Covering Numbers

Definition Let (A, d) be a metric space. Given $W \subset A$ and a positive number ϵ, a subset $C \subset W$ is called a ϵ-cover of W if for any $w \in W$, there is $c \in C$ such that $d(w, c) < \epsilon$.

Definition A ϵ-covering number of W denoted by $\mathcal{N}(\epsilon, W, d)$, is the minimal cardinality of an ϵ-cover of W.

Definition Let F be a set of functions from a domain X and let k be a positive integer. An uniform ϵ-covering number is defined as

$$\mathcal{N}_\infty(\epsilon, F, k) = \max\{\mathcal{N}(\epsilon, F_{|x}, d_\infty) : x \in X^k\}.$$

The Pseudo Dimension

Definition 11.1 Let F be a set of real-valued functions mapping from a domain X and suppose that $S = \{x_1, x_2, \ldots, x_m\} \subseteq X$. Then S is pseudo-shattered by F if there are real number $r_1, r_2, \ldots, r_m$ such that for each $b \in \{0, 1\}^m$ there is a function $f_b \in F$ with $\text{sign}(f_b(x_i) - r_i) = b_i$ for $1 \leq i \leq m$. We say that $r = (r_1, r_2, \ldots, r_m)$ witnesses the shattering.

Definition 11.2 Suppose that F is a set of real-valued functions mapping from a domain X. Then F has pseudo-dimension d if d is the maximum cardinality of a subset S of X that is pseudo-shattered by F. If no such maximum exists, we say that F has infinite pseudo-dimension. The pseudo-dimension of F is denoted $\text{Pdim}(F)$.

The Fat-Shattering Dimension

Definition 11.10 Let F be a set of real-valued functions mapping from a domain X and suppose that $S = \{x_1, x_2, \ldots, x_m\} \subseteq X$. Suppose also that γ is a positive real number. Then S is γ-shattered by F if there are real numbers $r_1, r_2, \ldots, r_m$ such that for each $b \in \{0,1\}^m$ there is a function $f_b \in F$ with

$$f_b(x_i) \geq r_i + \gamma \text{ if } b_i = 1,\; f_b(x_i) \leq r_i - \gamma \text{ if } b_i = 0,\; \text{for } 1 \leq i \leq m.$$

Definition 11.11 Suppose that F is a set of real-valued functions mapping from a domain X and that $\gamma > 0$. Then F has γ-dimension d if d is the maximum cardinality of a subset S of X that is γ-shattered by F. If no such maximum exists, we say that F has infinite γ-dimension. The γ-dimension of F is denoted $\text{fat}_F(\gamma)$.

Relating Fat-Shattering Dimension and Pseudo-Dimension

Theorem 11.13 Suppose that F is a set of real-valued functions. Then,

1. For all $\gamma > 0$, $\text{fat}_F(\gamma) \leq \text{Pdim}(F)$.
2. If a finite set S is pseudo-shattered then there is γ_0 such that for all $\gamma < \gamma_0$, S is γ-shattered.
3. The function $\text{fat}_F(\gamma)$ is non-increasing with γ.
4. $\text{Pdim}(F) = \lim_{\gamma \downarrow 0} \text{fat}_F(\gamma)$ (where both sides may be infinite).

Neural Network Learning: Theoretical Foundations
Chapter 14 and 15

Martin Anthony and Peter L. Bartlett

Gi-Soo Kim
September 2 , 2017

Large Margin SEM Algorithms

- In analyzing classification learning algorithms for real-valued function classes, it is useful to consider algorithms that, given a sample and a parameter $\gamma > 0$, return hypotheses minimizing the sample error with respect to γ, which is defined as

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

where

$$\mathrm{margin}(f(x_i), y_i) = \begin{cases} f(x_i) - 1/2 & \text{if } y_i = 1 \\ 1/2 - f(x_i) & \text{if } y_i = 0 \end{cases}$$

Large Margin SEM Algorithms

Definition 13.1 Suppose that F is a set of real functions defined on the domain X. Then a large margin sample error minimization algorithm (or large margin SEM algorithm) L for F takes as input a margin parameter $\gamma > 0$ and a sample $z \in \bigcup_{m=1}^{\infty} Z^m$, and returns a function from F such that for all $\gamma > 0$, all m, and all $z \in Z^m$,

$$\hat{\text{er}}_z^{\gamma}(L(\gamma, z)) = \min_{f \in F} \hat{\text{er}}_z^{\gamma}(f).$$

