**CSC 588: Machine learning theory** **Spring 2021**

Lecture 6: The Growth Function and VC Examples

*Lecturer: Chicheng Zhang* *Scribe: Katherine Best*

# 1 Growth Function and VC Dimension

Previously we have discussed the VC dimension and the notion of a hypothesis class $\mathcal{H}$ shattering some set of data points $S$. Now we will introduce the growth function as it relates to these concepts.

**Definition 1.** *Suppose we have a dataset $x_1, \ldots, x_n$ of size $n$ . Then growth function $S(\mathcal{H}, n)$ is defined as follows:*

$$S(\mathcal{H}, n) = \max_{x_1,\ldots,x_n} |\Pi_{\mathcal{H}}(\{x_1, \ldots, x_n\})|.$$

**Example 1.** *Suppose we have two distinct points $x_1, x_2 \in \mathbb{R}$ with $x_1 < x_2$ and let $\mathcal{H}$ be the threshold class. Note that, if we let $t$ be the threshold value, we can generate three distinct labeling for these points by choosing $t < x_1$, $x_1 < t < x_2$, or $x_2 < t$. So, since $\mathcal{H}$ can generate three possible labelings on this dataset, we have $S(\mathcal{H}, 2) = 3$.*

Now note that if we have $S(\mathcal{H}, n) = 2^n$ then there exists a dataset of $n$ points for which we can generate all $2^n$ possible labelings. So, $S(\mathcal{H}, n) = 2^n$ implies that there is a dataset of size $n$ shattered by $\mathcal{H}$. Using this observation, we can rewrite the VC dimension definition using the growth function:

$$\mathrm{VC}(\mathcal{H}) = \max\{n : S(\mathcal{H}, n) = 2^n\}.$$

Next, we consider the following theorem, which provides some motivation for why we are interested in the VC dimension.

**Theorem 2.** *Suppose we have a hypothesis class $\mathcal{H}$ with $VC(\mathcal{H}) = d < \infty$ and a set of $n$ i.i.d. examples $(x_1, y_1), \ldots, (x_n, y_n)$ drawn from $\mathcal{D}$. Then, with probability $1 - \delta$*

$$\sup_{h\in\mathcal{H}} |\operatorname{err}(h, S) - \operatorname{err}(h, \mathcal{D})| \le c\sqrt{\frac{d \ln(n/d) + \ln(2/\delta)}{n}}.$$

This theorem states that if a hypothesis class has a finite VC dimension then the empirical error will concentrate around the generalization error for all classifiers in this class. With this theorem, we can apply the same analysis for empirical risk minimization that we talked about last time. We can conclude that the for the $\hat{h}$ obtained via ERM,

$$\operatorname{err}(\hat{h}, \mathcal{D}) \le \min_{h\in\mathcal{H}} \operatorname{err}(h, \mathcal{D}) + \tilde{\mathcal{O}}(\sqrt{d/n})$$

which implies that $\mathcal{H}$ is agnostic PAC learnable by ERM. So, we can think of this theorem as a generalization of the previously developed theorem of agnostic PAC learnabiliy for finite hypothesis classes.

Note that we use the notation $\tilde{\mathcal{O}}$ to be big O notation where log factors are ignored.

Next we will talk about some of the properties of the growth function and VC dimension.

**Lemma 3.** *If $|\mathcal{H}| < \infty$, then the following two statements hold:*

1. $S(\mathcal{H}, n) \le |\mathcal{H}|$
2. $\mathrm{VC}(\mathcal{H}) \le \log |\mathcal{H}|$.

To see why the first statement is true, consider that there are only $|\mathcal{H}|$ classifiers in $\mathcal{H}$. So we can generate at most $|\mathcal{H}|$ different labelings. Now to see why the second statement holds, recall that $\mathrm{VC}(\mathcal{H}) = \max\{n : S(\mathcal{H}, n) = 2^n\}$. By statement 1 we have that $S(\mathcal{H}, n) = 2^n \le |\mathcal{H}|$. So if we take the log of both sides we obtain $n \le \log |\mathcal{H}|$. Thus, $\mathrm{VC}(\mathcal{H})$ must be less than $\log |\mathcal{H}|$.

## 2 Examples

1. *threshold class*

   VC($\mathcal{H}$) = 1 (shown in previous lecture)
   $S(\mathcal{H}, n) = n + 1$

   To see why we have this second result, consider the distinct points $x_1, \ldots, x_n$ and begin with a threshold to the left of all the points. Now as we move the threshold in the positive direction, every time we cross a point the number of labelings induced increases for one. Thus, we have a total of $n+1$ distinct labelings.

2. *interval class in* $[0, 1]$ This class is formally defined by

   $$\mathcal{H} = \{h_{a,b}(x) = 2\mathbb{I}(a \leq x \leq b) - 1 : 0 \leq a \leq b \leq 1\}.$$

   We want to find VC($\mathcal{H}$) and $S(\mathcal{H}, n)$.

