Neural Network Learning: Theoretical Foundations
Chapter 18,19

Speaker : Semin Choi

Department of Statistics, Seoul National University, South Korea

November 15, 2017

Introduction

- We have seen that the d_1-covering numbers are crucial:

$$P^m\{\text{some } f \text{ in } F \text{ has } |er_P(f) - \hat{er}_z(f)| \geq \epsilon\}$$
$$\leq 4\mathcal{N}_1(\epsilon/16, F, 2m)\exp(-\epsilon^2 m/32)$$

- As in chapter 12, we present two bounds, one in terms of the fat-shattering dimension, and one in terms of the pseudo-dimension.

Bounding with the Fat-Shattering Dimension

A Bound on the d_1-packing numbers in terms of the fat-shattering dimension

Theorem 18.1.

Suppose that F is a set of real functions from a domain X to the bounded interval $[0, 1]$ and that $0 < \epsilon \leq 1$. Then

$$\mathcal{M}_1(\epsilon, F, m) < 2b^{3(\lceil \log_2 y \rceil + 1)},$$

where $b = \lfloor 4/\epsilon \rfloor$ and, with $d = \text{fat}_F(\epsilon/8) \geq 1$,

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

Bounding with the Fat-Shattering Dimension

A Bound on the d_1-covering numbers in terms of the fat-shattering dimension

Theorem 18.2.

Let F be a set of real functions from a domain X to the bounded interval $[0, 1]$. Let $0 < \epsilon \leq 1$ and let $d = \text{fat}_F(\epsilon/8)$. Then for $m \geq d \geq 1$,

$$\mathcal{N}_1(\epsilon, F, m) < 2\left(\frac{4}{\epsilon}\right)^{3d \log_2(16em/(d\epsilon))}.$$

- It can be proved by Theorem 18.1 and the relationship between covering and packing numbers.

Bounding with the Fat-Shattering Dimension

Proof of Theorem 18.1.

- $Q_\alpha(F) = \{Q_\alpha(f) : f \in F\}$ where $Q_\alpha(f)(x) = \alpha \lfloor f(x)/\alpha \rfloor$.
- Fix $\epsilon, m, 0 < \alpha < \epsilon$.
- $\mathcal{M}_1(\epsilon, F, m) \leq \mathcal{M}_1(\epsilon - \alpha, Q_\alpha(F), m)$.
- $\text{fat}_{Q_\alpha(F)}(\epsilon) \leq \text{fat}_F(\epsilon - \alpha/2)$ for $\alpha < 2\epsilon$.
- Lemma 18.3 :
  $\mathcal{M}_1(3\epsilon/4, Q_{\epsilon/4}(F), m) \leq 2b^{3(\lceil \log_2 y \rceil + 1)}$ with $d = \text{fat}_{Q_{\epsilon/4}(F)}(\epsilon/4)$.
- This implies Theorem 18.1.

Bounding with the Pseudo-Dimension

- The fat-shattering dimension is always no more than the pseudo-dimension.
- If a function class F has finite pseudo-dimension, then Theorem 18.2 trivially yields an upper bound on covering numbers in terms of the pseudo-dimension.
- However, for classes of finite pseudo-dimension, a quite different bound can be obtained.
- We define the pseudo-metric $d_{L_1(P)}$ on the function class F by

$$d_{L_1(P)}(f,g) = \mathbb{E}(|f(x) - g(x)|) = \int |f(x) - g(x)| dP.$$

Theorem 18.5.

Let F be a nonempty set of real functions mapping from a domain X into the real interval $[0,1]$, and suppose F has finite pseudo-dimension d. Then,

$$\mathcal{M}(\epsilon, F, d_{L_1(P)}) < 2\left(\frac{2e}{\epsilon}\log\left(\frac{8e}{\epsilon}\right)\right)^d$$

for any probability distribution P on X, and for all $0 < \epsilon \leq 1$.

Bounding with the Pseudo-Dimension

$$\begin{aligned}\mathcal{M}_1(\epsilon, F, m) &= \max\{\mathcal{M}(\epsilon, F_{|x}, d_1) : x \in X^m\} \\ &= \max\{\mathcal{M}(\epsilon, F, d_{P_x}) : x \in X^m\} \\ &< 2\left(\frac{2e}{\epsilon}\log\left(\frac{8e}{\epsilon}\right)\right)^d\end{aligned}$$

where P_x is the distribution that is uniform on the entries of x and vanishes elsewhere.

- Hence, $\mathcal{M}_1(\epsilon, F, m) < 2\left(\frac{2e}{\epsilon}\log\left(\frac{8e}{\epsilon}\right)\right)^d$.

Theorem 18.4.

