with function class $\mathcal{F}$

2. This is useful because conditioned on data, we can get bounds that depend on effective size of $\mathcal{F}$ on data $x_1, \ldots, x_n$.

3. Eg. threshold is learnable and effective size on n points is at most $n+1$ but $\mathcal{F}$ is uncountably infinite

2 Infinite $\mathcal{F}$: Binary Classes and Growth Function

First let us simplify the Rademacher complexity for binary classification problem. Note that for binary classification problem where $\mathcal{Y} \in \{\pm 1\}$, the loss can be rewritten as $\ell(y', y) = \mathbb{1}_{\{y \neq y'\}} = \frac{1-y\cdot y'}{2}$. Hence

$$\begin{aligned}2\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t)\right\}\right] &= 2\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^n \epsilon_t \frac{1 - f(x_t)\cdot y_t}{2}\right\}\right] \\ &= \mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^n \epsilon_t y_t f(x_t)\right]\end{aligned}$$

Now consider the inner term in the expectation above, ie. $\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^n \epsilon_t y_t f(x_t)\right]$. Note that given any fixed choice of $y_1, \ldots, y_n \in \{\pm 1\}$, $\epsilon_1 y_1, \ldots, \epsilon_n y_n$ are also Rademahcer random variables. Hence for the binary classification problem,

$$2\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\left\{\frac{1}{n}\sum_{t=1}^n \epsilon_t \ell(f(x_t), y_t)\right\}\right] = \mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^n \epsilon_t f(x_t)\right]$$

In the above statement we moved from Rademacher complexity of loss class $\ell \circ \mathcal{F}$ to the Rademacher complexity of the function class $\mathcal{F}$ for binary classification task. This is a precursor to what we will refer to as contraction lemma which we will show later.

Why is symmetrization useful? Think what we gain for an infinite class ...

3 Effective size of function class on Data

Why is the introduction of Rademacher averages important ? To analyze the term, $\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^n \epsilon_t \ell(f(x_t),y_t)\right]$ consider the inner expectation, that is conditioned on sample consider the term $\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^n \epsilon_t \ell(f(x_t),y_t)\right]$. Note that $\frac{1}{n}\sum_{t=1}^n \epsilon_t \ell(f(x_t),y_t)$ is still average of 0 mean random variables (conditioned on data) and we can apply Hoeffding bound for each fixed $f\in\mathcal{F}$ individually. Now $\mathcal{F}$ might be an infinite class, but, conditioned on input instances $(x_1,y_1),\ldots,(x_n,y_n)$, one can ask, what is the size of the projection set

$$\mathcal{F}_{|x_1,\ldots,x_n} = \{f(x_1),\ldots,f(x_n) : f\in\mathcal{F}\}$$

For any binary class $\mathcal{F}$, first note that this set can have a maximum cardinality of 2^n however it could be much smaller. In fact we can have,

$$\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{t=1}^n \epsilon_t\ell(f(x_t),y_t)\right] = \mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{\mathbf{f}\in\mathcal{F}_{|x_1,\ldots,x_n}}\frac{1}{n}\sum_{t=1}^n \epsilon_t\ell(\mathbf{f}[t],y_t)\right] \le \mathbb{E}_S\left[\sqrt{\frac{\log|\mathcal{F}_{|x_1,\ldots,x_n}|}{n}}\right]$$

where the last step is using the finite Lemma. Now one can define the growth function for a hypothesis class $\mathcal{F}$ as follows.

$$\Pi_{\mathcal{F}}(\mathcal{F},n) = \sup\{|\mathcal{F}_{|x_1,\ldots,x_n}| : x_1,\ldots,x_n\in\mathcal{X}\}$$

Example : thresholds
What does the growth function of the class of threshold function look like ?
Well sort any given n points in ascending order, using thresholds, we can get at most $n+1$ possible labeling on the n points. Hence $\Pi_{\mathcal{F}}(n) = n+1$. From this we conclude that for the learning thresholds problem,

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \le \sqrt{\frac{\log(n)}{n}}$$

4 Growth Function and VC dimension

Growth function is defined as,

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

Clearly we have from the previous results on bounding minimax rates for statistical learning in terms of cardinality of growth function that :

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \le \sqrt{\frac{2\log\Pi(\mathcal{F},n)}{n}}$$

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 $\{x_1, \ldots, x_n\}$ we have that $|\mathcal{F}_{|x_1,\ldots,x_n}| = 2^n$ then we say that such a set is shattered. Alternatively VC dimension is the size of the largest set that can be shattered by $\mathcal{F}$. We also define VC dimension of a class $\mathcal{F}$ restricted to instances $x_1, \ldots, x_n$ as

$$\mathrm{VC}(\mathcal{F}; x_1, \ldots, x_n) = \max\left\{t : \exists i_1, \ldots, i_t \in [n] \text{ s.t. } \left|\mathcal{F}_{|x_{i_1},\ldots,x_{i_n}}\right| = 2^t\right\}$$

That is the size of the largest shattered subset of n. Note that for any $n \geq \mathrm{VC}(\mathcal{F})$, $\sup_{x_1,\ldots,x_n} \mathrm{VC}(\mathcal{F}_{|x_1,\ldots,x_n}) = \mathrm{VC}(\mathcal{F})$.