Large Margin SEM Algorithms as Learning Algorithms

Theorem 13.2 Suppose that F is a set of real-valued functions defined on the domain X and that L is a large margin SEM algorithm for F. Suppose that $\epsilon \in (0,1)$ and $\gamma > 0$. Then given any probability distribution P on Z for all m, we have

$$P^m\{\mathrm{er}_P(L(\gamma, z)) \geq \mathrm{opt}_P^\gamma(F) + \epsilon\} \leq 2\mathcal{N}_\infty(\gamma/2, \pi_\gamma(F), 2m)e^{-\epsilon^m/72} + e^{-2\epsilon^2 m/9},$$

where $\mathrm{opt}_P^\gamma(F) = \inf_{f\in F} \mathrm{er}_P^\gamma(f)$.

Bounding Covering Number with the Pseudo-Dimension

Theorem 12.2 Let F be a set of real-valued functions from a domain X to the bounded interval $[0, B]$. Let d be a pseudo-dimension of F. Then for any $\epsilon > 0$,

$$\mathcal{N}_\infty(\epsilon, F, m) \leq \sum_{i=1}^{d} \binom{m}{i} \left(\frac{B}{\epsilon}\right)^i$$

which is less than $(emB/(\epsilon d))^d$ for $m \geq d$.

Bounding Covering Number with the Fat Shattering Dimension: A general upper bound

Theorem 12.8 Let F be a set of functions from a domain X to the bounded interval $[0, B]$. Let $d = \text{fat}_F(\epsilon/4)$. Then any $\epsilon > 0$ and for all $m \geq d$

$$\mathcal{N}_\infty(\epsilon, F, m) < 2\left(\frac{4mB^2}{\epsilon^2}\right)^{d\log_2(4eBm/(d\epsilon))}.$$

Bounding Covering Number with the Fat Shattering Dimension: A general lower bound

Theorem 12.10 Let F be a set of real-valued functions and let $\epsilon > 0$. Let $d = \mathrm{fat}_F(\epsilon/4)$. Then for all $m \geq \mathrm{fat}_F(16\epsilon)$,

$$\mathcal{N}_\infty(\epsilon, F, m) \geq \mathcal{N}_1(\epsilon, F, m) \geq e^{\mathrm{fat}_F(16\epsilon)/8}.$$

Large Margin SEM Algorithms as Learning Algorithms

Theorem 13.4 Suppose that F is a set of real-valued functions defined on the domain X with finite fat-shattering dimension, and that L is a large margin SEM algorithm for F. Then L is a classification learning algorithm for F. Given $\delta \in (0,1)$ and $\gamma > 0$, suppose $d = \text{fat}_{\pi_\gamma(F)}(\gamma/8) \geq 1$. Then the estimation error of L satisfies

$$\epsilon_L(m,\delta,\gamma) \leq \left[\frac{72}{m}\left\{d\log_2\left(\frac{32em}{d}\right)\log(128m) + \log\left(\frac{6}{\delta}\right)\right\}\right]^{1/2}$$

Furthermore, the sample complexity of L satisfies, for any $\epsilon \in (0,1)$,

$$m_L(\epsilon,\delta,\gamma) \leq \frac{144}{\epsilon^2}\left(27d\log^2\left(\frac{3456d}{\epsilon^2}\right) + \log\left(\frac{6}{\delta}\right)\right).$$

Large Margin SEM Algorithms as Learning Algorithms

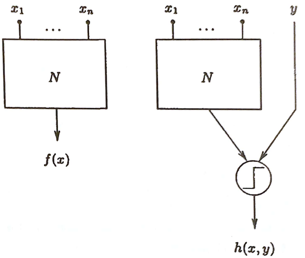
Theorem 13.6 If F is a set of real-valued functions with finite pseudo-dimension, and L is a large margin SEM algorithm for F. Let $d = \text{Pdim}(F)$. For all $\delta \in (0,1)$, all M, and $\gamma > 0$, its estimation error satisfies

$$\epsilon_L(m, \delta, \gamma) \leq \left[\frac{72}{m}\left\{d \log\left(\frac{8em}{d}\right) + \log\left(\frac{3}{\delta}\right)\right\}\right]^{1/2}.$$

outline

outline

Theorem 14.1 Let N be any neural network with a single real-valued output unit, and form a neural network N' as follows.