Let F be a nonempty set of real functions mapping from a domain X into the real interval $[0, 1]$ and suppose that F has finite pseudo-dimension d. Then,

$$\mathcal{N}_1(\epsilon, F, m) \leq \mathcal{M}_1(\epsilon, F, m) \leq e(d+1)\left(\frac{2e}{\epsilon}\right)^d$$

for all $\epsilon > 0$.

Comparing the Different Approaches

Bounding with the Fat-Shattering Dimension

- Theorem 12.8.

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

- Theorem 18.2.

$$\mathcal{N}_1(\epsilon, F, m) \leq \left(\frac{1}{\epsilon}\right)^{\text{fat}_F(\epsilon/8)\log_2(m/(\epsilon \text{fat}_F(\epsilon/8)))}$$

Bounding with the Pseudo-Dimension

- Theorem 12.2.

$$\mathcal{N}_1(\epsilon, F, m) \leq \mathcal{N}_\infty(\epsilon, F, m) \leq \left(\frac{m}{\epsilon}\right)^{\text{Pdim}(F)}$$

- Theorem 18.4.

$$\mathcal{N}_1(\epsilon, F, m) \leq \left(\frac{1}{\epsilon}\right)^{\text{Pdim}(F)}$$

Introduction

- Approximate-SEM algorithm : a function from $\cup_{m=1}^{\infty} Z^m \times \mathbb{R}^+$ to F s.t.

$$\hat{\text{er}}_z(\mathcal{A}(z,\epsilon)) < \inf_{f \in \mathcal{F}} \hat{\text{er}}_z(f) + \epsilon$$

- In chapter 16 : if a function class is totally bounded w.r.t. the L_∞ metric, then it is learnable by an algorithm derived from any approximate-SEM algorithm.
- In this chapter : if F has finite fat-shattering dimension, any approximate-SEM algorithm can be used to construct a learning algorithm for F.
- We also give lower bounds on the sample complexity of any learning algorithm, in terms of the fat-shattering dimension of the function class.
- A function class is learnable iff if it has finite fat-shattering dimension.

Classes with Finite Fat-Shattering Dimension

Theorem 19.1.

Suppose that F is a class of functions mapping from a domain X into the real interval $[0, 1]$, and suppose also that F has finite fat-shattering dimension.
Let $\mathcal{A}$ be any approximate-SEM algorithm for F and define, for $z \in Z^m$, $L(z) = \mathcal{A}(z, \epsilon_0/6)$, where $\epsilon_0 = 16/\sqrt{m}$.
Then, L is learning algorithm for F, and its sample complexity satisfies

$$m_L(\epsilon, \delta) \leq m_0(\epsilon, \delta) = \frac{256}{\epsilon^2}\left(18\text{fat}_F(\epsilon/256)\log^2\left(\frac{128}{\epsilon}\right) + \log\left(\frac{16}{\delta}\right)\right)$$

for all $\epsilon, \delta > 0$.

We say that the learning algorhtim L described in Theorem 19.1 is based on the approximate-SEM algorithm $\mathcal{A}$.

Classes with Finite Pseudo-Dimension

Theorem 19.2.

Suppose that F is a class of functions mapping from a domain X into the interval $[0, 1]$ of real numbers, and that F has finite pseudo-dimension.
Let $\mathcal{A}$ be any approximate-SEM algorithm for F and let L be as described in the statement of Theorem 19.1.
Then, L is a learning algorithm for F and its sample complexity is bounded as follows:

$$m_L(\epsilon, \delta) \leq m_0(\epsilon, \delta) = \frac{128}{\epsilon^2}\left(2\text{Pdim}(F)\log\left(\frac{34}{\epsilon}\right) + \log\left(\frac{16}{\delta}\right)\right)$$

for all $0 < \epsilon, \delta < 1$.

Results for Neural Networks

Corollary 19.3.

Suppose that a feed-forward network N has W weights and k computation units arranged in L layers.
Suppose that each computation unit has a fixed piecewise-polynomial activation function with p pieces and degree no more than l.
Let F be the class of functions computed by N.
Then any approximate-SEM algorithm for F can be used to define a learning algorithm for F, and for fixed p and l, the sample complexity of this learning algorithm is

$$O\left(\frac{1}{\epsilon^2}\left((WL\log W + WL^2)\log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$$

- By Theorem 8.8, Theorem 14.1 and Theorem 19.2.

Results for Neural Networks

Corollary 19.4.

