Neural Network Learning: Theoretical Foundation Chap. 6-7

김성현

서울대학교 통계학과

2017년 7월 22일

Outline

Outline

Feed-Forward Neural Network

- Feed-forward neural network consists of linear combinations and activation functions.
- Feed-Forward Condition: Connections do not form any loops.
- The units can be labelled with integers in such a way that if there is a connection from the unit labelled i to the computation unit labelled j then i < j .
- A feed-forward network is said to be a linear threshold network if each activation function is *sgn*.

Feed-Forward Neural Network

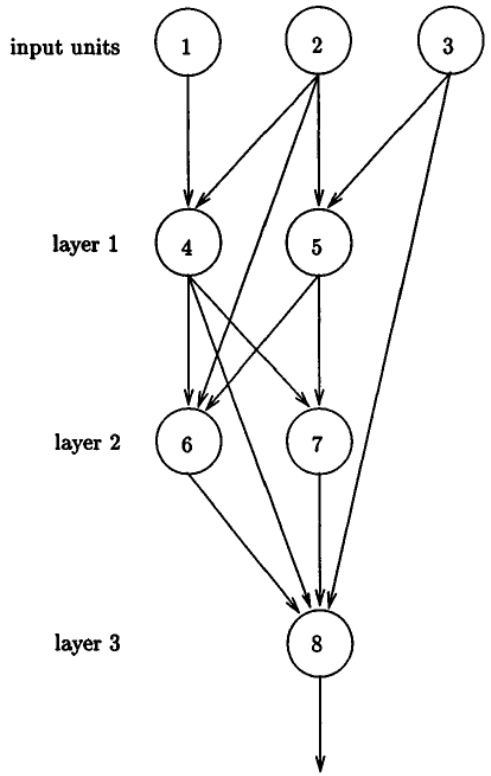

Figure: Feed-Forward Neural Network

Outline

Theorem 6.1

Suppose that N is a feed-forward linear threshold network having a total of W variable weights and thresholds, and k computation units. Let H be the class of functions computable by N on real inputs. Then for $m \geq W$ the growth function of H satisfies

$$\Pi_H(m) \leq \left(\frac{emk}{W}\right)^W,$$

and hence $VCdim(H) < 2Wlog_2(2k/ln2)$.

Outline

Theorem 6.2

Let N be a two-layer linear threshold network, fully connected between adjacent layers, with $n \geq 3$ input units, k computation units in the first layer (and one output unit in the second layer). Suppose that $k \leq 2^{n+1}/(n^2 + n + 2)$. Then the class H of functions computable by N on binary inputs is such that

$$VCdim(H) \geq nk + 1 \geq 3W/5,$$

where $W = nk + 2k + 1$ is the total number of weights and thresholds.

Theorem 6.3

Let W be any positive integer greater than 32. Then there is a three-layer feed-forward linear threshold network N_W with at most W weights and thresholds, for which the following holds. If H is the class of functions computable by N_w on binary inputs, then

$$VCdim(H) > \frac{W}{132} log_2 \frac{k}{16},$$

where k is the number of computation units.

Theorem 6.4

Let N be a two-layer feed-forward linear threshold network, fully connected between adjacent layers, having k computation units and $n \geq 3$ inputs, where $k \leq 2^{n/2-2}$. Let H be the set of functions computable by N on $\mathbb{R}^n$. Then

$$VCdim(H) \geq \frac{nk}{8} log_2 \frac{k}{4} \geq \frac{W}{32} log_2 \frac{k}{4},$$

where $W = nk + 2k + 1$ is the total number of weights and thresholds

Outline

Theorem 6.5

Suppose $s : \mathbb{R} \to \mathbb{R}$ satisfies $lim_{\alpha \to \infty} s(\alpha) = 1$ and $lim_{\alpha \to -\infty} s(\alpha) = 0$. Let N be a feed-forward linear threshold network, and N' a network with the same structure as N, but with the threshold activation functions replaced by the activation function s in all non-output computation units. Suppose that S is any finite set of input patterns. Then, any function computable by N on S is also computable by N'.

Theorem 6.5

Hence the lower bound results Theorem 6.2, Theorem 6.3 and Theorem 6.4 also hold for such networks, and in particular for standard sigmoid networks.