The network N' has one extra input unit and one extra computation unit. The extra computation unit is a linear threshold unit receiving input from output unit of N and the new input unit. If H' is the set of $\{0, 1\}$-value functions computed by N' and F the set of functions computed by N, then

$$Pdim(F) \leq VCdim(H').$$

Proof of Theorem 14.1 Use the fact that $Pdim(F) = VCdim(B_F)$ where

$$B_F = \{(x, y) \mapsto sgn(f(x) - y) : f \in F\}$$

Pseudo-dimension of neural networks

Theorem 14.2 Let F be the set of functions computed by a feed-forward network with W parameters and k computation units, in which each computation unit has the standard sigmoid activation function. Then,

$$Pdim(F) \leq ((W+2)k)^2 + 11(W+2)k\log_2(18(W+2)k^2).$$

<u>RECALL</u>

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, provided $m \geq W$,

$$VCdim(H) \leq (Wk)^2 + 11Wk\log_2(18Wk^2)$$

Bounds of fat-shattering dimension in terms of number of parameters W

- Split the network into two parts: the 1st layer & later layers.
- Let X be the input space $\mathbb{R}^n$, Y_1 be the output set of the 1st layer (ex. $\mathbb{R}^k$).
- Let $F_1 : X \longrightarrow Y_1$ be the class of vector valued functions computed by the 1st layer, and $G : Y_1 \longrightarrow \mathbb{R}$ be the class of functions computed by the remainder of the network.
- Then the set of functions computable by the whole network is

$$G \circ F_1 = \{g \circ f : g \in G, f \in F_1\}$$

Bounds of fat-shattering dimension in terms of number of parameters W

Definition Define the uniform, L_∞ distance between functions $h, g \in G$ as

$$d_{L_\infty}(g, h) = \sup_{y \in Y_1} |g(y) - h(y)|.$$

lemma 14.3 Let X be a set and (Y_1, ρ) be a metric space. Supp. $L \geq 0$, F_1 is a class of functions mapping from X to Y_1, G is a class of functions mapping from Y_1 to $\mathbb{R}$, satisfying Lipschitz condition: for all $g \in G$ and all $y, z \in Y_1$,

$$|g(y) - g(z)| \leq L\rho(y, z).$$

For $y = (y_1, \cdots, y_m)$ and $z = (z_1, \cdots, z_m)$ from Y_1^m, let

$$d_\infty^\rho(y, z) = \max_{1 \leq i \leq m} \rho(y_i, z_i).$$

Then,

$$N_\infty(\varepsilon, G \circ F, m) \leq \max_{x \in X^m} N\left(\varepsilon/2L, F_1|_x, d_\infty^\rho\right) N\left(\varepsilon/2, G, d_{L_\infty}\right).$$

Proof of lemma 14.3

Proof of lemma 14.3 Fix $x \in X^m$. Supp. that $\hat{F}_1$ is an $\frac{\varepsilon}{2L}$-cover of $F_1|_x$ w.r.t. d_∞^ρ and $\hat{G}$ is an $\frac{\varepsilon}{2}$-cover of G w.r.t. to d_{L_∞}. Let

$$\hat{G}|_{\hat{F}_1} = \{(\hat{g}(\hat{f}_1), \cdots, \hat{g}(\hat{f}_m)) : \hat{f} = (\hat{f}_1, \cdots, \hat{f}_m) \in \hat{F}_1, \hat{g} \in \hat{G}\}.$$

Then we can show $\hat{G}|_{\hat{F}_1}$ is an ε-cover of $(G \circ F)|_x$ w.r.t. to d_∞.
Choose $f \in F_1$ and $g \in G$, and pick $\hat{f} \in \hat{F}_1$ and $\hat{g} \in \hat{G}$ s.t.

$$d_\infty^\rho(f|_x, \hat{f}) \leq \frac{\varepsilon}{2L} \text{ and } d_{L_\infty}(g, \hat{g}) \leq \frac{\varepsilon}{2}.$$

