3.4 VC dimension of linear classifiers and neural networks

- In this section we want to compute or bound the VC dimension of important hypotheses classes

- We start with the class of linear hypotheses on $\mathcal{X} = \mathbb{R}^d$

$$\mathcal{L}_d = \left\{ h_{\mathbf{w},b}(\mathbf{x}) := \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x} + b) \mid \mathbf{w} \in \mathbb{R}^d,\ b \in \mathbb{R} \right\}$$

  and later also consider feedforward neural networks

- As an intermediate step we also consider the subclass of separating hyperplanes $H_{\mathbf{w},0}$ through the origin $\mathbf{0} \in \mathbb{R}^d$

$$\mathcal{L}_d^0 := \left\{ h_{\mathbf{w}}(\mathbf{x}) := \operatorname{sgn}(\mathbf{w} \cdot \mathbf{x}) \mid \mathbf{w} \in \mathbb{R}^d \right\} \subset \mathcal{L}_d$$

VC dimension of linear hypotheses

Lemma 3.34:

A set $M = \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subseteq \mathcal{X}$ is shattered by $\mathcal{L}_d^0$ **if and only if** the vectors $\mathbf{x}_j \in \mathbb{R}^d$, $j = 1, \dots, m$, are linearly independent.

Theorem 3.35:

We have $\mathrm{VCD}(\mathcal{L}_d^0) = d$ as well as $\mathrm{VCD}(\mathcal{L}_d) = d + 1$.

Consequences:

- The sample complexity $m_{\mathcal{L}_d}$ grows only linearly with number of features d

$$m_{\mathcal{L}_d} \in \mathcal{O}(d)$$

- Therefore, linear classifiers are suitable for large number of features $d \gg 1$, e.g., learning classification rules in text analysis

Proof of lemma 3,34:

- let $M = x_1, \dots x_m \subseteq \mathbb{R}^d$ with $x_j; j = 1, \dots, m$ being linearly independent. then: $A = \begin{bmatrix} x_1^T \\ \vdots \\ x_m^T \end{bmatrix} \in \mathbb{R}^{m \times d}$. then the matrix has a rank of (m) which have to be smaller than d to be linearly independent $(\text{Rank}(A) = m \leq d)$ hence for any labeling $b = (b_1, \dots b_m)^T \in \{-1, +1\}^m \in \mathbb{R}^m$ means transpose

we can find a solution $w \in \mathbb{R}^d$ of linear system $Aw = b$ because matrix has full rank

Thus $\therefore$ the hypothesis of $h_b(x) = \text{sgn}(w_b \cdot x)$ yields

$h_b(x_1) = h_1, \ldots h_b, h_b(x_m) = b_m$. Since $b \in \{-1,+1\}^m$ were

arbitrary $\mathcal{L}_d^0$ shatters M. ($|\mathcal{L}_d^0|_M = 2^{|M|}$).

Rank number of independent numbers of colums of matrix

$\Rightarrow$ now let $\mathcal{L}_d^0$ shatter $M = \{x_1, \ldots x_m\}$ we then argue

that the $x_i, j = 1, \ldots m$ can ~~not~~ be linearly independent

by contradiction

if the points of $x_i \in m$ are linearly indpendent then

$\exists\, a = (a_1, \dots, a_m)^T \neq 0$ such that $\underbrace{a_1 x_1 + \dots + a_m x_m = 0}$

according to linear Algebra

then let $I_+ = \{i ; a_i > 0\}$ and

$I_- = \{i : a_i < 0\}$ & then we make an extension:

case 1: $I_+ = \phi$ (null) then let $b \in \{-1, +1\}^m$ with

$b_i = +1$ for $i \in I_-$. this labeling of m cannot be reproduced by a hypothesis $h_w \in \mathcal{L}_d^0$ because

$$0 > \sum_{i \in I_-} a_i \underbrace{(w \cdot x_i)}_{\geq 0 \text{ because } b_i = 0} = w \cdot \Big(\sum_{i \in I_-} a_i x_i\Big)$$

vector here is 0 which is contradiction

because of contradiction, this contradicts that M is shattered by $\mathcal{L}_d^0$.

- Case 2: $I_+ \neq \phi$, then let $b = \{-1, +1\}^m$ with $b_i = +1 \ \forall i \in I_+$ & $b_i = -1 \ \forall i \in I_-$. again this labeling of M can't be reproduced by a $h_w \in \mathcal{L}_d^0$, because then

$$0 < \sum_{i \in I_+} \underbrace{a_i}_{>0} \underbrace{(w \cdot x_i)}_{>0} = w \cdot \underbrace{\left(\sum_{i \in I_+} a_i x_i \right)}_{= -\sum_{i \in I_-} a_i x_i}$$

$$= -\sum_{I \in I_-} \underbrace{a_i}_{<0} \underbrace{(w x_i)}_{<0} < 0 \quad \text{a contradiction.}$$

this contradicts that M is shattered by L_d

Thus $x_1, \dots x_m$ can't be linearly independent.

otherwise, we do have $\mathcal{L}$ that can't be reproduced.

