Stat 928: Statistical Learning Theory **Lecture: 24**

Growth Functions and the VC dimension

Instructor: Sham Kakade

1 Growth function

Consider the case $\mathcal{Y} = \{\pm 1\}$ (classification). Let ϕ be the 0-1 loss function and $\mathcal{F}$ be a class of ± 1-valued functions. We can relate the Rademacher average of $\phi_{\mathcal{F}}$ to that of $\mathcal{F}$ as follows.

Recall the following definitions:

$$\mathfrak{R}_m(\phi_{\mathcal{F}}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \epsilon_i \phi(f(X_i), Y_i)\right]$$

where the expectation is with respect to the ϵ_i's, X_i's and Y_i's. The conditional Rademacher average is:

$$\mathfrak{R}_m(\phi_{\mathcal{F}}|X_1^m) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \epsilon_i \phi(f(X_i), Y_i)\middle| X_1^m\right]$$

where the expectation is with respect to the ϵ_i's and Y_i's. Note that:

$$\mathfrak{R}_m(\phi_{\mathcal{F}}) = \mathbb{E}[\mathfrak{R}_m(\phi_{\mathcal{F}}|X_1^m)]$$

where the expectation is with respect to the X_i's.

Lemma 1.1. *Suppose $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$ and let $\phi(y', y) = \mathbf{1}[y' \neq y]$ be the 0-1 loss function. Then we have,*

$$\mathfrak{R}_m(\phi_{\mathcal{F}}) = \frac{1}{2}\mathfrak{R}_m(\mathcal{F}|X_1^m)\,.$$

Proof. Note that we can write $\phi(y', y)$ as $(1 - yy')/2$. Then we have,

$$\begin{aligned}
\mathfrak{R}_m(\phi_{\mathcal{F}}|X_1^m, Y_1^m) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \epsilon_i \frac{1 - Y_i f(X_i)}{2}\middle| X_1^m, Y_1^m\right] \\
&= \frac{1}{2}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m (-\epsilon_i Y_i) f(X_i)\middle| X_1^m, Y_1^m\right] \qquad (1) \\
&= \frac{1}{2}\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^m \epsilon_i f(X_i)\middle| X_1^m\right] \qquad (2) \\
&= \frac{1}{2}\mathfrak{R}_m(\mathcal{F}|X_1^m)\,.
\end{aligned}$$

Equation (1) follows because $\mathbb{E}[\epsilon_i|X_1^m, Y_1^m] = 0$. Equation (2) follows because $\epsilon_i Y_i$'s jointly have the same distribution as ϵ_i's. The proof follows from:

$$\mathfrak{R}_m(\phi_{\mathcal{F}}|X_1^m) = \mathbb{E}[\mathfrak{R}_m(\phi_{\mathcal{F}}|X_1^m, Y_1^m)] = \mathbb{E}[\frac{1}{2}\mathfrak{R}_m(\mathcal{F}|X_1^m)] = \frac{1}{2}\mathfrak{R}_m(\mathcal{F}|X_1^m)$$

where the expectation is with respect to the Y_i's. □

Note that the Rademacher average of the class $\mathcal{F}$ on the set $X_1, \ldots, X_m$ can also be written as

$$\mathfrak{R}_m(\mathcal{F}_{|X_1^m}) = \mathbb{E}\left[\sup_{a \in \mathcal{F}_{|X_1^m}} \frac{1}{m}\sum_{i=1}^{m} \epsilon_i a_i\right] ,$$

where $\mathcal{F}_{|X_1^m}$ is the function class $\mathcal{F}$ restricted to the set $X_1, \ldots, X_m$. That is,

$$\mathcal{F}_{|X_1^m} := \{((f(X_1), \ldots, f(X_m)) \,|\, f \in \mathcal{F}\} .$$