   To find VC($\mathcal{H}$) consider the case where we have two distinct points $x_1$ and $x_2$ with $x_1 < x_2$. We can show that this dataset is shatterable by $\mathcal{H}$. If both points have a positive label then we can choose the interval $a < x_1 < x_2 < b$ and if both points have a negative label then we can choose $0 < a < b < x_1$. For the case where only one point has a positive label we can choose the interval to contain only the positive point. So if we assume without loss of generality that $x_1$ has a positive label and $x_2$ has a negative label, then we can choose $a < x_1 < b < x_2$. Thus, all possible labels for this dataset can be generated by $\mathcal{H}$ and we can conclude that $\mathcal{H}$ shatters this dataset.

   However, we can also show that $\mathcal{H}$ is incapable of shattering any dataset with three points. To see this consider a dataset with three points $x_1 < x_2 < x_3$ and the labeling $(+, -, +)$. Note that if $h_{a,b}(x_1) = +1$ then we must have $x \leq x_1$ and if $h_{a,b}(x_3) = +1$ then we must have $b \geq x_3$. However, this necessarily implies that $a < x_2 < b$ which would give us that $h_{a,b}(x_2) = +1$. Thus, datasets of size three are not shatterable by $\mathcal{H}$.

   Now since there is a dataset of size 2 that is shatterable and no dataset of size 3 that is shatterable, we can conclude that VC($\mathcal{H}$) = 2.

   Next we consider $S(\mathcal{H}, n)$. Note that there are $\binom{n}{2}$ ways to choose the boundary points such that they include at least two points, $\binom{n}{1}$ labelings with only one positive point, and $\binom{n}{0}$ labelings with no positive points. Thus, the growth function is

   $$S(\mathcal{H}, n) = \binom{n}{2} + \binom{n}{1} + \binom{n}{0}.$$

3. *homogeneous linear classifiers in* $\mathbb{R}^d$ This class is formally defined by

   $$\mathcal{H} = \{h_w(x) = 2\mathbb{I}(\mathbf{w} \cdot \mathbf{x} > 0) - 1 : \mathbf{w} \in \mathbb{R}^d\}.$$

   Note that this class of classifiers is the set of linear classifiers with normal vector $\mathbf{w}$ that pass through the origin.

   To find VC($\mathcal{H}$) we will begin by building some intuition with the $d = 2$ case. If we have one point in the first quadrant and one point in the second quadrant, as is shown in Figure 1, then there are classifiers in $\mathcal{H}$ that can generate any of the possible labelings of these points. If they have the same label then the separating line shown in black (with normal **w** also shown in black) in Figure 1 will produce this classification. Note rotating a separating line (and associated normal) 180 degrees will result in the same separating line but the normal will be pointing in the opposite direction, so the labelings generated will be the negative of what they were originally. So, the separating line in black could generate either the (+,+) labeling or the (-,-) labeling depending on which way **w** is pointing. Similarly, we can pick **w** normal to the separating line shown in red to generate either set of labels where the signs are different. That is, (+,-) or (-,+). Thus, we can generate all possible labelings for this dataset of size 2. So we know that VC($\mathcal{H}$) $\geq$ 2.

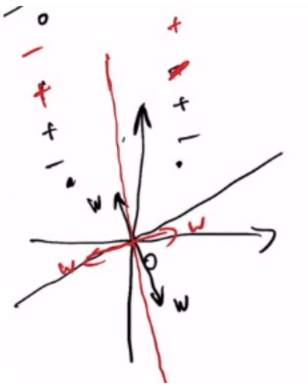

Figure 1: labelings for 2 points in 2 dimensions

Next, we consider the case where we have three points in two dimensions. Note that in this case if we are not able to draw a line that puts all three points on one side of the line, then we cannot generate labelings with all the same sign (such as (+,+,+)). If, however, we can find a **w** such that we can put all points on one side of the line normal to **w**, then we cannot generate the labeling (+,-,+). This is illustrated in Figure 2. Thus, we can conclude that any three points are not shatterable by $\mathcal{H}$. So for

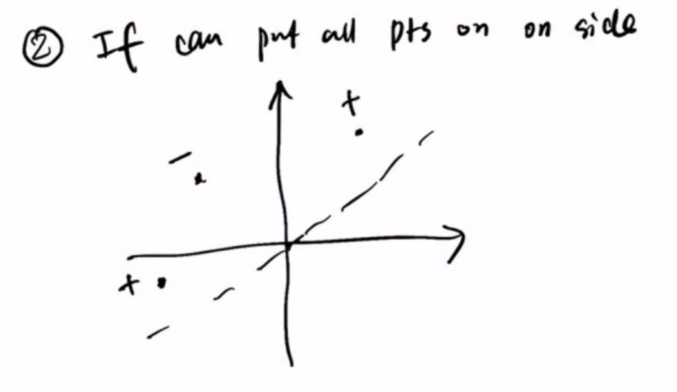

Figure 2: three points on one side of a line

$d = 2$ we have $VC(\mathcal{H}) = 2$.

