CS 7545: Machine Learning Theory **Fall 2018**

Lecture 25: VC Dimension of Neural Networks

Lecturer: Jacob Abernethy *Scribes: Peng Li, Shinya Hirata*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

25.1 Review

1. **Growth Function:** Given binary hypothesis class $\mathcal{H}$, we define growth function of $\mathcal{H}$ as:

$$\Pi_{\mathcal{H}}(m) = \max_{\substack{|S|=m \\ S=S_1...S_m}} |\{h(x_1)...h(x_m) : h \in \mathcal{H}\}|$$

2. **VC-dimmension:** Given binary hypothesis class $\mathcal{H}$, we define VC-dimmension of $\mathcal{H}$ as:

$$\text{VC-dim}(\mathcal{H}) = \max_d \{d : \Pi_{\mathcal{H}}(m) = 2^d\}$$

3. **Sauer's Lemma:**

$$\Pi_{\mathcal{H}}(m) \leq (\frac{em}{d})^d, \text{where } d = \text{VC-dim}(\mathcal{H})$$

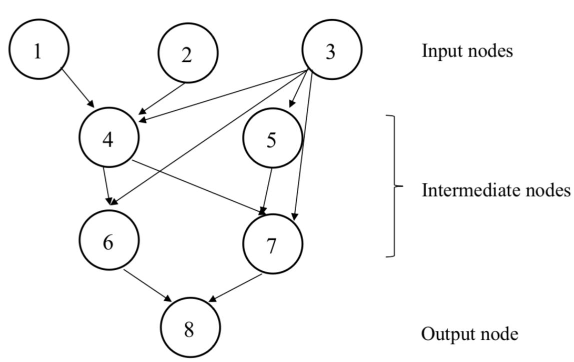
4. **Linear Threshold Function:** $f : \mathbb{R}^d \to \{0, 1\}$ is a linear threshold function if:

$$f(x) = sign(\sum_{i=1}^{d} w_i x_i), \text{where } sign(\sim) := \mathbb{1}[\sim \geq 0]$$

If $\mathcal{H} = \{\text{linear threshold on } \mathbb{R}^d\}$, VC-dim$(\mathcal{H}) = d$ (If intercept term is in the definition of linear threshold function, the VC-dimension is d+1).

25.2 Neural Networks

Definition 25.1 (Linear Threshold Neural Network) *A linear threshold neural network is a function $f : \mathbb{R}^d \to \{0, 1\}$, computed as follows:*



We are given d input nodes, k intermediate nodes, one output node, and an edge set E connecting those nodes so that they form a Directed Acyclic Graph. Let us assign a topological ordering to those nodes (i.e. if $(v', v) \in E \Rightarrow v' < v$).

The parameters of network are the weights assigned to each edge: $\{w_e : e \in E\}$

Given an initial input vector $x \in \mathbb{R}^d$, the output value of node l is described by the function:

$$f_l(x) = sign(\sum_{v',v \in E} w_{v',v} z_{v'})$$

where x is the input vector, $w_{v',v}$ is the weight of edge (v', v) and

$$z_{v'} = \begin{cases} \text{input value } x_{v'} & \text{(if } v' \text{ is an input node)} \\ f_{v'}(x) & \text{(otherwise)} \end{cases}$$

The output of the network is $f_k(x)$, which can be computed sequentially.

25.3 VC-dim of Neural Networks

Definition 25.2 (Hypothesis Class of Neural Networks) *Given architecture of a nueral network, the hypothesis class is,*

$$\mathcal{H} := \{\textit{networks parametrized by } w_E \in \mathbb{R}^{|E|}\}$$

Theorem 25.3 *Let $w = |E|$ (Number of weights), we have*

$$\Pi_{\mathcal{H}}(m) \leq (\frac{emk}{w})^w$$

Hence,

$$\textit{VC-dim}(\mathcal{H}) \leq O(wlogk) \quad (*)$$

Notes: This indicates that the VC-dimension is almost the number parameters $\times$ the number of nodes. Since $k < w$, basically, it is linear in the number of parameters.

Proof: The proof of statement (*) is left as an exercise. Define $S = \{x_1 ... x_m\} \in \mathcal{X}$, i.e. m points in $\mathcal{X}$. Define $\ell = d + k + 1$ to be the total number of nodes in our graph. Then, $f_v(x_i)$ would be the output of node v for i'th point.

Define $D_\ell(S)$ as:

$$D_\ell(S) = \left| \left\{ \begin{bmatrix} f_1(x_1) & \cdots & f_1(x_m) \\ \vdots & \ddots & \vdots \\ f_\ell(x_l) & \cdots & f_\ell(x_m) \end{bmatrix} \text{for all networks parameterized by some } w_E \in \mathbb{R}^{|E|} \right\} \right|$$

■

Claim 25.4 $\Pi_{\mathcal{H}}(m) \leq \max_{|S|=m} D_\ell(S)$

Proof: The left hand side counts all the permutations of the output layer (last row of the matrix) for m points across our hypothesis set. Meanwhile, the right hand side counts all possible permutations of all the neuron outputs for m points across our hypothesis set. So, RHS≤LHS. ■

Claim 25.5 $D_\ell(S) \leq D_{\ell-1}(S)(\frac{em}{d_\ell})^{d_\ell}$, *where $d_\ell = |\{(v', \ell) \in E\}|$ (i.e. the number of incoming edges).*

Proof: Let us split the parameters w_E into two set:

- ℓ's parameters $(w_{\cdot,\ell})$
- Earlier parameters $(w_{v',v}$ s.t. $v' < v < \ell)$.

