Neural Network Learning: Theoretical Foundations

Chap.8, 9

Martin Anthony and Peter L. Bartlett

Presenter: Sarah Kim

2017.07.29

Contents

Contents

Reviews

- 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 $\{f_1, \ldots f_k\} \subseteq G$ that has regular zero-set intersetions, we have

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

- 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 $\mathrm{sgn}(F)$. 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

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

for $m \geq d/k$.

8.2 Function Classes that are Polynomial in their Parameters

- Consider classes of functions that can be expressed as boolean combinations of thresholded real-valued functions, each of which is polynomial in its parameters.
- Lemma 8.1 Suppose $f: \mathbb{R}^d \to \mathbb{R}$ is a polynomial of degree l. Then the number of connected components of $\{a \in \mathbb{R}^d : f(a) = 0\}$ is no more than $l^{d-1}(l+2)$.
- Corollary 8.2 For $l \in \mathbb{N}$, the set of degree l polynomials defined on $\mathbb{R}^d$ has solution set components bound $B = 2(2l)^d$.

- Theorem 8.3 Let F be a class of functions mapping from $\mathbb{R}^d \times X$ to $\mathbb{R}$ so that, for all $x \in X$ and $f \in F$, the function $a \mapsto f(a, x)$ is a polynomial on $\mathbb{R}^d$ of degree no more than l. Suppose that H is a k-combination of $\mathrm{sgn}(F)$. Then if $m \geq d/k$,

$$\Pi_H(m) \leq 2\left(\frac{2emkl}{d}\right)^d,$$

and hence $\mathrm{VCdim}(H) \leq 2d \log_2(12kl)$.

- Theorem 8.4 Suppose h is a function from $\mathbb{R}^d \times \mathbb{R}^n$ to $\{0,1\}$ and let

$$H = \{x \mapsto h(a,x) : a \in \mathbb{R}^d\}$$

be the class determined by h. Suppose that h can be compoted by an algorithm that takes as input the pair $(a,x) \in \mathbb{R}^d \times \mathbb{R}^n$ and returns $h(a,x)$ after no more than t operations of the following types:

  - the arithmetic operations $+, -, \times$, and $/$ on real numbers,
  - jumps conditioned on $>, \geq, <, \leq, =$, and $\neq$ comparisions of real numbers, and
  - output 0 or 1.

  Then $\mathrm{VCdim}(H) \leq 4d(t+2)$.

- Theorem 8.5 For all $d, t \geq 1$, there is a class H of functions, parametrized by d real numbers, that can be computed in time $O(t)$ using the model of computation defined in Thoerem 8.4, and that has $\mathrm{VCdim}(H) \geq dt$.

8.3 Piecewise-Polynomial Networks

- Theorem 8.6 Suppose N is a feed-forward linear threshold network with a total of W weights, and let H be the class of functions computed by this network. Then $\mathrm{VCdim}(H) = O(W^2)$.

- This theorem can easily be generalized to network with piecewise-polynomial activation functions. A piecewise-polynomial function $f : \mathbb{R} \to \mathbb{R}$ can be written as $f(\alpha) = \sum_{i=1}^{p} 1_{A(i)}(\alpha) f_i(\alpha)$, where $A(1), \ldots, A(p)$ are disjoint real intervals whose union is $\mathbb{R}$, and $f_1, \ldots, f_p$ are polynomials. Define the degree of f as the largest degree of the polynomials f_i.

- Theorem 8.7 Suppose N is a feed-forward network with a total of W weights and k computation units, in which the output unit is a linear threshold unit and every other computation unit has a piecewise-polynomial activation function with p pieces and degree no more than l. Then, if H is the class of functions computed by N, $\mathrm{VCdim}(H) = O(W(W + kl \log_2 p))$.

- Theorem 8.8 Suppose N is a feed-forward network of the form described in Theorem 8.7, with W weights, k computation units, and all non-output units having piecewise-polynomial activation functions with p pieces and degree no more than l. Suppose in addition that the computation units in the network are arranged in L layers, so that each unit has connections only from units in earlier layers. Then if H is the class of functions computed by N,

$$\Pi_H(m) \leq 2^L (2emkp(l+1)^{L-1})^{WL},$$

and

$$\mathrm{VCdim}(H) \leq 2WL \log_2(4WLpk/\ln 2) + 2WL^2 \log_2(l+1) + 2L.$$

For fixed p, l, $\mathrm{VCdim}(H) = O(WL \log_2 W + WL^2)$.

- Theorem 8.9 Suppose $s : \mathbb{R} \to \mathbb{R}$ has the following properties:
  1. $\lim_{\alpha \to \infty} s(\alpha) = 1$ and $\lim_{\alpha \to -\infty} s(\alpha) = 0$, and
  2. s is differentiable at some point $\alpha_0 \in \mathbb{R}$, with $s'(\alpha_0) \neq 0$.

  For any $L \geq 1$ and $W \geq 10L - 14$, there is a feed-forward network with L layers and a total of W parameters, where every computation unit but the output unit has activation function s, the output unit being a linear threshold unit, and for which the set H of functions computed by the network has

$$\text{VCdim}(H) \geq \left\lfloor \frac{L}{2} \right\rfloor \left\lfloor \frac{W}{2} \right\rfloor$$

8.4 Standard Sigmoid Networks

Discrete inputs and bounded fan-in

- Consider networks with the standard sigmoid activation, $\sigma(\alpha) = 1/(1 + e^{-\alpha})$.
- We define the fan-in of a computation unit to be the number of input units or computation units that feed into it.
- Theorem 8.11 Consider a two-layer feed-forward network with input domain $X = \{-D, -D + 1, \dots, D\}^n$ (for $D \in \mathbb{N}$) and k first-layer computation units, each with the standard sigmoid activation function. Let W be the total number of parameters in the network, and suppose that the fan-in of each first-layer unit is no more than N. Then the class H of functions computed by this network has $\text{VCdim}(H) \leq 2W \log_2(60ND)$.

