STAT 598Y Statistical Learning Theory **Instructor: Jian Zhang**

Lecture 5: Growth Function and VC Dimension

We have considered the case when $\mathcal{H}$ is finite or countably infinite. In practice, however, the function class $\mathcal{H}$ could be uncountable. Under this situation, the previous method does not work. The key idea is to group functions based on the sample.

Given a sample $\mathcal{D}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, and define $S = \{x_1, \ldots, x_n\}$. Consider the set

$$\mathcal{H}_S = \mathcal{H}_{x_1,\ldots,x_n} = \{(h(x_1), \ldots, h(x_n) : h \in \mathcal{H}\}.$$

The size of this set is the total number of possible ways that $S = \{x_1, \ldots, x_n\}$ can be classified. For binary classification the cardinality of this set is always finite, no matter how large $\mathcal{H}$ is.

Definition (Growth Function). *The growth function is the maximum number of ways into which n points can be classified by the function class:*

$$G_{\mathcal{H}}(n) = \sup_{x_1,\ldots,x_n} |\mathcal{H}_S|.$$

Growth function can be thought as a measure of the "size" for the class of functions $\mathcal{H}$. Several facts about the growth function:

- When $\mathcal{H}$ is finite, we always have $G_{\mathcal{H}}(n) \leq |\mathcal{H}| = m$.
- Since $h(x) \in \{0, 1\}$, we have $G_{\mathcal{H}}(n) \leq 2^n$. If $G_{\mathcal{H}}(n) = 2^n$, then there is a set of n points such that the class of functions $\mathcal{H}$ can generate any possible classification result on these points.

Definition (Shatterring). *We say that $\mathcal{H}$ shatters S if $|\mathcal{H}_S| = 2^{|S|}$.*

Definition (VC Dimension). *The VC dimension of a class $\mathcal{H}$ is the largest $n = d_{VC}(\mathcal{H})$ such that*

$$G_{\mathcal{H}}(n) = 2^n.$$

In other words, VC dimension of a function class $\mathcal{H}$ is the cardinality of the largest set that it can shatters.

Example. Consider all functions of the form $\mathcal{H} = \{h(x) = I(x \leq \theta), \theta \in \mathbb{R}\}$. Then it can shatter 2 points but for any three points it cannot shatter. □

Example. Consider all linear classifiers in a 2-d space, i.e. $\mathcal{X} = \mathbb{R}^2$. In this case, all linear classifiers can shatter a set of 3 points. No set of four points can be shattered by linear classifiers. So the VC dimension in this case is 2. □

Example. Consider all linear classifiers in a p-dimensional Euclidean space, i.e. $\mathcal{X} = \mathbb{R}^p$. Given $x_1, \ldots, x_n \in \mathbb{R}^p$, we define the augmented data vector

$$z_i = [1, x_i]^T \in \mathbb{R}^{p+1}, \; i = 1, \ldots, n.$$

Then the set of all linear classifiers can be written as

$$\mathcal{H} = \left\{h : \; h(z) = \mathsf{sign}\left(\theta^T z\right), \theta \in \mathbb{R}^{p+1}\right\}.$$

Define

$$\mathbf{Z} = [z_1, z_2, \ldots, z_n] \in \mathbb{R}^{(p+1)\times n}$$

and we argue that $x_1, \ldots, x_n$ is shattered by $\mathcal{H}$ if and only if the n columns of $\mathbf{Z}$ are linearly independent.

- If columns $z_1, \ldots, z_n$ are linearly independent, we have $n \leq p + 1$ and for any possible classification assignment $\mathbf{y} \in \{\pm 1\}^n$ the linear system $\mathbf{Z}^T \theta = \mathbf{y}$ must have a solution. Thus, there is a linear classifier in $\mathcal{H}$ (by taking the solution of the linear equation) which can produce such arbitrary class assignment $\mathbf{y}$.

- Suppose columns $z_1, \ldots, z_n$ are not linearly independent. For $\mathcal{H}$ to shatter the set there must exist a $\theta \in \mathbb{R}^{p+1}$ with $\mathsf{sign}(z_1^T\theta)$, ..., $\mathsf{sign}(z_n^T\theta)$ taking any possible vector in $\{\pm 1\}^n$. In other words, this means that the vector $\mathbf{Z}^T\theta$ can be in any of the 2^n orthants of $\mathbb{R}^n$. However, this contradicts the fact that $z_1, \ldots, z_n$ are linearly dependent.