Note that $\mathcal{F}_{|X_1^m}$ is finite and

$$|\mathcal{F}_{|X_1^m}| \leq \min\{|\mathcal{F}|, 2^m\} .$$

Thus we can define the *growth function* as

$$\Pi_{\mathcal{F}}(m) := \max_{x_1^m \in \mathcal{X}^m} |\mathcal{F}_{|x_1^m}| .$$

2 Rademacher Averages and Growth Function

Theorem 2.1. *Let $\mathcal{F}$ be a class of ± 1-valued functions. Then we have,*

$$\mathfrak{R}_m(\mathcal{F}) \leq \sqrt{\frac{2 \ln \Pi_{\mathcal{F}}(m)}{m}} .$$

Proof. We have,

$$\begin{aligned}
&\mathfrak{R}_m(\mathcal{F})\mathbb{E}[\mathfrak{R}_m(\mathcal{F}|X_1^m)] \\
&\quad = \mathbb{E}\left[\mathbb{E}\left[\sup_{a \in \mathcal{F}_{|X_1^m}} \frac{1}{m}\sum_{i=1}^{m} \epsilon_i a_i \,\middle|\, X_1^m\right]\right] \\
&\quad \leq \mathbb{E}\left[\sqrt{m}\frac{\sqrt{2 \ln |\mathcal{F}_{|X_1^m}|}}{m}\right] \\
&\quad \leq \mathbb{E}\left[\sqrt{m}\frac{\sqrt{2 \ln \Pi_{\mathcal{F}}(m)}}{m}\right] \\
&\quad = \sqrt{\frac{2 \ln \Pi_{\mathcal{F}}(m)}{m}}
\end{aligned}$$

Since $f(x_i) \in \{\pm 1\}$, any $a \in \mathcal{F}_{|X_1^m}$ has $\|a\| = \sqrt{m}$. The first inequality above therefore follows from Massart's finite class lemma. The second inequality follows from the definition of the growth function $\Pi_{\mathcal{F}}(m)$. □

Note that plugging in the trivial bound $\Pi_{\mathcal{F}}(m) \leq 2^m$ does not give us any interesting bound. This is quite reasonable since this bound would hold for any function class no matter how complicated it is. To measure the complexity of $\mathcal{F}$, let us look at the first natural number such that $\Pi_{\mathcal{F}}(m)$ falls below 2^m. This brings us to the definition of the *Vapnik-Chervonenkis* dimension.

$\mathcal{X}$	$\mathcal{F}$	$\mathrm{VCdim}(\mathcal{F})$
$\mathbb{R}^2$	convex polygons	∞
$\mathbb{R}^2$	axis-aligned rectangles	4
$\mathbb{R}^2$	convex polygons with d vertices	$2d+1$
$\mathbb{R}^d$	halfspaces	$d+1$

3 Vapnik-Chervonenkis Dimension

The Vapnik-Chervonenkis dimension (or simply the VC-dimension) of a function class $\mathcal{F} \subseteq \{\pm 1\}^{\mathcal{X}}$ is defined as

$$\mathrm{VCdim}(\mathcal{F}) := \max\{m > 0 \,|\, \Pi_{\mathcal{F}}(m) = 2^m\} \ .$$

An equivalent definition is that $\mathrm{VCdim}(\mathcal{F})$ is the size of the largest set shattered by $\mathcal{F}$. A set $\{x_1, \ldots, x_m\}$ is said to be *shattered* by $\mathcal{F}$ if for any labelling $\vec{b} = (b_1, \ldots, b_m) \in \{\pm 1\}^m$, there is a function $f \in \mathcal{F}$ such that

$$(f(x_1), \ldots, f(x_m)) = (b_1, \ldots, b_m) \ .$$

Note that a function $f \in \{\pm\}^{\mathcal{X}}$ can de identified with the subset of $\mathcal{X}$ on which it is equal to $+1$. So, we often talk about the VC-dimension of a collection of subsets of $\mathcal{X}$. The table below gives the VC-dimensions for a few examples.

