Neural Network Learning: Theoretical Foundations
Chapter 12 and 13

Martin Anthony and Peter L. Bartlett

Presented by Ilsang Ohn
August 25, 2017

outline

outline

Bounding Covering Numbers with Dimensions

- Pseudo-dimension and fat-shattering dimension, are generalizations of the VC-dimension
- Covering numbers are generalizations of the growth function.
- The pseudo-dimension and fat-shattering dimension are used to bound covering numbers and hence to bound the sample complexity and estimation error classification learning.

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 $\gamma > 0$, $\mathrm{fat}_F(\gamma) \leq \mathrm{Pdim}(F)$.
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 $\mathrm{fat}_F(\gamma)$ is non-increasing with γ.
4. $\mathrm{Pdim}(F) = \lim_{\gamma \downarrow 0} \mathrm{fat}_F(\gamma)$ (where both sides may be infinite).

Packing Numbers

Definition Let (A, d) be a metric space. Given $W \subset A$ and a positive number ϵ, a subset $P \subset W$ is said to be ϵ-separated or to be an ϵ-packing of W, if for all distinct $x, y \in P$, $d(x, y) > \epsilon$.

Definition A ϵ-packing number of W denoted by $\mathcal{M}(\epsilon, W, d)$, is the maximum cardinality of an ϵ-separated subset of W.

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

$$\mathcal{M}_p(\epsilon, H, k) = \max\{\mathcal{M}(\epsilon, H_{|x}, d_p) : x \in X^k\}.$$

for $p = 1, 2, \infty$.

Packing Numbers

Theorem 12.1 Let (A, d) be a metric space. Then for all positive ϵ, and for every subset $W \subset A$, the covering numbers and packing numbers satisfy

$$\mathcal{M}(2\epsilon, W, d) \leq \mathcal{N}(\epsilon, W, d) \leq \mathcal{M}(\epsilon, W, d)$$

Proof

1. If M is a 2ϵ-separated subset of W and N is a ϵ-cover of W, then N must select a point within ϵ distance of each of the points in M. These points will necessarily be distinct since points in M are at least 2ϵ apart. Thus $|M| \leq |N|$.
2. If M is a maximal ϵ-separated subset of W then M has to be an ϵ-cover. Because if it is not, then there is a point $w \in W$ such that there is no point of M within a distance of ϵ from w. In that case, w can be added to M while still keeping it ϵ-separated. This violates the maximality of M. Thus, $\mathcal{N}(\epsilon, W, d) \leq |M|$.

Bounding 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 $m \geq d$.

Proof of Theorem 12.2

For a positive real number α, define Q_α as

$$Q_\alpha(u) = \alpha \left\lfloor \frac{u}{\alpha} \right\rfloor$$

Lemma 12.3 Let F be a set of real-valued functions from a domain X to the interval $[0,1]$. Then for any $\epsilon > 0$, any positive integers m and any $0 < \alpha \leq \epsilon$,

$$\mathcal{M}_\infty(\epsilon, F, m) \leq \mathcal{M}_\infty\left(\alpha \left\lfloor \frac{\epsilon}{\alpha} \right\rfloor, Q_\alpha(F), m\right)$$

where $Q_\alpha(F) = \{Q_\alpha(f) : f \in F\}$ with the function $Q_\alpha(f)$ defined as

$$(Q_\alpha(f))(x) = Q_\alpha(f(x))$$

which maps from X into the finite subset $\{0, \alpha, 2\alpha, \ldots, \lfloor 1/\alpha \rfloor \alpha\}$. In particular

$$\mathcal{M}_\infty(\epsilon, F, m) \leq \max_{x \in X^m} \left| Q_\epsilon(F)_{|x} \right|$$