1. To show $\mathrm{VC}(\mathcal{F}) \geq d$ show that you can at least pick d points $x_1, \ldots, x_d$ that can be shattered.
2. To show that $\mathrm{VC}(\mathcal{F}) \leq d$ show that no configuration of $d + 1$ points can be shattered.

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 1. *VC dimension of half-spaces in $\mathbb{R}^d$ is $d + 1$*

Proof. We consider half-spaces that map vector in $\mathbb{R}^d$ to $\{\pm 1\}$. That is

$$\mathcal{F} = \{\mathbf{x} \mapsto \mathrm{sign}\left(\mathbf{f}^\top \mathbf{x} + f_0\right) : \mathbf{f} \in \mathbb{R}^d, f_0 \in \mathbb{R}\}$$

We prove the statement as follows.

1. $\mathrm{VC}(\mathcal{F}) \geq d + 1$:
   We can shatter the points $\mathbf{e}_1, \ldots, \mathbf{e}_d, \mathbf{0}$. To see this, note that given any $y_1, \ldots, y_{d+1} \in \{\pm 1\}^{d+1}$, if we consider $f \in \mathcal{F}$ given by $f_0 = y_{d+1}$ and for all $i \in [d]$, $\mathbf{f}[i] = y_i - y_{d+1}$. Hence note that, $f(\mathbf{0}) = \mathrm{sign}\left(\mathbf{f}^\top \mathbf{0} + f_0\right) = \mathrm{sign}(y_{d+1}) = y_{d+1}$. Also, for any $i \in [d]$, $f(\mathbf{e}_i) = \mathrm{sign}\left(\mathbf{f}^\top \mathbf{e}_i + f_0\right) = \mathrm{sign}\left(y_i - y_{d+1} + y_{d+1}\right) = y_i$.

2. $\mathrm{VC}(\mathcal{F}) < d + 2$:
   By Radon theorem, any set of $d + 2$ points in $\mathbb{R}^d$ can be partitioned into two disjoint subsets whose convex hulls have a non-empty intersection. Label one of these partitions $+1$ and other -1. No half-space can successfully label points in the intersection.

$\square$

Claim 2. *Learnability with binary hypothesis class $\mathcal{F}$ implies* $\mathrm{VC}(\mathcal{F}) < \infty$.

Proof. First note that learnability in the statistical learning framework implies learnability in the realizable PAC setting. Hence to prove the claim, it suffices to show that if a hypothesis class has infinite VC dimension, then it is not even learnable in the realizable PAC setting. To this end, 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 $\mathcal{F}$. That is, on points $x_1, \ldots, x_{2n}$, effectively the function class $\mathcal{F}$ can take all possible labels or in other words, if we restrict input space to just these $2n$ points $x_1, \ldots, x_{2n}$ then $\mathcal{F}$ on this input space is same as $\{\pm 1\}^{\{x_1,\ldots,x_{2n}\}}$ the set of all possible functions. Hence using the no free lunch theorem, restricting ourselves to this set of $2n$ points we can conclude that

$$\mathcal{V}_n^{\mathrm{PAC}}(\mathcal{F}) \geq \frac{1}{4}$$

□

Lemma 3 (VC'71 (originially 64!)/Sauer'72/Shelah'72)**.** *For any class $\mathcal{F} \subset \{\pm 1\}^{\mathcal{X}}$ with $\mathrm{VC}(\mathcal{F}) = d$, we have that,*

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

Proof. For notational ease let $g(d, n) = \sum_{i=0}^{d} \binom{n}{i}$. We want to prove that $\Pi(\mathcal{F}, n) \leq g(d, n) = g(d, n-1) + g(d-1, n-1)$. We prove this one by induction on $n + d$.

Base case : We need to consider two base cases. First, note that when VC dimension $d = 0$, then clearly for any $x, x' \in \mathcal{X}$, $f(x) = f(x')$ and so we can conclude that for such a class $\mathcal{F}$ effectively contains only one function and so $\Pi(\mathcal{F}, n) = g(0, n) = 1$. On the other hand, note that for any $d \geq 1$, if VC dimension of the function class $\mathcal{F}$ is d then it can at least shatter 1 point and so $\Pi(\mathcal{F}, 1) = g(d, 1) = 2$. These form our base case.