Then,

$$\max_{1 \leq i \leq m} \rho(f(x_i), \hat{f}_i) \leq \frac{\varepsilon}{2L}$$

and so,

$$\max_{1 \leq i \leq m} \Big(g(f(x_i)) - g(\hat{f}_i) \Big) \leq L \cdot \frac{\varepsilon}{2L} = \frac{\varepsilon}{2}$$

due to Lipschitz condition, which implies

$$\max_{1 \leq i \leq m} |g(f(x_i)) - \hat{g}(\hat{f}_i)| \leq \varepsilon.$$

Bounds of fat-shattering dimension in terms of number of parameters W

Consider F computed by a feed-forward real-output multi-layer network, with following properties:

- $l \geq 2$ layers, with connections only between adjacent layers
- W weights
- For some $b > 0$, each computation unit maps into $[-b, b]$, and each computation unit in the 1st layer has non-dicreasing activation function.
- $\exists V > 0$ and $L > \frac{1}{V}$ s.t. for each unit in all but 1st layer, vector w of weight associated with that unit has $||w||_1 \leq V$ and the unit's activation function $s : \mathbb{R} \longrightarrow [-b, b]$ satisfies Lipschitz condition $|s(\alpha_1) - s(\alpha_2)| \leq L|\alpha_1 - \alpha_2|$ for all entries $\alpha_1, \alpha_2 \in \mathbb{R}$.
- Assume no threshold, for convenience.

Theorem 14.5 For the class F of functions computed by the network above, if $\varepsilon \leq 2b$, then

$$N_\infty(\varepsilon, F, m) \leq \left(\frac{4embW(LV)^l}{\varepsilon(LV-1)}\right)^W.$$

Proof of Theorem 14.5 I

Proof of Theorem 14.5 Use lemma 14.3 !!

1. Functions in G satisfy Lipschitz condition.

   lemma For every $g \in G$ and $y_1, y_2 \in Y_1$,

   $$|g(y_1) - g(y_2)| \leq (LV)^{l-1} ||y_1 - y_2||_\infty.$$

   proof Decompose $g = g_l \circ \cdots \circ g_2$ and use Lipschitz condition on s.

   $$\begin{aligned} ||g_i(y_1) - g_i(y_2)||_\infty &\leq Lmax|w^T(y_1 - y_2)| \\ &\leq Lmax\{||w||_1||y_1 - y_2||_\infty\} \\ &= LV||y_1 - y_2||\infty, \end{aligned}$$

   where y_1 and y_2 are units in layer i.

2. Bound on $N(\varepsilon, F_1|_x, d_\infty^\rho)$.

lemma For $x \in X^m$,

$$N(\varepsilon, F_1|_x, d_\infty^\rho) \leq \left(\frac{2emb}{\varepsilon}\right)^{W-W_G},$$

where W_G is the number of weights in all but 1st layer.
proof
For $f \in F_1$, we can write $f(x) = (f_1(x), \cdots, f_k(x)) \in [-b, b]^k$.
Define $F_{1,j} = \{f_j : (f_1, \cdots, f_k) \in F_1\}$. Then $F_1|_x \subset F_{1,1}|_x \times \cdots \times F_{1,k}|_x$.

$$\Rightarrow N(\varepsilon, F_1|_x, d_\infty^\rho) \leq \prod_{j=1}^{k} N(\varepsilon, F_{1,j}|_x, d_\infty)$$

Supp. $X \subset \mathbb{R}^n$. Since the activation function of each 1st layer unit is non-decreasing,

$$\max_{x \in X^m} N(\varepsilon, F_{1,j}|_x, d_\infty) \leq \left(\frac{2emb}{\varepsilon n}\right)^n$$

due to Thm 11.3, 11.6 and 12.2. Result follows from $kn = W - W_G$.

Proof of Theorem 14.5 III

3. Bound on $N(\varepsilon, G, d_{L_\infty})$.

<u>lemma</u> If $\varepsilon \leq 2b$ and $LV > 1$,

$$N(\varepsilon, G, d_{L_\infty}) \leq \left(\frac{2LVW_G b(LV)^{l-1}}{\varepsilon(LV-1)}\right)^{W_G}$$

Bounds of fat-shattering dimension in terms of number of parameters W

Theorem 14.9 For the class F of functions computed by the network described above,

$$fat_F(\varepsilon) \leq 16W\Big(l\log(LV) + 2\log(32W) + \log(\frac{b}{\varepsilon(LV-1)})\Big)$$

Proof of Theorem 14.9 Use Theorem 14.5 and Theorem 12.10

RECALL

Theorem 12.10 Let F be a set of real-valued functions and let $\epsilon > 0$. Let $d = \text{fat}_F(\epsilon/4)$. Then for all $m \geq \text{fat}_F(16\epsilon)$,

$$\mathcal{N}_\infty(\epsilon, F, m) \geq \mathcal{N}_1(\epsilon, F, m) \geq e^{\text{fat}_F(16\epsilon)/8}.$$

Bounds of fat-shattering dimension in terms of size of parameters V

Key idea is to approximate a network with bounded weights by one with few weights!