Proof of Lemma 12.3 For $x = (x_1, \ldots, x_m)$, since $|Q_\alpha(b) - Q_\alpha(a)| \geq Q_\alpha(|b - a|)$,

$$\begin{aligned} d_\infty(f_x, g_x) \geq \epsilon &\Leftrightarrow |f(x_i) - g(x_i)| \geq \epsilon \text{ for some } i = 1, \ldots, m \\ &\Rightarrow |(Q_\alpha(f))(x_i) - (Q_\alpha(g))(x_i)| \geq \epsilon \left\lfloor \frac{\epsilon}{\alpha} \right\rfloor \text{ for some } i = 1, \ldots, m \\ &\Leftrightarrow d_\infty((Q_\alpha(f))_x, (Q_\alpha(g))_x) \geq \epsilon \left\lfloor \frac{\epsilon}{\alpha} \right\rfloor \end{aligned}$$

The second inequality follows on substituting $\alpha = \epsilon$ since

$$\mathcal{M}(\epsilon, Q_\epsilon(F)_{|x}, m) \leq \left| Q_\epsilon(F)_{|x} \right|$$

Proof of Theorem 12.2

Lemma (Theorem 12.4) Suppose that H is a set of functions from a finite set X with $|X| = m$ to a finite set $Y \subset \mathbb{R}$ with $|Y| = N$ and that $\mathrm{Pdim}(H) \leq d$. Then

$$|H| \leq \sum_{i=0}^{d} \binom{m}{i} (N-1)^i.$$

Without the condition that $\mathrm{Pdim}(H) \leq d$, $|H| = N^m = \sum_{i=0}^{m} \binom{m}{i} (N-1)^i$. Suppose that there are $S = \{x_1, \ldots, x_d, x_{d+1}\} \subset X$ and $h \in H$ such that $h(x_i) \neq h(x_j)$ for all $i \neq j \in \{1, \ldots, d+1\}$, then $\mathrm{Pdim}(H) \geq d+1$.

Proof of Theorem 12.2

Proof of Theorem 12.2 Applying Theorem 12.4 with $H = Q_\epsilon(F)_{|x}$ which maps into the finite set of cardinality $N = 1 + \lfloor 1/\epsilon \rfloor$, we obtain

$$\mathcal{M}_\infty(\epsilon, F, m) \leq \max_{x \in X^m} \left| Q_\epsilon(F)_{|x} \right| \leq \sum_{i=0}^{d} \binom{m}{i} \left\lfloor \frac{1}{\epsilon} \right\rfloor^i .$$

where $d = \text{Pdim}(Q_\epsilon(F)_{|x}) \leq \text{Pdim}(Q_\epsilon(F)) \leq \text{Pdim}(F)$ by Theorem 11.3 since $Q_\epsilon(\cdot)$ is non-decreasing.

Bounding with the Fat Shattering Dimension: A general upper bound

Theorem 12.7 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 for any $\epsilon > 0$,

$$\mathcal{M}_\infty(\epsilon, F, m) < 2(mb^2)^{\lceil \log_2 y \rceil}$$

where $b = \lfloor 2B/\epsilon \rfloor$ and $y = \sum_{i=1}^{d} \binom{m}{i} b^i$.

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 $m \geq d$

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

Proof of Theorem 12.7

Proof of Theorem 12.7 By Lemma 12.3 with $\alpha = \epsilon/2$

$$\mathcal{M}_\infty(\epsilon, F, m) \leq \mathcal{M}_\infty(\epsilon, Q_{\epsilon/2}(F), m).$$

By a simple rescaling, Lemma 12.9 (next slide) shows that

$$\mathcal{M}(\epsilon, Q_{\epsilon/2}(F), d_\infty) \leq 2(mb^2)^{\lceil \log_2 y' \rceil}$$

where

$$y' = \sum_{i=1}^{\mathrm{fat}_{Q_{\epsilon/2}(F)}(\epsilon/2)} \binom{m}{i} b^i \leq \sum_{i=1}^{\mathrm{fat}_F(\epsilon/4)} \binom{m}{i} b^i = y$$

Proof of Theorem 12.7

Lemma 12.9 Let $Y = \{0, 1, \ldots, b\}$, and suppose $|X| = m$ and $H \subset Y^X$ has $\text{fat}_H(1) = d$ Then

$$\mathcal{M}(2, H, d_\infty) \leq 2(mb^2)^{\lceil \log_2 y \rceil}$$

where $y = \sum_{i=1}^{d} \binom{m}{i} b^i$.