Induction : Assume that the statement holds for any class $\mathcal{F}$ with VC dimension $d' \leq d$ and any $n' \leq n - 1$ that $\Pi(\mathcal{F}, n') \leq g(d', n')$. We shall prove the that in this case, for any $\mathcal{F}$ with VC dimension $d' \leq d$, $\Pi(\mathcal{F}, n) \leq g(d', n)$ and similarly for any $n' \leq n$, and for any $\mathcal{F}$ with VC dimension at most $d + 1$, $\Pi(\mathcal{F}, n') \leq g(d + 1, n')$.

To this end, consider any class $\mathcal{F}$ of VC dimension at most d' and consider any set of n instances $x_1, \ldots, x_n$. Define hypothesis class

$$\tilde{\mathcal{F}} = \left\{ f \in \mathcal{F} : \exists f' \in \mathcal{F} \text{ s.t. } f(x_n) \neq f'(x_n),\ \forall i < n,\ f(x_i) = f'(x_i) \right\}$$

That is the hypothesis class consisting of all functions that have a pair with same exact value of $x_1, \ldots, x_{n-1}$ but opposite sign only on x_n. We first claim that,

$$\left|\mathcal{F}_{|x_1,\ldots,x_n}\right| = \left|\mathcal{F}_{|x_1,\ldots,x_{n-1}}\right| + \left|\tilde{\mathcal{F}}_{|x_1,\ldots,x_{n-1}}\right|$$

This is because $\tilde{\mathcal{F}}_{|x_1,\ldots,x_{n-1}}$ are exactly the elements that need to be counted twice (once for $+$ and once for $-$). We also claim that $\mathrm{VC}(\tilde{\mathcal{F}}; x_1, \ldots, x_{n-1}) \leq d' - 1$ because if not, by definition of $\tilde{\mathcal{F}}$ we know that $\tilde{\mathcal{F}}$ can shatter x_n and so we will have that

$$\mathrm{VC}(\tilde{\mathcal{F}}; x_1, \ldots, x_n) = \mathrm{VC}(\tilde{\mathcal{F}}; x_1, \ldots, x_{n-1}) + 1 = d' + 1$$

This is a contradiction as $\tilde{F}$ is a subset of $\mathcal{F}$ which itself has only VC dimension at most d'. Thus we conclude that for any class $\mathcal{F}$ of VC dimension at most d',

$$\begin{aligned}\Pi(\mathcal{F}, n) &= \sup_{x_1,\ldots,x_n} \left|\mathcal{F}_{|x_1,\ldots,x_n}\right| \\ &\leq \sup_{x_1,\ldots,x_n} \left\{\left|\mathcal{F}_{|x_1,\ldots,x_{n-1}}\right| + \left|\tilde{\mathcal{F}}_{|x_1,\ldots,x_{n-1}}\right|\right\}\end{aligned}$$

where $\mathrm{VC}(\tilde{\mathcal{F}}; x_1, \ldots, x_{n-1})$ is at most $d-1$. Using the above bound, the inductive hypothesis and the fact that $g(d', n) = g(d', n-1) + g(d'-1, n-1)$, we conclude that for any class $\mathcal{F}$ with VC dimension at most $d' \leq d$,

$$\begin{aligned}\Pi(\mathcal{F}, n) &\leq \sup_{x_1,\ldots,x_n} \left\{\left|\mathcal{F}_{|x_1,\ldots,x_{n-1}}\right| + \left|\tilde{\mathcal{F}}_{|x_1,\ldots,x_{n-1}}\right|\right\} \\ &\leq g(d', n-1) + g(d'-1, n-1) = g(d', n)\end{aligned}$$

Similarly for any $n' \leq n$, and for any $\mathcal{F}$ with VC dimension at most $d+1$, we can show by repeatedly using the inductive hypothesis, starting from $n' = 2$ up until $n' = n$ that for any $\Pi(\mathcal{F}, n') \leq g(d+1, n')$. This concludes out induction. □

Remark 4.1. *Note that* $\sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{n}{d}\right)^d$. *Hence we can conclude that for any binary classification problem with hypothesis class* $\mathcal{F}$ *in the statistical learning setting, if* $\mathrm{VC}_{\mathcal{F}} \leq d$ *then,*

$$\mathcal{V}_n^{\text{stat}}(\mathcal{F}) \leq \frac{1}{n} \sup_D \mathbb{E}_S \mathbb{E}_\epsilon \left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t)\right] \leq \sqrt{\frac{d \log\left(\frac{n}{d}\right)}{n}}$$

The above statement basically implies that if a binary hypothesis class $\mathcal{F}$ *has finite VC dimension, then it is learnable in the statistical learning (agnostic PAC) framework.* $\log n/d$ *in the above bound can be removed.*