- Theorem 8.12 Consider a two-layer feed-forward linear threshold network that has W parameters and whose first-layer units have fan-in no more than N. If H is the set of functions computed by this network on binary inputs, then $\mathrm{VCdim}(H) \leq 2W \log_2(60N)$. Furthermore, there is a constant c s.t. for all W there is a network with W parameters that has $\mathrm{VCdim}(H) \geq cW$.

General standard sigmoid networks

- Theorem 8.13 Let H be the set of functions computed by a feed-forward network with W parameters and k computation units, in which each computation unit other than the output unit has the standard sigmoid activation function (the output unit being a linear threshold unit). Then

$$\Pi_H(m) \leq 2^{(Wk)^2/2}(18Wk^2)^{5Wk}\left(\frac{em}{W}\right)^W$$

probided $m \geq W$, and

$$\mathrm{VCdim}(H) \leq (Wk)^2 + 11Wk\log_2(18Wk^2).$$

- Theorem 8.14 Let h be a function from $\mathbb{R}^d \times \mathbb{R}^n$ to $\{0, 1\}$, determining the class

$$H = \{x \mapsto h(a, x) : a \in \mathbb{R}^d\}.$$

Suppose that h can be computed by an algorithm that takes as input the pair $(a, x) \in \mathbb{R}^d \times \mathbb{R}^n$ and returns $h(a, x)$ after no more than t of the following oprations:

  - the exponential function $\alpha \mapsto e^\alpha$ on real numbers,
  - the arithmetic operations $+, -, \times$, and $/$ on real numbers,
  - jumps conditioned on $>, \geq, <, \leq, =$, and $\neq$ comparisions of real numbers, and
  - output 0 or 1.

Then $\mathrm{VCdim}(H) \leq t^2 d(d + 19 \log 2(9d))$. Furthermore, if the t steps include no more than q in which the exponential function is evaluated, then

$$\Pi_H(m) \leq 2^{(d(q+1))^2/2} (9d(q+1)2^t)^{5d(q+1)} \left(\frac{em(2^t - 2)}{d}\right)^d,$$

and hence $\mathrm{VCdim}(H) \leq (d(q+1))^2 + 11d(q+1)(t + \log_2(9d(q+1)))$.

Proof of VC-dimension bounds for sigmoid networks and algorithms

- Lemma 8.15 Let $f_1, \ldots, f_q$ be fixed affine functions of $a_1, \ldots, a_d$, and let G be the class of polynomials in $a_1, \ldots, a_d, e^{f_1(a)}, \ldots, e^{f_q(a)}$ of degree no more than l. Then G has solution set components bound

$$B = 2^{q(q-1)/2}(l+1)^{2d+q}(d+1)^{d+2q}.$$

- Lemma 8.16 Suppose G is the class of functions defined on $\mathbb{R}^d$ computed by a circuit satisfying the following conditions: the circuit contains q gates, the output gate computes a rational function of degree no more than $l \geq 1$, each non-output gate computes the exponential function of a rational function of degree no more than l, and the denominator of each rational function is never zero. Then G has solution set components bound $2^{(qd)^2/2}(9qdl)^{5qd}$.

Contents

9.2 Large Margin Classifiers

- Suppose F is a class of functions defined on the set X and mapping to the interval $[0, 1]$.

- Definition 9.1 Let $Z = X \times \{0, 1\}$. If f is a real-valued function in F, the margin of f on $(x, y) \in Z$ is

$$\mathrm{margin}(f(x), y) = \begin{cases} f(x) - 1/2 & \text{if } y = 1 \\ 1/2 - f(x) & \text{otherwise.} \end{cases}$$

  Suppose γ is a nonnegative real number and P is a probability distribution on Z. We define the error $er_P^\gamma(f)$ of f w.r.t. P and γ as the probability

$$er_P^\gamma(f) = P\{\mathrm{margin}(f(x), y) < \gamma\},$$

  and the misclassification probability of f as

$$er_P(f) = P\{\mathrm{sgn}(f(x) - 1/2) \neq y\}.$$

- Definition 9.2 A classification learning algorithm L for F takes as input a margin parameter $\gamma > 0$ and a sample $z \in \bigcup_{i=1}^{\infty} Z^i$, and returns a function in F s.t., for any $\epsilon, \delta \in (0,1)$ and any $\gamma > 0$, there is an integer $m_0(\epsilon, \delta, \gamma)$ s.t. if $m \geq m_0(\epsilon, \delta, \gamma)$ then, for any probability distribution P on $Z = X \times \{0,1\}$,

$$P^m \left\{ er_P(L(\gamma, z)) < \inf_{g \in F} er_P^{\gamma}(g) + \epsilon \right\} \geq 1 - \delta.$$

- Sample error of f w.r.t. γ on the sample z :

$$\hat{er}_z^{\gamma}(f) = \frac{1}{m} |\{ i : \text{margin}(f(x_i), y_i) < \gamma \}|$$

- Proposition 9.3 For any function $f\colon X \to \mathbb{R}$ and any sequence of labelled examples $((x_1, y_1), \ldots, (x_m, y_m))$ in $(X \times \{0,1\})^m$, if

$$\frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2 < \epsilon$$

then

$$\hat{er}_z^{\gamma}(f) < \epsilon/(1/2 - \gamma)^2$$

for all $0 \leq \gamma < 1/2$.