Now we consider the general case for dimension $d$. To begin, we claim that $\mathcal{H}$ can shatter and linearly independent points, $x_1, \ldots, x_d$. To see why this is the case consider the linear prediction on this set of examples:

$$\text{sign}\left(\begin{pmatrix} x_1^T \\ \vdots \\ x_d^T \end{pmatrix} \cdot \mathbf{w}\right) = \begin{pmatrix} \text{sign}(\langle \mathbf{w}, x_1 \rangle) \\ \vdots \\ \text{sign}(\langle \mathbf{w}, x_d \rangle) \end{pmatrix}$$

Note that since we have $d$ vectors$x_i$, each of which is linearly independent, we have that any other vector in $\mathbb{R}^d$ can be described as a linear combination of these vectors. That is,

$$\{\mathbf{X} \cdot \mathbf{w} : \mathbf{w} \in \mathbb{R}^d\} = \mathbb{R}^d.$$

Now if we consider the labeling $\mathbf{l} = (l_1, \dots, l_d) \in \{+1, -1\}^d$ we have that $\mathbf{l} = \text{sign}(\mathbf{X} \cdot \mathbf{w})$. We can find $\mathbf{w}$ that satisfies this by solving $\mathbf{w} = \mathbf{X}^{-1}\mathbf{l}$. Thus, a set of $d$ linearly independent points is shatterable.

Next we claim that any set of $d+1$ points is not shatterable by $\mathcal{H}$. To show this, note that for any $d+1$ points we can find $a_1, \dots, a_{d+1}$ not all zero and and such that there exists some $i^*$ with $a_{i^*} > 0$ such that

$$\sum_{i=1}^{d+1} a_i x_i = 0.$$

Now we define $\mathbf{l}$ as follows

$$l_i = \begin{cases} +1 & a_i > 0 \\ -1 & a_i \leq 0 \end{cases}$$

for all $i = 1, \dots, d+1$. If there exists $\mathbf{w}$ that achieves the labeling $\mathbf{l}$ then for all $a_i > 0$ corresponding to $l_i = +1$ then we must have $w_i x_i > 0$ and for $a_i \leq 0$ corresponding to $l_i = -1$ we must have $w_i x_i \leq 0$. Then the sum

$$\sum_{i=1}^{d+1} a_i \langle \mathbf{w}, x_i \rangle \geq 0.$$

Further, since we have some $a_{i^*} > 0$ there is some term $a_{i^*}\langle \mathbf{w}, x_i \rangle > 0$. This implies that the sum must be strictly positive:

$$\sum_{i=1}^{d+1} a_i \langle \mathbf{w}, x_i \rangle > 0.$$

This contradicts our assumption

$$\sum_{i=1}^{d+1} a_i x_i = 0.$$

Thus, we have shown that there is not set of $d+1$ points that are shatterable by $\mathcal{H}$. So, we can conclude that $\text{VC}(\mathcal{H}) = d$.

4. *non-homogeneous linear classifiers in* $\mathbb{R}^d$ This class is formally defined by

$$\mathcal{H} = \{h_{w,b}(x) = 2\mathbb{I}(\mathbf{w} \cdot \mathbf{x} + b > 0) - 1 : \mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}\}.$$

We can show that $\text{VC}(\mathcal{H}) = d+1$ but this is left as an exercise.

# 3 Sauer's Lemma and a Theorem

**Lemma 4.** *Suppose we have a hypothesis class* $\mathcal{H}$ *with* $\text{VC}(\mathcal{H}) = d$ *and that we have* $n$ *data points* $x_1, \dots, x_n$. *Then,*

$$S(\mathcal{H}, n) \begin{cases} = 2^n & n \leq d \\ \leq \sum_{i=0}^d \binom{n}{i} & n > d \end{cases}$$

Note that

$$\sum_{i=0}^{d} \binom{n}{i} \leq \begin{cases} n^{d+1} & n \geq 2 \\ \left(\frac{e \cdot n}{d}\right)^d & n \geq d+2 \end{cases}.$$

So, this growth function upper bound is polynomial instead of exponential in the size of in $n$.

**Theorem 5.** *Suppose we have a hypothesis class $\mathcal{H}$ with $\mathrm{VC}(\mathcal{H}) = d$ and $n$ points $S = \{x_1, \ldots, x_n\}$. Then,*

$$|\Pi_{\mathcal{H}}(S)| \leq |\{T \subseteq S : \mathcal{H} \textit{ shatters } T\}|.$$

Note that if $\mathcal{H}$ shatters $T$ then $|T| \leq d$. So,

$$\begin{aligned} |\Pi_{\mathcal{H}}(S)| &\leq |\{T \subseteq S : \mathcal{H} \text{ shatters } T\}| \\ &\leq \{T \subseteq S : |T| \leq d\} \\ &\leq \sum_{i=0}^{d} \binom{n}{i} \end{aligned}$$

where the last inequality follows from the fact that for each $|T| = i$ we have $\binom{n}{i}$ different choices of $T$.