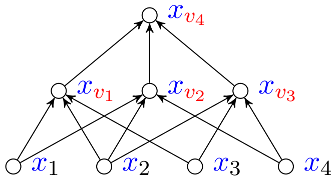
Notes 16: Neural networks

What is the VC dimension of a neural network?
Define **neural network** N as directed acyclic graph G with LTFs at internal nodes



G specifies the network architecture and is fixed
G has n input nodes $1, \dots, n$ and s internal nodes $v_1, \dots, v_s$
Input nodes (those without incoming edges) receive input signals $x_1, \dots, x_n \in \mathbb{R}$
Node/neuron v is **internal** if it has at least one incoming edge
Internal neuron v computes a linear threshold function on its predecessor neurons
$x_v = \mathbb{1}\big(\sum_{u \in \mathrm{Pred}(v)} w_{uv} \cdot x_u \geqslant \theta_v\big)$ where $\mathrm{Pred}(v) = \{\text{predecessors of } v\}$
v is activated (i.e. $x_v = 1$) if the weighted sum of incoming signals exceeds threshold θ_v
When G has a single output node (that has no outgoing edges)
the network N computes a function $f_N : \mathbb{R}^n \to \{0, 1\}$ (given w_{uv} and θ_v)
If learning algorithm A searches for weights and thresholds to minimize training error
A's hypothesis class is $\mathcal{H}_N = \{f_N \mid w_{uv} \in \mathbb{R}, \theta_v \in \mathbb{R}\}$
$\mathrm{VCDim}(\mathcal{H}_N) \leqslant ?$

Will answer this question for a more general class of neural networks:
Redefine neural network N as directed acyclic graph G with concept classes at internal nodes
$\mathcal{C}_j$ over $\mathbb{R}^{\mathrm{Pred}(v_j)}$ is the concept class at internal node v_j
Internal neuron v_j computes $x_{v_j} = \mathbb{1}\big(x_{\mathrm{Pred}(v_j)} \in c_j\big)$ for some $c_j \in \mathcal{C}_j$
Original definition has $\mathcal{C}_j = \{\text{LTFs}\}$ for all v_j; New definition allows other activation functions
Hypothesis class $\mathcal{H}_N = \{f_N \mid c_j \in \mathcal{C}_j\}$ (now $f_N : \mathbb{R}^n \to \{0, 1\}$ implicitly depends on c_j's)

Theorem 1. *Growth function of $\mathcal{H}_N$ is at most the product of growth functions of $\mathcal{C}_j$ over internal nodes $v_1, \dots, v_s$ of G,*

$$\Pi_{\mathcal{H}_N}(m) \leqslant \Pi_{\mathcal{C}_1}(m) \cdots \Pi_{\mathcal{C}_s}(m) \qquad \text{for all } m \in \mathbb{N}$$

Proof. Order internal nodes $v_1, \dots, v_s$ by the order they get evaluated (i.e. topological order)
e.g. in above diagram, v_4 comes after $v_1, \dots, v_3$ because x_{v_4} depends on $x_{v_1}, \dots, v_{v_3}$
Fix m input samples $S = \{x^1, \dots, x^m\}$ where every $x^i \in \mathbb{R}^n$
How many different labelings/dichotomies $T \in \Pi_{\mathcal{H}_N}(S)$ are induced as $c_j \in \mathcal{C}_j$ vary?

Imagine choosing $c_1, \dots, c_s$ sequentially and suppose $c_1, \dots, c_{j-1}$ have been fixed
For every $u \in \mathrm{Pred}(v_j)$, the function $f_u : \mathbb{R}^n \to \mathbb{R}$ of the subnetwork ending at u is fixed
Every sample x^i yields a vector $(f_u(x^i))_{u \in \mathrm{Pred}(v_j)}$ of evaluations of these functions
Call this vector $f_{\mathrm{Pred}(v_j)}(x^i)$; It belongs to $\mathbb{R}^{\mathrm{Pred}(v_j)}$
Collection of these vectors $S_j = \big\{f_{\mathrm{Pred}(v_j)}(x^i) \mid x^i \in S\big\}$ has size $\leqslant m$
Varying c_j may induce different dichotomies $T_j \in \Pi_{\mathcal{C}_j}(S_j)$ on S_j

Choosing all $c_1, \dots, c_s$ yields a labeling T of S, together with a sequence $(T_1, \dots, T_s)$ as above
Distinct labelings T and T' must correspond to different sequences $(T_1, \dots, T_s)$ and $(T'_1, \dots, T'_s)$
Because a sequence $(T_1, \dots, T_s)$ contains enough information to recover T
via computing $f_{v_j}(x^i) = \mathbb{1}\big(f_{\mathrm{Pred}(v_j)}(x^i) \in T_j\big)$ iteratively for $j = 1, \dots, s$
Every T_j is induced by $c_j \in \mathcal{C}_j$ on S_j of size $\leqslant m$ $\implies$ At most $\Pi_{\mathcal{C}_1}(m) \cdots \Pi_{\mathcal{C}_s}(m)$ sequences $\square$

Corollary 2. *If* $\mathrm{VCDim}(\mathcal{C}_j) \leqslant d$ *for all* $1 \leqslant j \leqslant s$*, then* $\mathrm{VCDim}(\mathcal{H}_N) \leqslant 2ds \log(es)$ *when* $s \geqslant 2$

Proof. By above Theorem and Sauer–Shelah lemma, when $m \geqslant d$,

$$\Pi_{\mathcal{H}_N}(m) \leqslant \Pi_{\mathcal{C}_1}(m) \cdots \Pi_{\mathcal{C}_s}(m) \leqslant \left(\left(\frac{em}{d} \right)^d \right)^s$$

$\mathrm{VCDim}(\mathcal{H}_N) < m \quad \iff \quad \Pi_{\mathcal{H}_N}(m) < 2^m$, so we want $\left(\frac{em}{d} \right)^{ds} < 2^m \quad \iff \quad ds \log \left(\frac{em}{d} \right) < m$

How to choose m?

Clearly $m \geqslant ds$ is needed, but then $\log(em/d) \geqslant \log(es)$, so $m \geqslant ds \log(es)$

Turns out $m = 2ds \log(es)$ suffices when $s \geqslant 2$ (exercise) □

Back to original question, if G has fan-in r (i.e. every internal node takes signals from r other nodes)

$\mathrm{VCDim}(\{\text{LTFs over } \mathbb{R}^r\}) = r + 1 \quad \implies \quad \mathrm{VCDim}(\mathcal{H}_N) \leqslant 2(r+1)s\log(es)$

Neural networks in practice typically have internal nodes with real-valued outputs, not just $\{0, 1\}$

Above Theorem does not apply to these networks

The end of Notes15 considers

$$\mathcal{H}_R = \left\{ \mathrm{sign} \left(\sum_{1 \leqslant t \leqslant R} \alpha_t h_t \;\middle|\; \alpha_t \in \mathbb{R}, h_t \in \mathcal{H} \text{ for } 1 \leqslant t \leqslant R \right) \right\}$$

where $\mathcal{H}$ denotes the hypothesis class of weak learner A in AdaBoost

Proposition in Notes15 can be proved using above Theorem and calculations in above Corollary

Question: Which neural network corresponds to $\mathcal{H}_R$? What are the $\mathcal{C}_j$'s?