Proof of Lemma 12.9 Fix $b \geq 3$ as the result trivially holds otherwise. For given X and $G \subset Y^X$, define $T_{X,G}$ as

$$T_{X,G} = \{(A, r) : G \text{ 1-shatters } \emptyset \neq A \subset X, \text{ witnessed by } r : A \to Y\}$$

For $k \geq 2$ and $m \geq 1$, define $t(k, m)$ as

$$t(k, m) = \min\{|T_{X,G}| : |X| = m, G \subset Y^X, |G| = k, \; G \text{ is 2-separated}\}$$

or take $t(k, m)$ to be infinite if the minimum is over the empty set.

Proof of Theorem 12.7

Proof of Lemma 12.9 Note that the number of pairs (A, r) with $A \neq \emptyset$ and $|A| \leq d$ is less than

$$y = \sum_{i=1}^{d} \binom{m}{i} b^i$$

If $t(k, m) \geq y$, then every 2-separated set G with $|G| = k$ 1-shatters some A with $|A| > d$ i.e., $\text{fat}_G(1) > d$. But $\text{fat}_H(1) = d$, so if $t(k, m) \geq y$ then $\mathcal{M}(2, H, d_\infty) < k$.

It suffices to prove that

$$t\left(2(mb^2)^{\lceil \log_2 y \rceil}, m\right) \geq y$$

for all $d \geq 1$ and all $m \geq 1$.

Proof of Theorem 12.7

Proof of Lemma 12.9 Prove $t\left(2(mb^2)^{\lceil \log_2 y \rceil}, m\right) \geq y$ for all $d \geq 1$ and all $m \geq 1$.

- Let G be a 2-separated set with $|G| = k = 2(mb^2)^{\lceil \log_2 y \rceil}$. Split G into $K/2$ arbitrary pairs.
- One can show (pigeonhole) that there are $x_0 \in X, i, j$ with $j \geq i + 2$ such that at least $k/(mb^2)$ of these pairs, say (g_1, g_2), satisfy $(g_1(x_0), g_2(x_0)) = (i, j)$. Let G_1 be a set of such g_1's and G_2 a set of such g_2's. Then $|G_1| = |G_2| > k/(mb^2)$ and they are 2-separated on $X \setminus \{x_0\}$.
- Hence there are at least $t(\lfloor k/mb^2 \rfloor, m-1)$ pairs (A, r) such that G_1 (G_2) 1-shatters $A \in X \setminus \{x_0\}$ witnessed by r.
- If both G_1 and G_2 1-shatter A witnessed by r, then G 1-shatters $A \cup \{x_0\}$, witnessed by r' with $r'(x) = r(x)$ if $x \in X \setminus \{x_0\}$ and $r'(x_0) = \lfloor (i+j)/2 \rfloor$. Hence
$$t(k, m) \geq 2t\left(\left\lfloor \frac{k}{mb^2} \right\rfloor, m-1\right).$$
The proof follows by induction.

Proof of Theorem 12.7

Lemma If $\alpha < 2\epsilon$ then

$$\text{fat}_{Q_\alpha(F)}(\epsilon) \leq \text{fat}_F(\epsilon - \alpha/2)$$

and, in particular,

$$\text{fat}_{Q_{\epsilon/2}(F)}(\epsilon/2) \leq \text{fat}_F(\epsilon/4)$$

Proof

$$\begin{aligned} (Q_\alpha(f_b))(x_i) - r_i &\geq \epsilon && \text{if } b_i = 1 \\ (Q_\alpha(f_b))(x_i) - r_i &\leq -\epsilon && \text{if } b_i = 0 \end{aligned}$$

implies

$$\begin{aligned} f_b(x_i) - r_i &\geq \epsilon && \text{if } b_i = 1 \\ f_b(x_i) - r_i &\leq -\epsilon + \alpha && \text{if } b_i = 0 \end{aligned}$$

Bounding 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 $d = \text{fat}_F(\epsilon/4)$. Then for all $m \geq \text{fat}_F(16\epsilon)$,