Outline

Theorem 7.1

Define

$$s(x) = \frac{1}{1+e^{-x}} + cx^3 e^{-x^2} \sin x$$

for $c > 0$. Then $s(\cdot)$ is analytic, and for any sufficiently small $c > 0$, we have

$$lim_{x \to \infty} s(x) = 1,$$

$$lim_{x \to -\infty} s(x) = 0,$$

$$\frac{d^2}{dx^2} s(x) \begin{cases} < 0 & \text{if x>0} \\ > 0 & \text{if x<0} \end{cases}$$

Theorem 7.1 (Conti.)

Let N be a two-layer network with one real input, two first-layer computation units using this activation function, and one output unit, so that functions in H_N are of the form

$$x \mapsto sgn(w_0 + w_1 s(a_1 x) + w_2 s(a_2 x)),$$

with $x, w_0, w_1, w_2, a_1, a_2 \in \mathbb{R}$. Then $VCdim(H_N) = \infty$.

Lemma 7.2

The class $F = \{x \mapsto sgn(sin(ax)) : a \in \mathbb{R}^+\}$ of functions defined on $\mathbb{N}$ has $VCdim(F) = \infty$.

- Now, let

$$h_a(x) = s(ax) + s(-ax) - 1 = 2c(ax)^3 e^{-a^2x^2} sin(ax).$$

For $a > 0$ and $x > 0$, $sgn(h_a(x)) = sgn(sin(ax))$, so Lemma 7.2 implies that $VCdim(H_N) = \infty$.

Outline

Definition 7.3

Let H be a class of $\{0,1\}$-valued functions defined on a set X, and F a class of real-valued functions defined on $\mathbb{R}^d \times X$. We say that H is a k-combination of $sgn(F)$ if there is a boolean function $g : \{0,1\}^k \to \{0,1\}$ and functions $f_1, ..., f_k$ in F so that for all h in H there is a parameter vector $a \in \mathbb{R}^d$ such that

$$h(x) = g(sgn(f_1(a,x)), ..., sgn(f_k(a,x)))$$

for all x in X.

Definition 7.3(Conti.)

We say that a function f in F is continuous in its parameters (C^p in its parameters) if, for all x in X, $f(\cdot, x)$ is continuous (respectively, C_p).

Definition 7.4

A set $\{f_1, ..., f_k\}$ of differentiable functions mapping from $\mathbb{R}^d$ to $\mathbb{R}$ is said to have regular zero-set intersections if, for all nonempty subsets $\{i_1, ..., i_l\} \subset \{1, ..., k\}$, the Jacobian of $(f_{i_1}, ..., f_{i_l}) : \mathbb{R}^d \to \mathbb{R}^l$ has rank l at every point a of the solution set

$$\{a \in \mathbb{R}^d : f_{i_1}(a) = ... = f_{i_l}(a) = 0\}.$$

Definition 7.4 (example)