Let $U = \{u_1, u_2, ...u_{d_v}\}$ be the the nodes that have a directed edge towards ℓ, i.e. $\forall u \in U, (u, \ell) \in E$. Then for all $x_i \in S$,

$$f_\ell(x_i) = sign\Big(\sum_{u \in U} w_{u,\ell} \cdot f_u(x_i)\Big)$$

If we fix all the earlier parameters, then by extension we would also be fixing $f_u(x_i) \quad \forall u \in U$. At this point $f_\ell(x_i)$, would simply be a linear threshold function on d_v-dimensional input vector $\{f_{u_1}(x_i), f_{u_2}(x_i), ...f_{u_{d_v}}(x_i)\}$. This d_v-dimensional input vector is in turn the output of the neural netwrok induced by taking the previous $\ell - 1$ nodes and their parameters. By this argument, we can say that

$$\Pi_\ell(m) \leq \Pi_{\ell-1}(m) \cdot (\frac{em}{d_\ell})^{d_\ell},$$

where $\Pi_\ell(m)$ is the growth function of a network consisting of the first $\ell - 1$ nodes, and $(\frac{em}{d_\ell})^{d_\ell}$ is the result of applying Saeur's Lemma on the aforementioned linear threshold function. Therefore, we can claim via induction that (remember that k is the number of intermediate nodes)

$$\Pi_\ell(m) \leq \prod_{\ell=1}^{k} (\frac{em}{d_\ell})^{d_\ell} \tag{25.1}$$

Claim 25.6 *Given $w = \sum_{\ell=1}^{k} d_\ell$, we can obtain an upper bound on $\Pi_\ell(m)$ by setting $d_\ell = \frac{w}{k} \quad \forall \ell$.*

Proof: Notice that since logarithm is a monotonic function,

$$\arg \max_{\{d_1,...d_k\}} \Pi_\ell(m) = \arg \max_{\{d_1,...d_k\}} \frac{1}{w} \log \Pi_\ell(m) + \log \frac{w}{em}.$$

This allows us to see that

$$\begin{aligned}
\frac{1}{w} \log \Pi_\ell(m) + \log \frac{w}{em} &\leq \frac{1}{w} \log \prod_{\ell=1}^{k} (\frac{em}{d_\ell})^{d_\ell} + \log \frac{w}{em} \\
&= \sum_{\ell=1}^{k} \frac{d_\ell}{w} \log \frac{em}{d_\ell} + \sum_{\ell=1}^{k} \frac{d_\ell}{w} \log \frac{w}{em} \\
&= \sum_{\ell=1}^{k} \frac{d_\ell}{w} \log \frac{w}{d_\ell} \\
&= Entropy(\frac{d_1}{w}, \frac{d_2}{w}, ...\frac{d_k}{w}).
\end{aligned}$$

Entropy is maximized by uniformly distributing the probabilities, and hence $\Pi_\ell(m)$ is maximized by setting $d_\ell = \frac{w}{k} \quad \forall \ell$. ■

Applying this to result to equation (25.1), we get

$$
\begin{aligned}
\Pi_\ell(m) &\leq \prod_{\ell=1}^{k}(\frac{em}{d_\ell})^{d_\ell} \\
&\leq \prod_{\ell=1}^{k}(\frac{em}{\frac{w}{k}})^{\frac{w}{k}} \\
&=(\frac{emk}{w})^{w}.
\end{aligned}
$$

It is left as a reader exercise to prove using this fact that

$$\text{VC-dim}(\mathcal{H}) \leq O(wlogk)$$

■

25.4 VC-Dim of Other Neural Networks

In the above section, we discussed the VC-Dimension of neural networks that use a simple sign function as its activation. Here we examine the VC-dimension in other contexts.

A very common activation function in neural networks is the sigmoid. Let us define a very slightly modified sigmoid function where c is a small constant:

$$\sigma(x) = \frac{1}{1+\exp(-x)} + cx^3\exp(-x^3)\sin x$$

If we define a hypothesis set $\mathcal{H}$ in the following way,

$$\mathcal{H} = \{h : h(x) = w_0 + w_1\sigma(a_1x) + w_2\sigma(a_2x), \quad w_0, w_1, w_2, a_1, a_2 \in \mathbb{R}\}$$

then an interesting property is that $\mathcal{H}$ has VC-Dimension of ∞.

Another interesting property is that if a neural network has d parameters, and performs up to t operations on the input to generate the final output value, then that neural network has VC-Dim of $O(t^2d^2\text{Polylog}(t,d))$. Here, an operation is defined as one of the following:

- addition(+), subtraction(-), multiplication(x), division(÷)
- exponentiation(exp)
- The following indicator functions $(>, <, \geq, \leq, =, \neq)$