Proof of Th 3.33:

VCD($\mathcal{H}^{0}_{d}$) $= d$ follows by Lemma 3.34 & VCD($\mathcal{L}_d$) $=$ $d+1$ which can be showed by lemma 3.34 as well. To the end note that any $h_{w,b} \in \mathcal{L}_d$,

$h_{w,b}(x) = \text{Sgn}(b + wx)$ which correspondes to

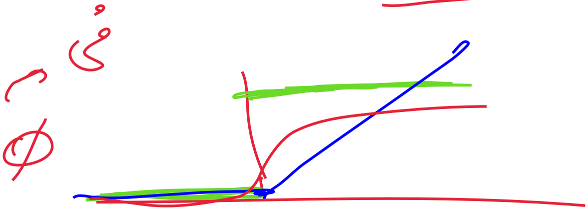
$h_{w,b}(x) = h_{(w,b)}\left((x,1)^T\right)$, $h_{(w,b)} \in \mathcal{H}^{0}_{d+1}$

thus VCD($\mathcal{L}_d$) $\leq$ VCD($\mathcal{L}^{0}_{d+1}$) $= d+1$, since $\mathcal{L}_d$ shatter any set $M = \{x_1, \dots x_m\} \cup \{0\} \subseteq \mathbb{R}^d$

with linearly independable $+ x_j \in R^d$ because

$A = \begin{bmatrix} x_1^T & 1 \\ \vdots & \vdots \\ x_m^T & 1 \end{bmatrix}$ has full rank $(A) = m+1$ then

we set $VC(\mathcal{H}_d) = d+1$

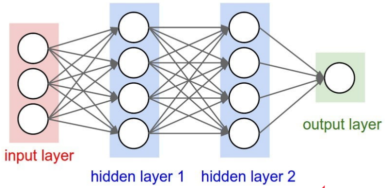


blue is Relu function
red is Sigmoid function
green is Sign "

Sign function is better in terms of classification
better than ReLu & Sigmoid since the start $[0, \infty)$
$\rightarrow$ all for output such Softmax

Repetition: Feedforward neural networks (FNN)

- Recall a FNN consists of L layers of n_k neurons processing and passing information from layer to layer



- Each neuron $v_{k,i}$, $k = 1, \dots, L$, $i = 1, \dots, n_k$ is a linear hypothesis

$$y_{k,i} = v_{k,i}(\mathbf{y}_{k-1}) := \phi\left(\sum_{j=1}^{n} w_j \; y_{k-1,j} + b\right)$$

layers

with activation function $\phi\colon \mathbb{R} \to \mathbb{R}$.

collects of weights in layers

- The output of the kth layer $V_k = \{v_{k,1}, \dots, v_{k,n_k}\}$ can then be written by

$$\mathbf{y}_k = \phi \circ f_{\mathbf{W}_k,\mathbf{b}_k}(\mathbf{y}_{k-1}), \qquad f_{\mathbf{W}_k,\mathbf{b}_k}(\mathbf{y}) := \mathbf{W}_k \; \mathbf{y} + \mathbf{b}_k$$

where ϕ is applied componentwise and we introduced the layerwise weight matrices $\mathbf{W}_k \in \mathbb{R}^{n_k \times n_{k-1}}$ and bias vectors $\mathbf{b}_k \in \mathbb{R}^{n_k}$

- The whole neural network is then a hypothesis $h\colon \mathcal{X} \to \mathcal{Y}$ of the form

$$h(\mathbf{x}) = \rho \circ f_{\mathbf{W}_L,\mathbf{b}_L} \circ \phi \circ f_{\mathbf{W}_{L-1},\mathbf{b}_{L-1}} \circ \phi \circ \cdots \circ \phi \circ f_{\mathbf{W}_1,\mathbf{b}_1}(\mathbf{x}),$$

where $\phi\colon \mathbb{R} \to \mathbb{R}$ as well as $\rho\colon \mathbb{R} \to \mathcal{Y}$ are the chosen activation functions

- Given a certain architecture (V, E) with $V = (V_0, \ldots, V_L)$ consisting of
  - collection of layers $V = (V_0, \ldots, V_L)$ where $V_k = \{v_{k,1}, \ldots, v_{k,n_k}\}$ and $|V_0| = d$, $|V_L| = 1$,
  - collection of communication edges between adjacent layers:

$$E \subseteq \{(v_{k,i}, v_{k+1,j})\colon v_{k,i} \in V_k \text{ and } v_{k+1,j} \in V_{k+1}\}.$$

and chosen activation functions $\phi\colon \mathbb{R} \to \mathbb{R}$, $\rho\colon \mathbb{R} \to \mathcal{Y}$ we introduce the class of all FNN $h\colon \mathcal{X} \to \mathcal{Y}$ with just this architecture