Bounds of fat-shattering dimension in terms of size of parameters V

Definition For a subset S of a vector space H, the convex hull of S, $co(S) \subset H$, is defined as

$$co(S) = \left\{ \sum_{i=1}^{N} \alpha_i s_i : N \in \mathbb{N}, s_i \in S, \alpha_i > 0, \sum_{i=1}^{N} \alpha_i = 1 \right\}$$

Theorem 14.10 Let F be a vector space with a scalar product and let $||f|| = \sqrt{(f, f)}$. Supp. $G \subset F$ and that for some $B > 0$, $||g|| \leq B$ for all $g \in G$. Then for all $f \in co(G)$, all $k \in \mathbb{N}$, and all $c > B^2 - ||f||^2$, $\exists g_1, \cdots, g_k \in G$ satisfying

$$||\frac{1}{k} \sum_{i=1}^{k} g_i - f||^2 \leq \frac{c}{k}.$$

Bounds of fat-shattering dimension in terms of size of parameters V

Theorem 14.11 Supp. $b > 0$ and that F is a class of $[-b, b]$-valued functions defined on X, and $N_2(\varepsilon, F, m)$ is finite for all $m \in \mathbb{N}$ and $\varepsilon > 0$. Then provided $\varepsilon_1 + \varepsilon_2 \leq \varepsilon$,

$$\log_2 N_2(\varepsilon, co(F), m) \leq \lceil \frac{b^2}{\varepsilon_1^2} \rceil \log_2 N_2(\varepsilon_2, F, m)$$

Proof of Theorem 14.11 I

Proof of Theorem 14.11 Take $N_2(\varepsilon_2, F, m) = N$. Then for any $x \in X^m$, $\exists \varepsilon_2$-cover S of $F|_x$ s.t. $|S| = N$.
Define $T_k \subset \mathbb{R}^m$ as

$$T_k = \{\frac{1}{k}\sum_{i=1}^{k} s_i : s_i \in S\}.$$

Then $|T_k| \leq N^k$. Choose any $f \in co(F)$ and suppose $f = \sum_{i=1}^{l} \alpha_i f_i$ with $\alpha_i > 0$, $\sum_{i=1}^{l} \alpha_i = 1$ and $f_i \in F$.
Since S is an ε_2-cover of $F|_x$, $\exists \hat{f}_1, \cdots, \hat{f}_l \in S$ s.t.

$$d_2(f_i|_x, \hat{f}_i) \leq \varepsilon_2.$$

$$\Rightarrow d_2(f|x, \sum_{i=1}^{l} \alpha_i \hat{f}_i) \leq \varepsilon_2.$$

By Theorem 14.10, $\exists g_1, \cdots, g_k \in S$ s.t.

$$d_2(\frac{1}{k}\sum_{i=1}^{k} g_i, \sum_{i=1}^{l} \alpha_i \hat{f}_i) \leq \frac{b}{\sqrt{k}}.$$

Proof of Theorem 14.11 II

By triangle inequality,

$$d_2(\frac{1}{k}\sum_{i=1}^{k} g_i, f|_x) \leq \varepsilon_2 + \frac{b}{\sqrt{k}}.$$

Hence, T_k is an $(\varepsilon_2 + \frac{b}{\sqrt{k}})$-cover of $co(F)|_x$. Choose $k = \lceil \frac{b^2}{\varepsilon_1^2} \rceil$.