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

Proof of Theorem 12.10

Lemma Let $d = \text{fat}_F(16\epsilon)$. If $m \geq d$, then

$$\mathcal{N}_1(\epsilon, F, m) \geq \mathcal{N}_1(2\epsilon, F, d).$$

Proof Let $m = kd + r$ where $k \geq 1$ and $0 \leq r < d$. Let z be the sample of length m obtained by concatenating k copies of x and adjoining the first r elements of x. For $f, g \in F$,

$$\begin{aligned} d_1(f_{|z}, g_{|z}) &= \frac{1}{m} \sum_{i=1}^{m} |f(z_i) - g(z_i)| \\ &= \frac{k}{kd + r} \sum_{i=1}^{d} |f(x_i) - g(x_i)| + \frac{1}{kd + r} \sum_{i=1}^{r} |f(x_i) - g(x_i)| \\ &\geq \frac{kd}{kd + r} d_1(f_{|x}, g_{|x}) \end{aligned}$$

Since $kd/(kd + r) > 1/2$, $d_1(f_{|z}, g_{|z}) < \epsilon$ implies $d_1(f_{|x}, g_{|x}) < 2\epsilon$.

Proof of Theorem 12.10

Lemma If $d = \text{fat}_F(16\epsilon)$, then $\mathcal{N}_1(2\epsilon, F, d) \geq e^{d/8}$

Proof Fix a sample x of length d that is 16ϵ-shattered. There is $r \in \mathbb{R}^d$ such that for every $b \in \{0,1\}^d$, there is $f_b \in F$ such that

$$f_b(x_i) \geq r_i + 16\epsilon \text{ if } b_i = 1, \; f_b(x_i) \leq r_i - 16\epsilon \text{ if } b_i = 0 \text{ for } i = 1 \ldots, d$$

Let $G = \{f_b : b \in \{0,1\}^d\}$ be such a set of 2^d functions.
Suppose C is a 2ϵ cover of $F_{|x}$. For each $c \in C$, there is $g \in G$ satisfying $d_1(c_{|x}, g_{|x}) < 2\epsilon$ and so

$$\left\{g' \in G : d_1(g'_{|x}, c_{|x}) < 2\epsilon\right\} \subset \left\{g' \in G : d_1(g'_{|x}, g_{|x}) < 4\epsilon\right\}$$

One can show that $\left|\left\{g' \in G : d_1(g'_{|x}, g_{|x}) < 4\epsilon\right\}\right| \leq 2^d e^{-d/8}$ which means that each element of C covers at most $2^d e^{-d/8}$ elements of G. Hence

$$|C| \geq \frac{|G|}{2^d e^{-d/8}} = \epsilon^{d/8}.$$

Fat-shattering dimension characterizes covering numbers

Theorem 12.11 Let F be a set of functions from a domain X to the bounded interval $[0, B]$. Then for any $\epsilon > 0$, if $m \geq \text{fat}_F(\epsilon/r) \geq 1$,

$$\frac{\log_2 \epsilon}{8} \text{fat}_F(16\epsilon) \leq \log_2 \mathcal{N}_1(\epsilon, F, m)$$
$$\leq \log_2 \mathcal{N}_\infty(\epsilon, F, m) \leq 3\text{fat}_F(\epsilon/4) \log_2^2 \left(\frac{4eBm}{\epsilon} \right).$$

REMARK. If a class has finite fat-shattering dimension, then the covering number is a sub-exponential function of m.

Example

Theorem 12.12 Let F be a set of functions of total variation at most V, mapping from the interval $[0,1]$ into $[0,1]$. Then for any $\epsilon > 0$ and for all m,

$$\mathcal{N}_\infty(\epsilon, F, m) < 2\left(\frac{4m}{\epsilon^2}\right)^{(1+2V/\epsilon)\log_2(2em/V)}.$$

Proof. Recall that $\text{fat}_F(\gamma) = 1 + \lfloor V/(2\gamma) \rfloor$. Then by Theorem 12.8 with $B = 1$ and $d = 1 + \lfloor 2V/\epsilon \rfloor$, we have

$$\mathcal{N}_\infty(\epsilon, F, m) \le 2\left(\frac{4mB^2}{\epsilon^2}\right)^{d\log_2(4eBm/(d\epsilon))} < 2\left(\frac{4m}{\epsilon^2}\right)^{(1+2V/\epsilon)\log_2(2em/V)}.$$

Example

Theorem 12.13 Let F be a set of real-valued functions. Let $\gamma > 0$ and let $d = \text{fat}_F(\gamma/8)$. Then

$$\mathcal{N}_\infty(\gamma/2, \pi_\gamma(F), 2m) \leq 2(128m)^{d \log_2(32em/d)}$$

where $\pi_\gamma(u) = \max(1/2 - \gamma, \min(1/2 + \gamma, u))$.