Since if $n > p+1$ it is not possible to have $\mathbf{Z}$'s columns linearly independent, but for $n \le p+1$ we can always find such $x_1, \ldots, x_n$ to make it happen, we have $d_{VC}(\mathcal{H}) = p+1$. $\square$

A somewhat surprising result shows that the growth function $G_{\mathcal{H}}$ either grows exponentially in n or only increases polynomially in n, depends on whether n is greater than its VC dimension $d_{VC}(\mathcal{H})$ or not.

Theorem 5-1 (Sauer). *If $\mathcal{H}$ is a class of functions with binary outputs and its VC dimension is $d = d_{VC}(\mathcal{H})$. Then for all $n \in \mathbb{N}$,*

$$G_{\mathcal{H}}(n) \le \sum_{i=0}^{d} \binom{n}{i}.$$

Furthermore, for all $n \ge d$, we have

$$G_{\mathcal{H}}(n) \le \left(\frac{en}{d}\right)^d.$$

PROOF. For any $S = \{x_1, \ldots, x_n\}$, consider a table containing values of functions in $\mathcal{H}_S$ (i.e. we only consider distinct ones projected onto the sample S), each row for one such unique tuple. For example, if $S = \{x_1, x_2, \ldots, x_5\}$ we might have the following table T:

$h(x_1)$	$h(x_2)$	$h(x_3)$	$h(x_4)$	$h(x_5)$
-	+	-	+	+
+	-	-	+	+
+	+	+	-	+
-	+	+	-	-
-	-	-	+	-

Table 1: An example of $\mathcal{H}$ projected onto $S = \{x_1, \ldots, x_5\}$

Each row is one possible tuple for some $h \in \mathcal{H}$ evaluated on the sample S. Obviously the number of rows in T is the same as the cadinality of $|\mathcal{H}_S|$. Thus we can bound the growth function of $\mathcal{H}$ by the maximum number of rows in table T. Next we transform the table T by processing each column sequentially. For example, to process the first column, for each row, we replace a " $+$ " into a " $-$ " unless it produces a duplicated row in the table. Table 2 shows the table after processing the first column (left table) and the final table after processing all 5 columns (right table).

$h^*(x_1)$	$h(x_2)$	$h(x_3)$	$h(x_4)$	$h(x_5)$
-	+	-	+	+
-	-	-	+	+
-	+	+	-	+
-	+	+	-	-
-	-	-	+	-

$h^*(x_1)$	$h^*(x_2)$	$h^*(x_3)$	$h^*(x_4)$	$h^*(x_5)$
-	+	-	-	-
-	-	-	+	+
-	-	-	-	+
-	-	-	-	-
-	-	-	+	-

Table 2: transformed tables (left: after processing the first column; right: after processing all 5 columns)

Now we have the following observations:

1. The size of the tables are not changed for such transformations, and rows in the final table T^* are still unique. Thus we use the upper bound of the number of rows in T^* to bound the growth function $G_{\mathcal{H}}(n)$.

2. The final table T^* possess the property that if we replace any "+" to "−", it will result in a duplication. So the set of " + " elements in each row must be a subset of S that can be shattered by the table T^* (in fact, by the set of functions $\mathcal{H}^*$ corresponding to the table T^*).

3. If a subset $A \subset S$ can be shattered by a latter table T_{k+1}, then it must also be shattered by the previous table T_k. To see this, notice that if A does not contain the transformed column x_k, then the result holds trivially as all columns in A remain the same in T_k and T_{k+1}. If A contain the transformed column x_k, then for each +/- combination ($2^{|A|-1}$) of elements in $A \backslash \{x_k\}$, we must have two rows in T_{k+1} such that they have " + " and "-" values in the column x_k. Now in the previous table T_k, those two rows must also exist. The "+" row is obviously there, and it must also contain the "-" row since otherwise the "+" would not show up in the later table T_{k+1} by the processing procedure.

Since $d_{VC}(T^*) \leq d_{VC}(T) = d_{VC}(\mathcal{H}) = d$ by observation 3, each row in T^* has at most d "+" elements. Thus an upper bound of the total number of rows in T^* is $\sum_{i=0}^{d} \binom{n}{i}$, which is also an upper bound of the growth function $G_{\mathcal{H}}(n)$ by observation 1.

The second statement comes from the fact that for $n \geq d$,

$$\begin{aligned} \sum_{i=0}^{d} \binom{n}{i} &\leq \left(\frac{n}{d}\right)^d \sum_{i=0}^{d} \binom{n}{i} \left(\frac{d}{n}\right)^i \\ &= \left(\frac{n}{d}\right)^d \left(1 + \frac{d}{n}\right)^n \\ &\leq \left(\frac{en}{d}\right)^d. \end{aligned}$$

□