Output activation function

$$\mathcal{H}_{V,E,\phi,\rho} = \{[\rho \circ f_{\mathbf{W}_L,\mathbf{b}_L}] \circ [\phi \circ f_{\mathbf{W}_{L-1},\mathbf{b}_{L-1}}] \circ \cdots \circ [\phi \circ f_{\mathbf{W}_1,\mathbf{b}_1}] : \mathbf{W}_k \in \mathbb{R}^{n_k \times n_{k-1}}, \mathbf{b} \in \mathbb{R}^{n_k},$$
$$[\mathbf{W}_k]_{i,j} \neq 0 \text{ iff } (v_{k,i}, v_{k+1,j}) \in E\}$$

- If $\phi = \rho$, we write only $\mathcal{H}_{V,E,\phi}$.

The VC dimension of neural networks

which need to be trained

Theorem 3.36:

Let $p_{V,E} = \sum_{k=1}^{L} n_k + |E|$ denote the number of parameters of the hypothesis class $\mathcal{H}_{V,E,\mathrm{sgn}}$. We have

sum number of bias

larger might lead to overfitting

$$\mathrm{VCD}\left(\mathcal{H}_{V,E,\mathrm{sgn}}\right) \in \mathcal{O}(p_{V,E}\ \ln(p_{V,E})).$$

we never use this in practice

- FNN with sgn as activation function are **PAC-learnable**, but the learnability decreases with the size of the network

- Lower bounds on $\mathrm{VCD}\left(\mathcal{H}_{V,E,\mathrm{sgn}}\right)$ can also be proved, as well as the VC dimension for other choices of σ and $\rho = \mathrm{sgn}$

	$\mathrm{VCD}(\mathcal{H}_{V,E,\sigma,\mathrm{sgn}})$	
	Lower bound	Upper bound
sign	$\Omega(p \ln p)$	$\mathcal{O}(p \ln p)$
sigmoid	$\Omega(\lvert E\rvert^2)$	$\mathcal{O}(p^2)$
ReLU	$\Omega(L\ p \ln(p/L))$	$\mathcal{O}(L\ p \ln p)$

- To prove Theorem 3.36 we exploit the special structure of $\mathcal{H} := \mathcal{H}_{V,E,\mathrm{sgn}}$

- Let $V = (V_0, \dots, V_L)$ with $n_k = |V_k|$ then

$$\mathcal{H} = \mathcal{H}_L \circ \cdots \circ \mathcal{H}_1, \qquad \mathcal{H}_k = \mathcal{L}_{d_{k,1}} \times \cdots \times \mathcal{L}_{d_{k,n_k}},$$

VCD $(\mathcal{L}_{d_{k,1}}) = d_k + 1$

i.e., $\mathcal{H}_k \subset \{h \colon \mathbb{R}^{n_{k-1}} \to \{-1,+1\}^{n_k}\}$ and $d_{k,j} \leq n_{k-1}$ denotes the number of incoming edges at node $v_{k,j}$

- We then can use the following:

Proposition 3.37:

Let either

1. $\mathcal{H} := \mathcal{H}_2 \circ \mathcal{H}_1$ given $\mathcal{H}_1 \subseteq \mathcal{Y}^{\mathcal{X}}$ and $\mathcal{H}_2 \subseteq \mathcal{Z}^{\mathcal{Y}}$,
2. or $\mathcal{H} := \mathcal{H}_1 \times \mathcal{H}_2$ given $\mathcal{H}_i \subseteq \mathcal{Y}_i^{\mathcal{X}}$ for $i = 1, 2$.

Then we have

$$\underline{\tau_{\mathcal{H}}(m) \leq \tau_{\mathcal{H}_1}(m) \cdot \tau_{\mathcal{H}_2}(m)}, \qquad m \in \mathbb{N}.$$

growth function

Proof of Lemma 3.37:

let $H = H_1 \circ H_2$ then for any set $M = \{x_1, \dots x_m\} \subset X$

we have $|H_M| = |\{[h(x_1), \dots h(x_m)] : h \in H\}|$

$$= |\{[h_2(h_1(x_1)), \dots, h_2(h_1(x_m))] : h_1 \in H_1,\ h_2 \in H_2\}|$$

$$= \Big|\bigcup_{y \in H_{1,M}} \underbrace{[h_2(y_1), \dots, h_2(y_m)] : h_2 \in H_2}_{\tau_{H_2}(m)}\Big|$$

$$\le |H_{1,M}| \cdot \tau_{H_2}(m) \le \underline{\tau_{H_1}(m) \cdot \tau_{H_2}(m)}$$

or

$$\boxed{\sup_{M \subset Z, |M| = m} |H_M| \le \tau_{H_1}(m) \circ \tau_{H_2}(m)}$$

#

Same reasoning can be applied to:

$$\mathcal{H} = H_1 \rtimes H_2 = \left\{ h = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} ; h_1 \in H_1, h_2 \in H_2 \right\}$$