Proof We may assume $\pi_\gamma(F)$ maps into $[0, 2\gamma]$. Then by Theorem 12.8 with $B = 2\gamma$ and $\epsilon = \gamma/2$, we have

$$\mathcal{N}_\infty(\gamma/2, \pi_\gamma(F), m) \leq 2\left(\frac{4mB^2}{\epsilon^2}\right)^{d \log_2(4eBm/(d\epsilon))} = 2(64m)^{d \log_2(32em/d)}.$$

Remark The upper bound in Theorem 10.3

$$\begin{aligned} P^m\left(\exists f \in F : \text{er}_P(f) \geq \hat{\text{er}}_z^\gamma(f) + \epsilon\right) &\leq 2\mathcal{N}_\infty(\gamma/2, \pi_\gamma(F), 2m)e^{-\epsilon^2 m/8} \\ &\leq 4(128m)^{d \log_2(32em/d)} e^{-\epsilon^2 m/8} \end{aligned}$$

Comparing the Two Approaches

- We have seen that if F is uniformly bounded,

$$\mathcal{N}_\infty(\epsilon, F, m) \leq \left(\frac{c_1 m}{\epsilon \mathrm{Pdim}(F)}\right)^{\mathrm{Pdim}(F)}$$

  and

$$\mathcal{N}_\infty(\epsilon, F, m) \leq \left(\frac{c_2 m}{\epsilon^2}\right)^{\mathrm{fat}_F(\epsilon/4)\log_2(c_3 m/(\mathrm{fat}_F(\epsilon/4)\epsilon))} = \left(\frac{c_3 m}{\epsilon \mathrm{fat}_F(\epsilon/4)}\right)^{c_4 \mathrm{fat}_F(\epsilon/4)}$$

  for some constants c_1, c_2, c_3 and c_4.
- Theorem 11.13 (a):

$$\mathrm{fat}_F(\epsilon/4) \leq \mathrm{Pdim}(F).$$

  - If the two are equal then the first bound is better.
  - However, it is possible for $\mathrm{fat}_F(\epsilon/4)$ to be significantly less than $\mathrm{Pdim}(F)$. For example for the class F of bounded variation functions, $\mathrm{Pdim}(F)$ is infinite but $\mathrm{fat}_F(\epsilon/4)$ is finite.

outline

Large Margin SEM Algorithms

- For binary classification, SEM algorithms L, which have the property that for all z,

$$\hat{\text{er}}_z(L(z)) = \min_{h \in H} \hat{\text{er}}_z(h) = \frac{1}{m}|\{i : h(x_i) \neq y_i)\}|$$

are learning algorithms when the clss H has finite VC-dimension.

- 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

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

where

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

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

AIM. Show that the large margin SEM algorithms for a function class F are learning algorithms when F has finite fat-shattering dimension. i.e.,

For any probability distribution P on $Z = X \times \{0, 1\}$, the large margin SEM algorithm L taking as input $\gamma \in (0, 1/2]$ and a sample $z \in \bigcup_{m=1}^{\infty} Z^m$ satisfies, with probability at least $1 - \delta$,

- $\exists m_L(\epsilon, \delta, \gamma)$ s.t. $\forall \epsilon > 0$, $\mathrm{er}_P(L(z)) < \mathrm{opt}_P^{\gamma}(F) + \epsilon$ whenever $m \geq m_L(\epsilon, \delta, \gamma)$

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

- $\forall m$, $\exists \epsilon_L(m, \delta, \gamma)$ s.t. $\mathrm{er}_P(L(z)) < \mathrm{opt}_P^{\gamma}(F) + \epsilon_L(m, \delta, \gamma)$.