4 Growth Function and VC Dimension

Suppose $\mathrm{VCdim}(\mathcal{F}) = d$. Then for all $m \leq d$, $\Pi_{\mathcal{F}}(m) = 2^m$. The lemma below, due to Sauer, implies that for $m > d$, $\Pi_{\mathcal{F}}(m) = O(m^d)$, a polynomial rate of growth. This result is remarkable for it implies that the growth function exhibits just two kinds of behavior. If $\mathrm{VCdim}(\mathcal{F}) = \infty$ then $\Pi_{\mathcal{F}}$ grows exponentially with m. On the other hand, if $\mathrm{VCdim}(\mathcal{F}) = d < \infty$ then the growth function is $O(m^d)$.

Sauer's Lemma. *Let $\mathcal{F}$ be such that $\mathrm{VCdim}(\mathcal{F}) \leq d$. Then, we have*

$$\Pi_{\mathcal{F}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} \ .$$

Proof. We prove this by induction on $m + d$. For $m = d = 1$, the above inequality holds as both sides are equal to 2. Assume that it holds for $m-1$ and d and for $m-1$ and $d-1$. We will prove it for m and d. Define the function,

$$h(m, d) := \sum_{i=0}^{d} \binom{m}{i}$$

so that our induction hypothesis is: for $\mathcal{F}$ with $\mathrm{VCdim}(\mathcal{F}) \leq d$, $\Pi_{\mathcal{F}}(m) \leq h(m, d)$. Since

$$\binom{m}{i} = \binom{m-1}{i} + \binom{m-1}{i-1} \ ,$$

is is easy to verify that h satisfies the recurrence

$$h(m, d) = h(m-1, d) + h(m-1, d-1) \ .$$

Fix a class $\mathcal{F}$ with $\mathrm{VCdim}(\mathcal{F}) = d$ and a set $X_1 = \{x_1, \ldots, x_m\} \subseteq \mathcal{X}$. Let $\mathcal{F}_1 = \mathcal{F}_{|X_1}$ and $X_2 = \{x_2, \ldots, x_m\}$ and define the function classes,

$$\mathcal{F}_1 := \mathcal{F}_{|X_1}$$

$$\mathcal{F}_2 := \mathcal{F}_{|X_2}$$

$$\mathcal{F}_3 := \{f_{|X_2} \mid f \in \mathcal{F} \;\&\; \exists f' \in \mathcal{F} \text{ s.t.} \\ \forall x \in X_2,\ f'(x) = f(x) \;\&\; f'(x_1) = -f(x_1)\} \,.$$

Note that $\mathrm{VCdim}(\mathcal{F}') \le \mathrm{VCdim}(\mathcal{F}) \le d$ and we wish to bound $|\mathcal{F}_1|$. By the definitions above, we have

$$|\mathcal{F}_1| = |\mathcal{F}_2| + |\mathcal{F}_3| \,.$$

It is easy to see that $\mathrm{VCdim}(\mathcal{F}_2) \le d$. Also, $\mathrm{VCdim}(\mathcal{F}_3) \le d-1$ because if $\mathcal{F}_3$ shatters a set, we can always add x_1 to it to get a set that is shattered by $\mathcal{F}_1$. By induction hypothesis, $|\mathcal{F}_2| \le h(m-1,d)$ and $|\mathcal{F}_3| \le h(m-1,d-1)$. Thus, we have

$$|\mathcal{F}_{|x_1^m}| = |\mathcal{F}_1| \le h(m-1,d) + h(m-1,d-1) = h(m,d) \,.$$

