Introduction to Machine Learning	**Fall Semester**
Recitation 7	
Lecturer: Regev Schweiger	*Scribe: Regev Schweiger*

7.1 Review and Preliminaries

We will first prove a few helpful Lemmas that will help us set the stage to the main result of this recitation.

7.1.1 VC dimension and the Sauer-Shelah Lemma

Recall the definition of a projection of an hypothesis class $\mathcal{H}$ on a set S:

Definition For a hypothesis class $\mathcal{H}$ over $\mathcal{X}$ and for any $S = \{x_1, \ldots, x_m\} \subseteq X$, we define the projection of $\mathcal{H}$ on S, $\Pi_{\mathcal{H}}(S) \subseteq \{-1, 1\}^m$, as:

$$\Pi_{\mathcal{H}}(S) = \{\langle h(x_1), \ldots, h(x_m)\rangle : h \in \mathcal{H}\}$$

Recall also the definition of the VC dimension:

Definition *The VC dimension* of $\mathcal{H}$ is the maximum size of a set S shattered by $\mathcal{H}$:

$$VCdim(\mathcal{H}) = max\{d : \exists S : |S| = d \ \ and \ \ |\Pi_{\mathcal{H}}(S)| = 2^d\}.$$

Finally, recall the Sauer-Shelah Lemma:

Lemma 7.1 (Sauer-Shelah Lemma) *Let $VCdim(\mathcal{H}) = d$ and $|S| = m$, then*

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

From the Sauer-Shelah Lemma, we can get a bound that's easier to work with, in the case $m \geq d$.

Lemma 7.2 *Let $VCdim(\mathcal{H}) = d$ and $|S| = m$, and assume $m \geq d$. Then,*

$$|\Pi_{\mathcal{H}}(S)| \leq \left(\frac{em}{d}\right)^d$$

Proof:

$$\begin{aligned}
|\Pi_{\mathcal{H}}(S)| &\leq \sum_{i=0}^{d} \binom{m}{i} \\
&\leq \left(\frac{m}{d}\right)^d \sum_{i=0}^{d} \left(\frac{d}{m}\right)^d \binom{m}{i} \\
&\leq \left(\frac{m}{d}\right)^d \sum_{i=0}^{d} \left(\frac{d}{m}\right)^i \binom{m}{i} \\
&\leq \left(\frac{m}{d}\right)^d \sum_{i=0}^{m} \left(\frac{d}{m}\right)^i \binom{m}{i} \\
&\leq \left(\frac{m}{d}\right)^d \left(1+\frac{d}{m}\right)^m \leq \left(\frac{m}{d}\right)^d \cdot e^d
\end{aligned}$$

In the third inequality, we used the fact that $d/m \leq 1$; the fourth inequality is true because we extended the sum from up to i to be until m; and the last inequality is true because the sequence $(1 + d/m)^m$ converges to e^d from below. □

7.1.2 Properties of the growth function

The growth function is defined as the largest projection size for a sample of size m. That is:
Definition For a hypothesis class $\mathcal{H}$ over $\mathcal{X}$, the growth function $\pi_{\mathcal{H}}$ is defined as

$$\pi_{\mathcal{H}}(m) = \max_{|S|=m} |\Pi_{\mathcal{H}}(S)|.$$

Note that from Lemma 7.2, we have that when $m \geq d$, $\pi_{\mathcal{H}}(m) \leq (me/d)^d$.

We have so far limited ourselves to functions (hypotheses) that output $\mathcal{Y} = \{-1, 1\}$. We will need to extend our discussion to a general family function $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$, where $|\mathcal{Y}|$ is finite. First, let's discuss the effect of concatenation of two functions. That is, if $f_1, f_2 : \mathcal{X} \to \mathcal{Y}$, then we can concatenate their outputs to create a new function, denoted $f_1 \times f_2$, for which $f_1 \times f_2 : \mathcal{X} \to \mathcal{Y} \times \mathcal{Y}$, defined simply as $(f_1 \times f_2)(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}))$.