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

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

Proof With probability at least $1 - 2\mathcal{N}_\infty(\gamma/2, \pi_\gamma(F), 2m)e^{-\epsilon^m/72} - e^{-2\epsilon^2 m/9}$,

$$\mathrm{er}_P(L(\gamma, z)) < \hat{\mathrm{er}}_z^{\gamma}(L(\gamma, z)) + \frac{\epsilon}{3} \leq \hat{\mathrm{er}}_z^{\gamma}(f^*) + \frac{\epsilon}{3} < \mathrm{er}_P^{\gamma}(f^*) + \frac{2\epsilon}{3}$$

where $f^* \in F$ is such that $\mathrm{er}_P^{\gamma}(f^*) < \mathrm{opt}_P^{\gamma}(F) + \epsilon/3$.

Lemma 13.3 Suppose that f is a real-valued function defined on X, P is a probability distribution on Z, $\epsilon > 0$, $\gamma > 0$, and m is a positive integer. Then

$$P^m\left(\hat{\text{er}}_z^\gamma(f) \geq \text{er}_P^\gamma(f) + \epsilon\right) \leq e^{-2\epsilon^2 m}$$

Lemma (Theorem 10.4, Uniform convergence)

$$P^m\left(\exists f \in F : \text{er}_P(f) \geq \hat{\text{er}}_z^\gamma(f) + \epsilon\right) \leq 2\mathcal{N}_\infty(\gamma/2, \pi_\gamma(F), 2m)e^{-\epsilon^2 m/8}.$$

Proof of Theorem 13.2

- Let $f^* \in F$ be such that $\text{er}_P^\gamma(f^*) < \text{opt}_P^\gamma(F) + \epsilon/3$. Then $\hat{\text{er}}_z^\gamma(f^*) < \text{er}_P^\gamma(f^*) + \epsilon/3 < \text{opt}_P^\gamma(F) + 2\epsilon/3$ with probability at lest $1 - e^{-2\epsilon^2 m/9}$.
- With probability at least $1 - 2\mathcal{N}_\infty(\gamma/2, \pi_\gamma(F), 2m)e^{-\epsilon^2 m/72}$, $\text{er}_P(f) < \hat{\text{er}}_z^\gamma(f) + \epsilon/3$ for all $f \in F$.
- Hence with probability $1 - e^{-2\epsilon^2 m/9} - 2\mathcal{N}_\infty(\gamma/2, \pi_\gamma(F), 2m)e^{-\epsilon^2 m/72}$

$$\text{er}_P(L(\gamma, z)) < \hat{\text{er}}_z^\gamma(L(\gamma, z)) + \frac{\epsilon}{3} \leq \hat{\text{er}}_z^\gamma(f^*) + \frac{\epsilon}{3} < \text{opt}_P^\gamma(F) + \epsilon$$

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 $\epsilon \in (0,1)$,

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

Theorem 4.2 For H a set of $\{0,1\}$-valued functions with VC dimension d,

- $\epsilon_L(m, \delta) \leq \left[\frac{32}{m} \left\{ d \log \left(\frac{2em}{d} \right) + \log \left(\frac{4}{\delta} \right) \right\} \right]^{1/2}$
- $m_L(\epsilon, \delta) \leq \frac{64}{\epsilon^2} \left(2d \log \left(\frac{12}{\epsilon} \right) + \log \left(\frac{4}{\delta} \right) \right)$

Proof of Theorem 13.4

For $d = \mathrm{fat}_{\pi_\gamma(F)}(\gamma/8) \geq 1$,

$$\begin{aligned} & P^m(\mathrm{er}_P(L(\gamma, z)) > \mathrm{opt}_P^\gamma(F) + \epsilon) \\ & \leq 2\mathcal{N}_\infty(\gamma/2, \pi_\gamma(F), 2m) e^{-\epsilon^m/72} + e^{-2\epsilon^2 m/9} \qquad \text{(Thm 13.2)} \\ & \leq 3\max(1, \mathcal{N}_\infty(\gamma/2, \pi_\gamma(F), 2m)) e^{-\epsilon^m/72} \\ & < 6(128m)^{d\log_2(32em/d)} e^{-\epsilon^2 m/72} := \delta^* \qquad \text{(Thm 12.13)} \end{aligned}$$