Bounds of fat-shattering dimension in terms of size of parameters V

lemma 14.12 If G is a normed vector sapce with induced metric d and $F \subset G$, then

$$N(\varepsilon, F, d) = N(|\alpha|\varepsilon, \alpha F, d)$$

for $\forall \varepsilon > 0$ and $\alpha \in \mathbb{R}$.

lemma 14.13 Suppose F is a class of real-valued functions defined on X, and the function $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ satisfies the Lipschitz condition,

$$|\phi(x) - \phi(y)| \leq L|x - y|$$

for all $x, y \in \mathbb{R}$. Then,

$$N_2(\varepsilon, \phi \circ F, m) \leq N_2(\varepsilon/L, F, m).$$

Proof Use the fact

$$|(\phi \circ f)(x) - (\phi \circ g)(x)| \leq L|f(x) - g(x)|$$

Bounds of fat-shattering dimension in terms of size of parameters V

Theorem 14.14 Suppose $b > 0$ and that F_1 is a class of $[-b, b]$-valued functions defined on a set X and satisfying

- $F_1 = -F_1$
- F_1 contains the identically zero function

For $V \geq 1$, define,

$$F = \{\sum_{i=1}^{N} w_i f_i : N \in \mathbb{N}, f_i \in F_1, \sum_{i=1}^{N} |w_i| \leq V\}.$$

Then for $\varepsilon_1 + \varepsilon_2 < \varepsilon$,

$$log_2 N_2(\varepsilon, F, m) \leq \lceil \frac{V^2 b^2}{\varepsilon_1^2} \rceil log_2 N_2(\varepsilon_2 / V, F_1, m).$$

Proof of Theorem 14.14 Due to conditions on F_1,

$$\begin{aligned}\sum_{i=1}^{N} w_i f_i &= \sum_{i=1}^{N} w_i sgn(w_i) sgn(w_i) f_i \\ &= \sum_{i=1}^{N} \frac{w_i sgn(w_i)}{V} V sgn(w_i) f_i \\ &= V\Big[\sum_{i=1}^{N} \frac{w_i sgn(w_i)}{V} sgn(w_i) f_i + (1 - \sum_{i=1}^{N} \frac{w_i sgn(w_i)}{V})0\Big]\end{aligned}$$

$\Rightarrow F = Vco(F_1)$
Result follows from Theorem 14.11 and lemma 14.12.

Bounds of fat-shattering dimension in terms of size of parameters V

Two corollaries for bound on $N_2(\cdot,\cdot,\cdot)$ of 2-layer neural networks.

Corollary 14.15 Suppose that $b > 0$ and $s : \mathbb{R} \longrightarrow [-b, b]$ is a nondecreasing function. Let $V \geq 1$ and supp. that F is the class of functions from $\mathbb{R}^n$ to $\mathbb{R}$,

$$F = \{x \mapsto \sum_{i=1}^{N} w_i s(v_i^T x + v_{i0}) + w_0 : N \in \mathbb{N}, v_i \in \mathbb{R}^n, v_{i0} \in \mathbb{R}, \sum_{i=0}^{N} |w_i| \leq V\}$$

Then for $0 < \varepsilon \leq b$ and $m \geq n + 1$,

$$\log_2 N_2(\varepsilon, F, m) \leq \frac{5V^2 b^2 (n+3)}{\varepsilon^2} \log_2\left(\frac{4embV}{\varepsilon(n+1)}\right)$$

Proof of Corollary 14.15 Let

$$F_1 = \{x \mapsto s(v_i^T x + v_{i0})\}.$$

Then

$$N_2(\varepsilon, F_1, m) \leq N_\infty(\varepsilon, F_1, m) \leq \left(\frac{2emb}{\varepsilon(n+1)}\right)^{n+1}$$

for $m \geq n+1$.
By lemma 14.12, $N_2(\varepsilon, -F_1, m) = N_2(\varepsilon, F_1, m)$, so

$$N_2(\varepsilon, \underline{F_1 U - F_1 U\{0,1\}}, m) \leq 2N_2(\varepsilon, F_1, m) + 2.$$

By Theorem 14.14,

$$log_2 N_2(\varepsilon, F, m) \leq \lceil \frac{V^2 b^2}{\varepsilon_1^2} \rceil log_2\Big(2N_2(\varepsilon, F_1, m) + 2\Big)$$

Bounds of fat-shattering dimension in terms of size of parameters V

Corollary 14.16 Suppose that $b > 0, L > 0$ and $s : \mathbb{R} \longrightarrow [-b, b]$ satisfies $|s(\alpha_1) - s(\alpha_2)| \leq L|\alpha_1 - \alpha_2|$ for all $\alpha_1, \alpha_2 \in \mathbb{R}$.
For $V \geq 1$ and $B \geq 1$, let

$$F = \{\sum_{i=1}^{N} w_i f_i + w_0 : N \in \mathbb{N}, f_i \in F_1, \sum_{i=1}^{N} |w_i| \leq V\}$$

where

$$F_1 = \{x \mapsto s(\sum_{i=1}^{n} v_i x_i + v_0) : v_i \in \mathbb{R}, x \in [-B, B]^n, \sum_{i=0}^{n} |v_i| \leq V\}$$