Suppose that $b, L > 0$ and $s : \mathbb{R} \to [-b, b]$ 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

$$F = \left\{ \sum_{i=1}^{N} w_i f_i + w_0 : N \in \mathbb{N}, f_i \in F_1, \sum_{i=0}^{N} |w_i| \leq V \right\}$$

where

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

Then, any approximate-SEM algorithm can be used to define a learning algorithm L for F that has sample complexity satisfying

$$m_L(\epsilon, \delta) = O\left(\frac{1}{\epsilon^2} \left(\frac{V^6 B^2}{\epsilon^4} \log n + \log\left(\frac{1}{\delta}\right) \right) \right)$$

Lower Bounds

Theorem 19.5.

Suppose that F is a class of functions mapping from X to $[0,1]$. Then for $B \geq 2$, $0 < \epsilon < 1$ and $0 < \delta < 1/100$, any learning algorithm L for F has sample complexity satisfying

$$m_L(\epsilon, \delta, B) \geq \frac{\text{fat}_F(\epsilon/\alpha) - 1}{16\alpha},$$

for any $0 < \alpha < 1/4$.

- Finiteness of the fat-shattering dimension of a function class is a necessary and sufficient condition for the existence of a learning algorithm for the class.
- This result shows that

$$m(\epsilon, \delta, B) = \Omega\left(\frac{1}{\epsilon} + \text{fat}_F(4\epsilon)\right).$$

Remarks

It is easy to extend Theorem 19.1 and 19.2 to the case where the bound $B \geq 1$.

- Theorem 19.1.

$$m_L(\epsilon, \delta, B) \leq \frac{256B^4}{\epsilon^2}\left(18\text{fat}_F\left(\frac{\epsilon}{256B}\right)\log^2\left(\frac{128B}{\epsilon}\right) + \log\left(\frac{16}{\delta}\right)\right).$$

- Theorem 19.2.

$$m_L(\epsilon, \delta, B) \leq \frac{128B^4}{\epsilon^2}\left(2\text{Pdim}(F)\log\left(\frac{37B}{\epsilon}\right) + \log\left(\frac{16}{\delta}\right)\right).$$

Restricted Model

- We can define a restricted version of the learning framework for real prediction, in which the labelled examples presented to the learning algorithm are of the form $(x, f(x))$ for some $f \in F$.
- The following example describes the extreme case in which a single labelled example $(x, f(x))$ serves to uniquely identify the function f.

Example 19.6

For a positive integer d, let $S_0, ..., S_{d-1}$ be disjoint subsets of X with $\cup_j S_j = X$. Define the class of $[0, 1]$-valued functions

$$F_d = \{f_{b_0,...,b_{d-1}} : b_i \in \{0, 1\}, i = 0, ..., d-1\},$$

where

$$f_{b_0,...,b_{d-1}}(x) = \frac{3}{4} \sum_{j=0}^{d-1} 1_{S_j}(x) b_j + \frac{1}{8} \sum_{k=0}^{d-1} b_k 2^{-k}.$$

Clearly, for any $\gamma \leq 1/4$, $\text{fat}_{F_d}(\gamma) = d$. Hence, $F = \cup_{d=1}^{\infty} F_d$ has $\text{fat}_F(\gamma) = \infty$ for $\gamma \leq 1/4$, but any f in F can be identified from a single example $(x, f(x))$.

Restricted Model

- There are less restricted models that avoid these pathological cases.
  - The labels are noisy versions of the function values : labelled examples are of the form $(x, f(x) + \eta)$.
  - The labels are quantized versions of the function values.
- Typically, the fat-shattering dimension is the appropriate measure of complexity in such cases.
- We shall consider one such model at the end of the next chapter.

Relative Uniform Convergence Results

Theorem 19.7

Suppose that F is a set of $[0,1]$-valued functions defined on a set X and that P is a probability distribution on $Z = X \times [0,1]$. For $\alpha, \epsilon > 0$ and m a positive integer, we have

$$P^m\{\exists f \in F : \mathrm{er}_P(f) > (1+\alpha)\hat{\mathrm{er}}_z(f) + \epsilon\} \leq 4\mathcal{N}_1\left(\frac{\epsilon}{4(2+\alpha)}, F, 2m\right)\exp\left(\frac{-2m\alpha\epsilon}{(2+\alpha)^2}\right).$$

Theorem 19.8

For F and P as in Theorem 19.7, $\nu > 0$ and $0 < \beta < 1$,

$$P^m\left\{\exists f \in F : \frac{|\mathrm{er}_P(f) - \hat{\mathrm{er}}_z(f)|}{\mathrm{er}_P(f) + \hat{\mathrm{er}}_z(f) + \nu} > \beta\right\} \leq 4\mathcal{N}_1\left(\frac{\beta\nu}{8}, F, 2m\right)\exp\left(\frac{-m\nu\beta^2}{8}\right).$$