$\delta^* \leq \delta$ when

- $\epsilon \geq \left[\frac{72}{m}\left\{d\log_2\left(\frac{32em}{d}\right) + \log(128m) + \log\left(\frac{6}{\delta}\right)\right\}\right]^{1/2}$
- $m \geq \frac{72}{\epsilon^2}\left(\frac{d}{\log 2}(\log m)^2 + 14d\log m + 7d\log\left(\frac{32e}{d}\right) + \log\left(\frac{6}{\delta}\right)\right)$
  Bound above $\log m$ by using the inequality $\log a \leq ab - \log b - 1$ for $a, b, > 0$ and bound above $(\log m)^2$ by using the inequality $(\log a)^2 \leq 6ab + 3(\log(1/b))^2$ for $a > 0, 0 < b < 1$ and $ab \geq 1$. Therefore

$$\frac{m}{2} \geq \frac{72}{\epsilon^2}\left(\frac{3d}{\log 2}\log^2\left(\frac{1728d}{\epsilon^2\log 2}\right) + 14d\log\left(\frac{4032d}{e\epsilon^2}\right) + 7d\log\left(\frac{32e}{d}\right) + \log\left(\frac{6}{\delta}\right)\right)$$

Lower Bounds for Certain Function Classes

Theorem 13.5 Suppose that F is a set of functions mapping into the interval $[0, 1]$ and that F is closed under addition of constants. Then, if L is any classification learning algorithm for F, the sample complexity of L satisfies

$$m_L(\epsilon, \delta, \gamma) \geq \max\left(\frac{d}{320\epsilon^2}, 2\left\lfloor \frac{1-\epsilon^2}{\epsilon^2} \log \frac{1}{8\delta(1-2\delta)} \right\rfloor\right)$$

for $0 < \epsilon < 1$, $\delta < 1/64$ and $\gamma > 0$, where $d = \text{fat}_{\pi_{4\gamma}(F)}(2\gamma) \geq 1$.

Proof of Theorem 13.5

Theorem 5.4 Suppose that H is a set of $\{0,1\}$-valued functions with VC dimension d. For any learning algorithm L for H the sample complexity of L satisfies

$$m_L(\epsilon, \delta) \geq \max\left(\frac{d}{320\epsilon^2}, 2\left\lfloor \frac{1-\epsilon^2}{\epsilon^2} \log\left(\frac{1}{8\delta(1-2\delta)}\right)\right\rfloor\right)$$

for all $0 < \epsilon < 1$ and $\delta < 1/64$.

Proof of Theorem 13.5 Construct H as follows.

- Choose $S \subset X$ so that $|S| = d = \text{fat}_{\pi_{4\gamma}(F)}(2\gamma)$ and S is 2γ-shattered by $\pi_{4\gamma}(F)$ witnessed by $r \in [1/2 - 2\gamma, 1/2 + 2\gamma]^d$.
- Let $T \subset S$ be the set of x_i with $r_i \in [1/2 - 2\gamma, 1/2]$. WLOG, assume $|T| \geq d/2$. Then T is γ-shattered by $\pi_{2\gamma}(F)$ witnessed by $(1/2 - \gamma, \cdots, 1/2 - \gamma)$.
- Let $F_0 \subset F$ be the set of functions $f \in F$ such that for all $x \in T$, $|f(x) - 1/2| \geq \gamma$. It is possible since F is closed under addition of constants.
- The set H of $\{0,1\}$-valued functions on T defined by

  $$H = \{x \mapsto \text{sign}(f(x) - 1/2) : f \in F_0\}$$

  is the set of all $\{0,1\}$-valued functions on T, and hence $\text{VCdim}(H) \geq d/2$.