Lemma 7.3 *Let $\mathcal{F}_1 \subseteq \mathcal{Y}_1^{\mathcal{X}}$ and $\mathcal{F}_1 \subseteq \mathcal{Y}_1^{\mathcal{X}}$ be two function families from $\mathcal{X}$ to $\mathcal{Y}_1$ or $\mathcal{Y}$, respectively. Define $\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2$ to be the set of functions which are a concatenation of a function from $\mathcal{F}_1$ and from $\mathcal{F}_2$. That is,*

$$\mathcal{F} = \mathcal{F}_1 \times \mathcal{F}_2 = \{f_1 \times f_2 | f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$$

Then,

$$\pi_{\mathcal{F}}(m) \leq \pi_{\mathcal{F}_1}(m) \cdot \pi_{\mathcal{F}_2}(m)$$

Proof: Let $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \in \mathcal{X}^m$ be a sample of size m. The projection of $\mathcal{F}$ on S is:

$$\begin{aligned}\Pi_{\mathcal{F}}(S) &= \{\langle f(\mathbf{x}_1), \ldots, f(\mathbf{x}_m)\rangle : f \in \mathcal{F}\} \\ &= \{\langle (f_1 \times f_2)(\mathbf{x}_1), \ldots, (f_1 \times f_2)(\mathbf{x}_m)\rangle : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}\end{aligned}$$

So, by definition, $|\Pi_{\mathcal{F}}(S)| = |\Pi_{\mathcal{F}_1}(S)| \cdot |\Pi_{\mathcal{F}_2}(S)|$ (this is simply a cartesian product). Thus, by the definition of the growth function, for every S,

$$|\Pi_{\mathcal{F}}(S)| = |\Pi_{\mathcal{F}_1}(S)| \cdot |\Pi_{\mathcal{F}_2}(S)| \leq \pi_{\mathcal{F}_1}(m) \cdot \pi_{\mathcal{F}_2}(m)$$

However, note that this holds for an arbitrary S; in particular, this is true for the S for which $|\Pi_{\mathcal{F}}(S)| = \pi_{\mathcal{F}}(m)$; therefore,

$$\pi_{\mathcal{F}}(m) \leq \pi_{\mathcal{F}_1}(m) \cdot \pi_{\mathcal{F}_2}(m)$$

□

Next, we discuss what happens when we compose two family functions.

Lemma 7.4 *Let $\mathcal{F}_1 \subseteq \mathcal{Y}_1^{\mathcal{X}}$ and $\mathcal{F}_2 \subseteq \mathcal{Y}_2^{\mathcal{Y}_1}$ be two function families. Define $\mathcal{F} = \mathcal{F}_2 \circ \mathcal{F}_1$ to be the set of functions which are a composition of a function from $\mathcal{F}_1$ and from $\mathcal{F}_2$. That is,*

$$\mathcal{F} = \mathcal{F}_2 \circ \mathcal{F}_1 = \{f_2 \circ f_1 | f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$$

Then,

$$\pi_{\mathcal{F}}(m) \leq \pi_{\mathcal{F}_1}(m) \cdot \pi_{\mathcal{F}_2}(m).$$

Proof: Again, let $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \in \mathcal{X}^m$ be a sample of size m. The projection of $\mathcal{F}$ on S is:

$$\begin{aligned}\Pi_{\mathcal{F}}(S) &= \{\langle f(\mathbf{x}_1), \ldots, f(\mathbf{x}_m)\rangle : f \in \mathcal{F}\} \\ &= \{\langle f_2(f_1(\mathbf{x}_1)), \ldots, f_2(f_1(\mathbf{x}_m))\rangle : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\} \\ &= \bigcup_{u \in \Pi_{\mathcal{F}_1}(S)} \{\langle f_2(u_1), \ldots, f_2(u_m)\rangle : f_2 \in \mathcal{F}_2\}\end{aligned}$$