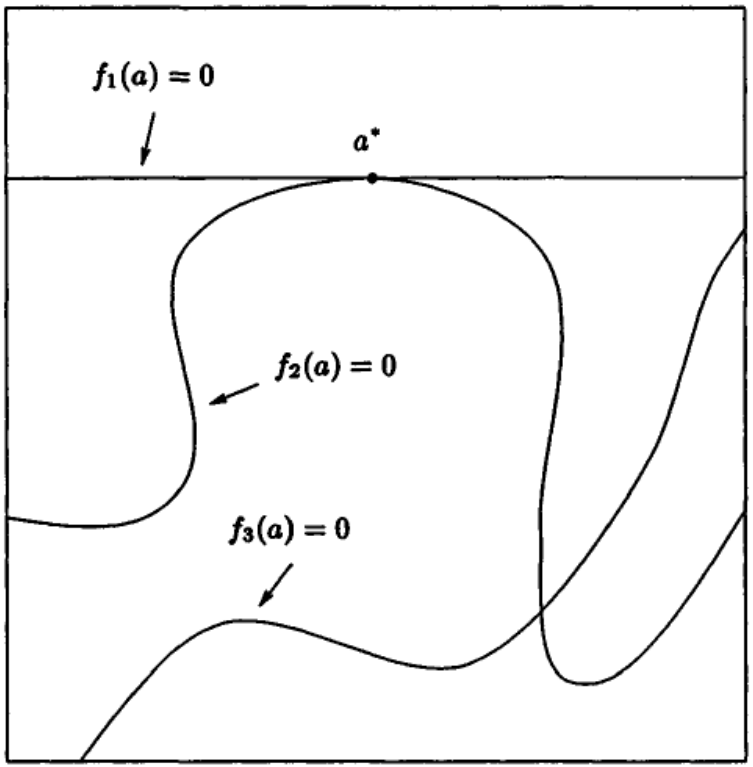

Figure: An example illustrating Definition 7.4

Definition 7.5

Let G be a set of real-valued functions defined on $\mathbb{R}^d$. We say that G has solution set components bound B if for any $1 \leq k \leq d$ and any $\{f_1, ..., f_k\} \subseteq G$ that has regular zero-set intersections, we have

$$CC\left(\bigcap_{i=1}^{k}\{a \in \mathbb{R}^d : f_i(a) = 0\}\right) \leq B.$$

Theorem 7.6

Suppose that F is a class of real-valued functions defined on $\mathbb{R}^d \times X$, and that H is a k-combination of $sgn(F)$. If F is closed under addition of constants, has solution set components bound B, and functions in F are C^d in their parameters, then

$$\Pi_H(m) \leq B \sum_{i=0}^{d} \binom{mk}{i} \leq B \left(\frac{emk}{d} \right)^d,$$

for $m \geq d/k$.

Outline

Lemma 7.7

Given a set $\{f_1, ..., f_k\}$ of C^d functions that map from $\mathbb{R}^d$ to $\mathbb{R}$, the set

$S = \{\lambda \in \mathbb{R}^k : \{f_1 - \lambda_1, ..., f_k - \lambda_k\}$ does not have regular zero-set intersections $\}$

has measure 0.

Lemma 7.8

Let F be a class of real-valued functions defined on $\mathbb{R}^d \times X$ that is closed under addition of constants. Suppose that the functions in F are continuous in their parameters and let H be a k-combination of $sgn(F)$. Then for some functions $f_1, ..., f_k$ in F and some examples $x_i, ..., x_m$ in X, the set

$$\{a \mapsto f_i(a, x_j) : i = 1, ..., k, j = 1, ..., m\}$$

has regular zero-set intersections

Lemma 7.8 (Conti.)

and the number of connected components of the set

$$\mathbb{R}^d - \bigcup_{i=1}^{k} \bigcup_{j=1}^{m} \{a \in \mathbb{R}^d : f_i(a, x_j) = 0\}$$

is at least $\Pi_H(m)$.

Lemma 7.9

Let $\{f_1, ..., f_k\}$ be a set of differentiable functions that map from $\mathbb{R}^d$ to $\mathbb{R}$, with regular zero-set intersections. For each i, define Z_i to be the zero-set of $f_i : Z_i = \{a \in \mathbb{R}^d : f_i(a) = 0\}$. Then

$$cc\left(\mathbb{R}^d - \bigcup_{i=1}^{k} Z_i\right) \leq \sum_{S \subseteq \{1,...,k\}} cc\left(\bigcap_{i \in S} Z_i\right)$$

Lemma 7.10

Define a set of functions $\{f_1, ..., f_k\}$ as in Lemma 7.9, and define sets $S_1, ..., S_{k-1}$ so that for $i = 1, ..., k-1$, either $S_i = \{a \in \mathbb{R}^d : f_i(a) = 0\}$ or $S_i = \{a \in \mathbb{R}^d : f_i(a) \neq 0\}$. Let C be a connected component of $\cap_{i=1}^{k-1} S_i$, and let C' be a connected component of $C \cap a \in \mathbb{R}^d : f_k(a) = 0$. Then $C - C'$ has no more than two connected components.