Proof of Theorem 13.5

Proof of Theorem 13.5 For any P on Z and any ϵ, if $m \geq m_L(\epsilon, \delta, \gamma)$

$$P^m(\mathrm{er}_P(L(\gamma, z)) < \mathrm{opt}_P^\gamma(F) + \epsilon) \geq 1 - \delta$$

where we have

$$\begin{aligned} \mathrm{opt}_P^\gamma(F) = \inf_{f \in F} \mathrm{er}_P^\gamma(f) &\leq \inf_{f \in F_0} \mathrm{er}_P^\gamma(f) \\ &= \inf_{h \in H} \mathrm{er}_P(h) \quad \because \forall f \in F_0, |f(x) - 1/2| \geq \gamma \end{aligned}$$

Thus $z \mapsto \mathrm{sign}(L(\gamma, z) - 1/2)$ is a learning algorithm for H with $m_L(\epsilon, \delta, \gamma)$.

From Theorems 13.4 and 13.5,

$$\frac{c_1 \text{fat}_{\pi_{4\gamma}(F)}(2\gamma)}{\epsilon^2} \leq m_L(\epsilon, \delta, \gamma) \leq \frac{c_2 \text{fat}_{\pi_\gamma(F)}(\gamma/8)}{\epsilon^2}.$$

- Only the behavior of functions in F near the threshold value $1/2$ influences the complexity of F for classification learning, whereas the fat-shattering dimension in these bounds measures the complexity of functions in $\pi_\gamma(F)$ over the whole of their $[1/2-\gamma, 1/2+\gamma]$ range.
- The condition that F is closed under addition of constants ensures that the complexity of F is uniform over this range.
- Let $F = \{f : \mathbb{N} \mapsto [1/2+\alpha, \infty)\}$ for $\alpha < 0$. Then $\text{fat}_{\pi_\gamma(F)}(\gamma/8)$ is infinite but there is a classification learning algorithm for F. The class F is complex, but the complexity of the functions in F is restricted to a range that does not include the threshold, and hence this complexity is irrelevant for classification learning.

Using the Pseudo-Dimension

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 $d = \mathrm{Pdim}(F)$. For all $\delta \in (0,1)$, all M, and $\gamma > 0$, its estimation error satisfies

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

Remark (Theorem 4.2) For H a set of $\{0,1\}$ valued functions with VC dimension d,

$$\epsilon_L(m,\delta) \leq \left[\frac{32}{m}\left\{d \log\left(\frac{2em}{d}\right) + \log\left(\frac{4}{\delta}\right)\right\}\right]^{1/2}$$

Let $H = \{x \mapsto \mathrm{sign}(f(x) - 1/2) : f \in F\}$. Since $\mathrm{VCdim}(H) \leq \mathrm{Pdim}(F)$ and $\mathrm{opt}_P(H) \leq \mathrm{opt}_P^\gamma(F)$, Theorem 13.6 is weaker than the VC-dimension results.

But using the fat-shattering dimension can give a significant improvement.

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

where $d = \mathrm{fat}_{\pi_\gamma(F)}(\gamma/8)$. In next chapter, we see examples of neural network classes that have finite fat-shattering dimension, but whose thresholded versions have infinite VC-dimension.

Relative Uniform Convergence Results

Theorem 13.5 implies that the rate of uniform convergence of $\text{er}_P(f)$ to $\hat{\text{er}}_z^\gamma(f)$ can be no faster than $1/\sqrt{m}$. But as the result of Section 5.5, $\text{er}_P(f)$ converges more quickly to $(1+\alpha)\hat{\text{er}}_z^\gamma(f)$ for any fixed $\alpha > 0$.

Theorem 13.7 Suppose that F is a set of real-valued functions defined on X. Then for given any probability distribution P on Z, any $\gamma > 0$ and any $\alpha, \beta > 0$,

$$P^m\left(\exists f \in F : \text{er}_P(f) > (1+\alpha)\hat{\text{er}}_z^\gamma(f) + \beta\right) \leq 4\mathcal{N}_\infty(\gamma/2, \pi_\gamma(F), 2m)e^{-\alpha\beta m/(4(1+\alpha))}.$$

Theorem 10.4

$$P^m\left(\exists f \in F : \text{er}_P(f) \geq \hat{\text{er}}_z^\gamma(f) + \epsilon\right) \leq 2\mathcal{N}_\infty(\gamma/2, \pi_\gamma(F), 2m)e^{-\epsilon^2 m/8}.$$