Then, for $\varepsilon \leq Vmin\{BL, b\}$,

$$log_2 N_2(\varepsilon, F, m) \leq 50\left(\frac{V^3 L^2 bB}{\varepsilon}\right)^2 log_2(2n+2).$$

Proof of Corollary 14.16

Proof of Corollary 14.16 By Theorem 14.14 and lemma 14.13,

$$
\begin{aligned}
log_2 N_2(\varepsilon, F_1, m) &\leq \lceil \frac{V^2 B^2 L^2}{\varepsilon_1^2} \rceil log_2 N_2(\varepsilon_2 / VL, \underline{GU - GU\{0,1\}}, m) \\
&\leq \lceil \frac{V^2 B^2 L^2}{\varepsilon_1^2} \rceil log_2 \Big(2N_2(\varepsilon_2 / VL, G, m) + 2 \Big) \\
&\leq \lceil \frac{V^2 B^2 L^2}{\varepsilon_1^2} \rceil log_2 (2n + 2)
\end{aligned}
$$

where $G = \{x \mapsto x_i : i \in \{1, \cdots, n\}\}, \varepsilon_1 + \varepsilon_2 \geq \varepsilon$.
Similarly, if $b \geq 1$,

$$
log_2 N_2(\varepsilon, F, m) \leq \lceil \frac{V^2 b^2}{\varepsilon_1^2} \rceil log_2 \Big(2N_2(\varepsilon_2 / V, F_1, m) + 2 \Big)
$$

for $\varepsilon_1 + \varepsilon_2 \leq \varepsilon$.

Bounds of fat-shattering dimension in terms of size of parameters V

Let

$$F_0 = \{x \mapsto x_i : x = (x_1, \cdots, x_n) \in [-b, b]^n, i \in \{1, \cdots, n\}\} \; U \; \{0, 1\},$$

and for $i \geq 1$,

$$F_i = \{s(\sum_{j=1}^{N} w_j f_j) : N \in \mathbb{N},\; f_j \in \cup_{k=0}^{i-1} F_k,\; \sum_{j=1}^{N} |w_j| \leq V\}.$$

Thus, F_l is the class of functions that can be computed by an $l-$layer feed-forward network, in which each unit has sum of magnitude of weights bounded by V.
Assume $s : \mathbb{R} \longrightarrow [-b, b]$ satisfies Lipschitz condition.

Theorem 14.17 For $l \geq 1$,

$$log_2 N_2(\varepsilon, F_l, m) \leq \frac{1}{2}(\frac{2b}{\varepsilon})^{2l}(2VL)^{l(l+1)} log_2(2n+2),$$

provided $b \geq 1,\; V \geq \frac{1}{2L},\; \varepsilon \leq VbL.$

Bounds of fat-shattering dimension in terms of size of parameters V

Theorem 14.19

$$fat_{F_l}(\varepsilon) \leq 4(\frac{32b}{\varepsilon})^{2l}(2VL)^{l(l+1)}log(2n+2)$$

provided $b \geq 1$, $V \geq \frac{1}{2L}$, $\varepsilon \leq 16VbL$.

Proof Theorem 14.17 and Theorem 12.10

Theorem 14.18 Suppose $b \geq 1$ and $s : \mathbb{R} \longrightarrow [-b, b]$ is a non-decreasing function. Let $V \geq 1$ and supp. that

$$F = \{x \mapsto \sum_{i=1}^{N} w_i s(v_i^T x + v_{i0}) + w_0 : N \in \mathbb{N}, v_i \in \mathbb{R}^n, v_{i0} \in \mathbb{R}, \sum_{i=0}^{N} |w_i| \leq V\}$$

Then for $0 < \varepsilon \leq b$,

$$fat_F(\varepsilon) \leq 2^{16}(n+3)(\frac{bV}{\varepsilon})^2 log(\frac{2^8 bV}{\varepsilon})$$

Proof Corollary 14.15 and Theorem 12.10

outline

Model Selection

The first two parts of the book considered the following 3 step approach to solving a pattern classification problem.