Lemma 7.11

Define a set of functions $\{f_1, ..., f_k\}$ and the zero-sets $Z_1, ..., Z_k$ as in Lemma 7.9. Let $I \subseteq \{I, ..., k\}$ and define $M = \cap_{i \in I} Z_i$. Define $b = k - |I|$ and let $\{M_1 ..., M_b\} = \{Z_i : i \notin I\}$. Then

$$CC\left(M - \bigcup_{j=1}^{b} M_j\right) \leq CC\left(M - \bigcup_{j=1}^{b-1} M_j\right) + CC\left(M \cap M_b - \bigcup_{j=1}^{b-1} M_j\right)$$

.

Outline

Intro

If, while computing $f(a)$, we calculate b_1, then b_2 and so on up to b_n, and we use these to calculate $f(a)$, then we can write f in the form

$$f(a) = \tilde{f}(a, b_1, ..., b_n)$$

where each b_i is a function only of a and $b_1, ..., b_{i-1}$.

Intro

In this definition, an intermediate calculation is expressed as $b = \phi(a)$ for some function ϕ. To ensure that the intermediate variables take the appropriate values, we use the trick of defining a function g for which the constraint $b = \phi(a)$ is satisfied when $g(a, b) = 0$.

Definition 7.12

For a set G of differentiable real-valued functions defined on $\mathbb{R}^d$ and a set G of differentiable real-valued functions defined on $\mathbb{R}^{d(n+1)}$, we say that $\tilde{G}$ computes G with n intermediate variables if, for any $1 \leq k \leq d$ and $f_1, ..., f_k \subseteq G$, there is a set

$$\left\{ \tilde{f}_1, g_{1,1}, ..., g_{1,n}, ..., \tilde{f}_k, g_{k,1}, ..., g_{k,n} \right\} \subseteq \tilde{(G)}$$

that satisfies the following conditions

Definition 7.12 (Conti.)

(i) For $i = 1, ..., k$, there are differentiable functions $\phi_{i,1}, ..., \phi_{i,n} : \mathbb{R}^{d(n+1)} \to \mathbb{R}$ which can be written

$$\phi_{i,1}(a, b) = \phi_{i,1}(a)$$

$$\phi_{i,j}(a, b) = \phi_{i,j}(a, b_{i,1}, ..., b_{i,j-1}) \quad for \quad j = 2, ..., n$$

where $a \in \mathbb{R}^d$, and $b = (b_{1,1}, ..., b_{1,n}, ..., b_{d,n}) \in \mathbb{R}^{dn}$. (The function $\phi_{i,j}$ defines the intermediate variable $b_{i,j}$, and the $\phi_{i,j}$ are ordered so that their values depend only on previously computed intermediate variables.)

Definition 7.12 (Conti.)

(ii) For $i = 1, ..., k$, the functions $g_{i,1}, ..., g_{i,n}$ can be written as $g_{i,j}(a, b) = g_{i,j}(a, b_{i,1}, ..., b_{i,j})$ for all a, b, and $j = l, ..., n$. (That is, the function $g_{i,j}$ depends only on previously computed intermediate variables, and on $b_{i,j}$.)

Definition 7.12 (Conti.)

(iii) For $i = 1, ..., k$, and $l = 1, ..., n$, if $b_{i,j} = \phi_{i,j}(a, b)$ for all a, b, and $j < l$, then for all a and b we have

$$g_{i,l}(a, b) = 0 \textit{ if and only if } b_{i,l} = \phi_{i,l}(a, b)$$

and

$$D_{b_{i,l}} g_{i,l}(a, \phi_{i,1}(a, b), ..., \phi_{i,l}(a, b)) \neq 0$$

(That is, the function $g_{i,j}$ implicitly defines the intermediate variable $b_{i,j}$. The derivative condition ensures that the zero-set intersections remain regular.)

Definition 7.12 (Conti.)

(iv) For all $a \in: \mathbb{R}^d$ and $b \in: \mathbb{R}^{dn}$, if $b_{i,j} = \phi_{i,j}(a, b)$ for $i = 1, ..., k$ and $j = 1, ..., n$ then

$$f_i(a) = \tilde{f}_i(a, \phi_{1,1}(a, b), ..., \phi_{k,n}(a, b))$$

for $i = 1, ..., n$.

Theorem 7.13

For function classes G and $\tilde{G}$ and a positive integer n, if $\tilde{G}$ computes G with n intermediate variables, then any solution set components bound for $\tilde{G}$ is also a solution set components bound for G.