We can now use a union bound to get

$$\begin{aligned}|\Pi_{\mathcal{F}}(S)| &\leq \left| \bigcup_{u \in \Pi_{\mathcal{F}_1}(S)} \{\langle f_2(u_1), \ldots, f_2(u_m)\rangle : f_2 \in \mathcal{F}_2\} \right| \\ &\leq \sum_{u \in \Pi_{\mathcal{F}_1}(S)} |\{\langle f_2(u_1), \ldots, f_2(u_m)\rangle : f_2 \in \mathcal{F}_2\}|\end{aligned}$$

The size of each one of these sets is bounded by the growth function of $\mathcal{F}_2$, while the number of terms is bounded by the growth function of $\mathcal{F}_1$:

$$\ldots \leq \sum_{u \in \Pi_{\mathcal{F}_1}(S)} \pi_{\mathcal{F}_2}(m) = |\Pi_{\mathcal{F}_1}(S)| \cdot \pi_{\mathcal{F}_2}(m) \leq \pi_{\mathcal{F}_1}(m) \cdot \pi_{\mathcal{F}_2}(m)$$

Again, since S was arbitrary, we get

$$\pi_{\mathcal{F}}(m) \leq \pi_{\mathcal{F}_1}(m) \cdot \pi_{\mathcal{F}_2}(m).$$

□

7.2 The VC-dimension of Neural Networks

We now have everything we need to prove the VC-dimension of a neural network (NN). We will discuss a simpler case, where the activation function is binary; i.e. the sign of the result. We will also assume, for simplicity, that the size of each layer is the same (except, of course, the output layer). The multilayer NN is therefore defined as follows; $\mathbf{z}_0 := \mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^d$ is the input. Then, for $t = 1, \ldots, L$,

$$\mathbf{z}_{t+1} = h(\mathbf{W}^{(t+1)}\mathbf{z}_t - \mathbf{b}^{(t+1)})$$

where $\mathbf{b}^{(t+1)}, \mathbf{z}_{t+1} \in \mathbb{R}^d$ are vectors of size d, and $\mathbf{W}^{(t+1)}$ is a $d \times d$ matrix (in the general multilayer NN, d might change between t-s). The activation function h, operates pointwise; that is,

$$h(\mathbf{z})_i = sign(\mathbf{z}_i)$$

Let $\mathcal{C}$ be the set of functions that are implementable as a NN as described above. We will show that $\mathcal{C}$ can be built by a composition and concatenation of basic function families, whose VC-dimension we know. This will give us an upper bound on $VCdim(\mathcal{C})$. Similar results are also true when the activation function is, e.g., the sigmoid, with similar proofs.

7.2.1 A single node

Let's begin by looking at a single node, e.g., node number i in layer t. Denote the i-th row of $\mathbf{W}^{(t)}$ by $\mathbf{w}^{(t)}_{i,:}$. Its output function, $f_{i,t}$, is given by:

$$(\mathbf{z}_{t+1})_i = f_{i,t+1} = sign\left(\left\langle \mathbf{w}^{(t+1)}_{i,:}, \mathbf{z}_t \right\rangle + (\mathbf{b}_{t+1})_i\right)$$

What is this function? We already know it - if we mark $\mathbf{w} = \mathbf{w}^{(t+1)}_{i,:}$ and $b = (\mathbf{b}_{t+1})_i$, then the function is $sign(\mathbf{w} \cdot \mathbf{z}_t + b)$ - this is exactly the family of linear separators! Mark by $\mathcal{H}$ the family of hyperplane separators in $\mathbb{R}^d$ (this was called C_4 in the lesson 3 scribe). Then, we have seen in

class (from Radon's theorem) that $VCdim(\mathcal{H}) = d + 1$. Therefore, the growth function of $\mathcal{H}$, by Lemma 7.2, is bounded by:

$$\pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d+1}\right)^{d+1}$$

To be continued...