Since $x_1, \ldots, x_m$ were arbitrary, we have

$$\Pi_{\mathcal{F}}(m) = \sup_{x_1^m \in \mathcal{X}^m} |\mathcal{F}_{|x_1^m}| \le h(m,d) \,.$$

and the induction step is complete. □

Corollary 4.1. *Let $\mathcal{F}$ be such that $\mathrm{VCdim}(\mathcal{F}) \le d$. Then, we have, for $m \ge d$,*

$$\Pi_{\mathcal{F}}(m) \le \left(\frac{me}{d}\right)^d \,.$$

Proof. Since $n \ge d$, we have

$$\begin{aligned}
\sum_{i=0}^{d} \binom{m}{i} &\le \left(\frac{m}{d}\right)^d \sum_{i=0}^{d} \binom{m}{i} \left(\frac{d}{m}\right)^i \\
&\le \left(\frac{m}{d}\right)^d \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^i \\
&\le \left(\frac{m}{d}\right)^d \left(1 + \frac{d}{m}\right)^m \\
&\le \left(\frac{m}{d}\right)^d e^d \,.
\end{aligned}$$

□

5 VC Dimension of halfspaces

Here we prove only the last claim: the VC-dimension of halfspaces in $\mathbb{R}^d$ is $d+1$.

Theorem 5.1. *Let $\mathcal{X} = \mathbb{R}^d$. Define the set of ± 1-valued functions associated with halfspaces,*

$$\mathcal{F} = \left\{ x \mapsto \mathrm{sgn}\,(w \cdot x - \theta) \,\middle|\, w \in \mathbb{R}^d, \theta \in \mathbb{R} \right\} \,.$$

Then, $\mathrm{VCdim}(\mathcal{F}) = d+1$.

Proof. We have to prove two inequalities

$$\mathrm{VCdim}(\mathcal{F}) \geq d+1 \,, \tag{3}$$

$$\mathrm{VCdim}(\mathcal{F}) \leq d+1 \,. \tag{4}$$

To prove the first inequality, we need to exhibit a particular set of size $d+1$ that is shattered by $\mathcal{F}$. Proving the second inequality is a bit more tricky: we need to show that *for all* sets of size $d+2$, there is labelling that cannot be realized using halfspaces.

Let us first prove (3). Consider the set $X = \{\mathbf{0}, \mathbf{e}_1, \ldots, \mathbf{e}_d\}$ which consists of the origin along with the vectors in the standard basis of $\mathbb{R}^d$. Given a labelling $b_0, \ldots, b_d$ of these points, set

$$\theta = -b_0 \,,$$

$$w_i = \theta + b_i \,,\ i \in [d] \,.$$

With these definitions, it immediately follows that $w \cdot \mathbf{0} - \theta = b_0$ and for all $i \in [d]$, $w \cdot \mathbf{e}_i - \theta = b_i$. Thus, X is shattered by $\mathcal{F}$. Since, $|X| = d+1$, we have proved (3).

Before we prove (4), we need the following result from convex geometry.

Radon's Lemma. *Let $X \subset \mathbb{R}^d$ be a set of size $d+2$. Then there exist two disjoint subsets X_1, X_2 of X such that $\mathrm{conv}(X_1) \cap \mathrm{conv}(X_2) \neq \emptyset$. Here $\mathrm{conv}(X)$ denotes the convex hull of X.*

Proof. Let $X = \{x_1, \ldots, x_{d+2}\}$. Consider the following system of $d+1$ equations in the variables $\lambda_1, \ldots, \lambda_{d+2}$,

$$\begin{pmatrix} x_1 & x_2 & \ldots & x_{d+2} \\ 1 & 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{d+2} \end{pmatrix} = \mathbf{0} \,. \tag{5}$$