1. Choose a suitable class of functions.
2. Gather data
3. Choose a function from the class.

Model Selection

Theorem 15.1 Let N_W be a 2-layer network with input set X, W parameters, a linear threshold output unit, and first-layer units with a fixed bounded piecewise-linear activation function. Let H_W be the class of functions computed by N_W. There is a constant c such that the following holds.
Suppose P is a probability distribution on $X \times \{0, 1\}$, and $z \in Z^m$ is chosen from P^m. Then if L_W is a SEM algorithm for H_W, wp at leat $1 - \delta$,

$$er_P(L_W(z)) < opt_P(H_W) + \left(\frac{c}{m}\left(W \log(W m) + \log(\frac{1}{\delta})\right)\right)^{1/2}.$$

Proof Theorem 8.8, Theorem 4.3, and Thoerem 4.2.

REMARK This result is applicable only if we fix the complexity of our class, W, before seeing any data.
$\Rightarrow$ Rather, we want that the learner chooses a suitable W after seeing the data.

Model Selection

Theorem 15.2 Let F_V be class of functions computed by a two-layer network,

$$F_V = \{x \mapsto \sum_{i=1}^{k} w_i s(v_i^T x + v_{i0}) + w_0 : k \in \mathbb{N}, \sum_{i=0}^{k} |w_i| \leq V\},$$

where $V > 0$, $x \in \mathbb{R}^n$, and $s : \mathbb{R} \longrightarrow [-1, 1]$ is non-decreasing. There is a constant c such that the following holds.
Fix $\gamma \in (0, 1]$ and suppose that P is a probability distribution on $X \times \{0, 1\}$, and that $z \in Z^m$ is chosen from P^m. Then, if L_V is a large margin SEM algorithm for F_V, wp at least $1 - \delta$,

$$er_P(L_V(z, \gamma)) < opt_P^{\gamma}(F_V) + \Big(\frac{c}{m}\Big(\frac{V^2 n}{\gamma^2} log^2(m) log(\frac{V}{\gamma}) + log(\frac{1}{\delta})\Big)\Big)^{1/2}$$

Proof Thoerem 14.18, Thoerem 13.2, and Thoerem 13.4.

REMARK Increasing γ decreases the estimation error term, but may increase the error term.

Model Selection

We want to choose the complexity parameters so as to minimize the upperbounds on misclassification probability.

Model Selection

Let L^c be a learning algorithm that returns $h \in \cup_W H_W$, corresponding to a pair (h, W) with $h \in H_W$, that minimizes

$$\hat{er}_z(h) + \left(\frac{c}{m}\left(W log(Wm) + log(\frac{W}{\delta})\right)\right)^{1/2},$$

over all values of $W \in \mathbb{N}$ and $h \in H_W$.

Theorem 15.3 There are constants c, c_1 such that wp at least $1 - \delta$,

$$er_P(L^c(z)) < \inf_W \left(opt_P(H_W) + \left(\frac{c_1}{m}\left(W log(Wm) + log(\frac{W}{\delta})\right)\right)^{1/2}\right).$$

Model Selection

Let L^c be a learning algorithm that returns $f \in \cup_V F_V$ corresponding to a triple (f, V, γ) with $f \in F_V$ and

$$\hat{er}_z^{\gamma}(f) + \Big(\frac{c}{m}\Big(\frac{V^2 n}{\gamma^2} log^2(m) log(\frac{V}{\gamma}) + log(\frac{V}{\gamma\delta})\Big)\Big)^{1/2}$$

within $1/m$ of its infimum over all values $\gamma \in (0, 1]$, $V \in \mathbb{R}^+$ and $f \in F_V$.

Thoerem 15.4 There are constants c, c_1 such that wp at least $1 - \delta$,

$$er_P(L^c(z)) < \inf_{V,\gamma}\Big(opt_P^{\gamma}(F_V) + \Big(\frac{c_1}{m}\Big(\frac{V^2 n}{\gamma^2} log^2(m) log(\frac{V}{\gamma}) + log(\frac{V}{\gamma\delta})\Big)\Big)^{1/2}\Big)$$

Model Selection: Remarks

- The model selection methods described in this chapter are similar to number of thechniques that are commonly used by neural network practitioners.
- Thoerem 15.6 $\longrightarrow$ weight decay