Since, there are more variables than equations, there is a non-trivial solution $\lambda^* \neq \mathbf{0}$. Define the set of indices,

$$P = \{i \,|\, \lambda_i^* > 0\} \ ,$$

$$N = \left\{j \,\middle|\, \lambda_j^* < 0\right\} \ .$$

Since $\lambda^* \neq \mathbf{0}$, both P and N and non-empty and

$$\sum_{i \in P} \lambda_i^* = \sum_{j \in N} (-\lambda_j^*) \neq 0 \,.$$

Moreover, since λ^* satisfies $\sum_{i=1}^{d+2} \lambda_i^* x_i = \mathbf{0}$, we have

$$\sum_{i \in P} \lambda_i^* x_i = \sum_{j \in N} (-\lambda_j^*) x_j \ .$$

Defining $X_1 = \{x_i \in X \,|\, i \in P\}$ and $X_2 = \{x_i \in X \,|\, i \in N\}$, we see that the point

$$\frac{\sum_{i \in P} \lambda_i^* x_i}{\sum_{i \in P} \lambda_i^*} = \frac{\sum_{j \in N} (-\lambda_j^*) x_j}{\sum_{j \in N} (-\lambda_j^*}$$

lies both in $\mathrm{conv}(X_1)$ as well as $\mathrm{conv}(X_2)$. □

Given Radon's lemma, the proof of (3) is quite easy. We have to show that given a set $X \in \mathbb{R}^d$ of size $d+2$, there is a labelling that cannot be realized using halfspaces. Obtain disjoint subsets X_1, X_2 of X whose existence is guaranteed by Radon's lemma. Now consider a labelling in which all the points in X_1 are labelled $+1$ and those in X_2 are labelled -1. We claim that such a labelling cannot be realized using a halfspace. Suppose there is such a halfspace H. Note that if a halfspace assigns a particular label to a set of points, then every point in their convex hull is also assigned the same label. Thus every point in $\mathrm{conv}(X_1)$ is labelled $+1$ by H while every point in $\mathrm{conv}(X_2)$ is labelled -1. But $\mathrm{conv}(X_1) \cap \mathrm{conv}(X_2) \neq \emptyset$ giving us a contradiction. □

We often work with ± 1-valued functions obtained by thresholding real valued functions at 0. If these real valued functions come from a finite dimensional vector space, the next result gives an upper bound on the VC dimension.

Theorem 5.2. *Let $\mathcal{G}$ be a finite dimensional vector space of functions on $\mathbb{R}^d$. Define,*

$$\mathcal{F} = \{x \mapsto \mathrm{sgn}(g(x)) \,|\, g \in \mathcal{G}\} \ .$$

If the dimension of $\mathcal{G}$ is k then $\mathrm{VCdim}(\mathcal{F}) \leq k$.

Proof. Fix an arbitrary set of $k+1$ points $x_1, \ldots, x_{k+1}$. We show that this set cannot be shattered by $\mathcal{F}$. Consider the linear transformation $T : \mathcal{G} \to \mathbb{R}^{k+1}$ defined as

$$T(g) = (g(x_1), \ldots, g(x_{k+1}) \ .$$

The dimension of the image of $\mathcal{G}$ under T is at most k. Thus, there exists a non-zero vector $\lambda \in \mathbb{R}^{k+1}$ that is orthogonal to it. That is, for all $g \in \mathcal{G}$,

$$\sum_{i=1}^{k+1} \lambda_i g(x_i) = 0 \ . \tag{6}$$

At least one of the sets,

$$P := \{i \,|\, \lambda_i > 0\} \ ,$$
$$N := \{j \,|\, \lambda_j < 0\} \ ,$$

is non-empty. Without loss of generality assume it is P. Consider a labelling of $x_1, \ldots, x_{k+1}$ that assigns the label $+1$ to all x_i such that $i \in P$ and -1 to the rest. If this labelling is realized by a function in $\mathcal{F}$ then there exists $g_0 \in \mathcal{G}$ such that

$$\sum_{i \in P} \lambda_i g_0(x_i) > 0 \ ,$$
$$\sum_{i \in N} \lambda_i g_0(x_i) \geq 0 \ .$$

But this contradicts (6). Therefore $x_1, \ldots, x_{k+1}$ cannot be shattered by $\mathcal{F}$. □