Neural Network Learning: Theoretical Foundation Chap. 4-5

Boyoung Kim

Seoul National University

July 14, 2017

Introduction : Learning by Minimizing Sample Error

- *Sample error minimization (SEM) algorithm* is any function $L : \bigcup_{m=1}^{\infty} Z^m \to H$ with the property :
  for any m and any $z \in Z^m$,
  $$L(z) = \underset{h \in H}{\operatorname{argmin}} \, \hat{er}_z(h).$$
- **Theorem 4.1** Suppose that H is a finite set of $\{0, 1\}$-valued functions. Then any SEM algorithm for H is a learning algorithm for H.
- Aim : The theorem also holds for many infinite function classes.
  $\Rightarrow$ If H has finite VC-dimension, the estimation error and sample complexity of any SEM algorithm can be bounded in terms of th VC-dimension of H.

Main theorem

- **Theorem 4.2** Suppose that H is a set of functions from a set X to $\{0,1\}$ and that H has finite VC dimension $d \geq 1$. Let L be any SEM algorithm for H. Then L is a learning algorithm for H. In particular, if $m \geq d/2$ then the estimation error of L satisfies

$$\epsilon_L(m,\delta) \leq \epsilon_0(m,\delta) = \left(\frac{32}{m}\left(d\ln\left(\frac{2em}{d}\right) + \ln\left(\frac{4}{\delta}\right)\right)\right)^{1/2}$$

and its sample complexity satisfies the inequality

$$m_L(\epsilon,\delta) \leq m_0(\epsilon,\delta) = \frac{64}{\epsilon^2}\left(2d\ln\left(\frac{12}{\epsilon}\right) + \ln\left(\frac{4}{\delta}\right)\right).$$

Uniform Convergence and Learnability

- The crucial step towards proving learnability is to obtain a result on the *uniform convergence* of sample errors to true errors.

- **Theorem 4.3** Suppose that H 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 $0 < \epsilon < 1$ and m a positive integer, we have

$$P^m\{|er_P(h) - \hat{er}_z(h)| \geq \epsilon \text{ for some } h \in H\} \leq 4\Pi_H(2m) exp\left(-\frac{\epsilon^2 m}{8}\right).$$

Proof of Uniform Convergence Result

- *Symmetrization* : bound the desired probability in terms of the probability of an event based on two samples.

- **Lemma 4.4** With the notation as above, let

$$Q = \{z \in Z^m : |er_P(h) - \hat{er}_z(h)| \geq \epsilon \text{ for some } h \in H\}$$

and

$$R = \{(r, s) \in Z^m \times Z^m : |\hat{er}_r(h) - \hat{er}_s(h)| \geq \frac{\epsilon}{2} \text{ for some } h \in H\}.$$

Then, for $m \geq 2/\epsilon^2$,

$$P^m(Q) \leq 2P^{2m}(R).$$

Proof of Uniform Convergence Result

- *Permutations* : involving a set of permutations on the labels of the double sample.
- Let Γ_m be the set of all permutations of $\{1, 2, \ldots, 2m\}$ that swap i and $m+i$. For instance, $\sigma \in \Gamma_3$ might give

$$\sigma(z_1, z_2, z_3, z_4, z_5, z_6) = (z_1, z_5, z_6, z_4, z_2, z_3).$$

- **Lemma 4.5** Let R be any subset of Z^{2m} and P any probability distribution on Z. Then

$$P^{2m}(R) = \mathbf{E} Pr(\sigma z \in R) \leq \max_{z \in Z^{2m}} Pr(\sigma z \in R),$$

  where the expectation is over z chosen according to P^{2m}, and the probability is over σ chosen uniformly from Γ_m.
- proof) For any $\sigma \in \Gamma_m$, $P^{2m}(R) = P^{2m}\{z : \sigma z \in R\}$.

Proof of Uniform Convergence Result

- *Reduction to a finite class* : reduce the problem to one involving a finite function class.
- **Lemma 4.6** For the set $R \subseteq Z^{2m}$ defined in Lemma 4.4, and permutation σ chosen uniformly at random from Γ_m,

$$\max_{z \in Z^{2m}} Pr(\sigma z \in R) \leq 2\Pi_H(2m) exp\left(-\frac{\epsilon^2 m}{8}\right).$$

- proof) let $S = \{x_1, \ldots, x_{2m}\}$ and $t = |H_{|s}|$, then $t \leq \Pi_H(2m)$. Then there are functions $h_1, \ldots, h_t \in H$. And use Hoeffding's lemma.

Application to the Perceptron

- Since n-input perceptron has a finite VC-dimension of $n+1$, as shown in chaper 3,
- We immediately get an estimation error bound and sample complexity bound for a SEM algorithm from theorem 4.2.

The Restricted Model

- t is called target function if $P\{(x, t(x)) : x \in X\} = 1$.

- **Theorem 4.8** Suppose that H is a set of functions from a set X to $\{0, 1\}$ and that H has finite VC dimension $d \geq 1$. Let L be such that for any m and for any $t \in H$, if $x \in X^m$ and z is the training sample corresponding to x and t, then the hypothesis $h = L(z)$ satisfies $h(x_i) = t(x_i)$ for $i = 1, 2, \ldots, m$. Then L is a learning algorithm for H in the restricted model, with sample complexity

$$m_L(\epsilon, \delta) \leq \frac{4}{\epsilon}\left(d \ln\left(\frac{12}{\epsilon}\right) + \ln\left(\frac{2}{\delta}\right)\right)$$

and with estimation error

$$\epsilon_L(m, \delta) \leq \frac{2}{m}\left(d \ln\left(\frac{2em}{d}\right) + \ln\left(\frac{2}{\delta}\right)\right).$$

- Such an algorithm in the theorem constitutes a SEM algorithm.

A better uniform convergence result

- Theorem 4.3 is not the best uniform convergence result that can be obtained, nor is the learnability result in Theorem 4.2.

- **Theorem 4.10** There is a positive constant c such that the following holds. Suppose that H is a set of functions from a set X to $\{0, 1\}$ and that H has finite VC dimension $d \geq 1$. Let L be any SEM algorithm for H. Then L is a learning algorithm for H and its sample complexity satisfies the inequality

$$m_L(\epsilon, \delta) \leq m'_0(\epsilon, \delta) = \frac{c}{\epsilon^2}\left(d + \ln\left(\frac{1}{\delta}\right)\right).$$

- $m_0(\epsilon, \delta)$ of Theorem 4.2 contains an additional $\ln(1/\epsilon)$ term multiplying the VC-dimension.

A better uniform convergence result

- proof) Use the following Lemma 4.11, which is the improvement of Lemma 4.6.
- **Lemma 4.11** For the set $R \subseteq Z^{2m}$ defined in Lemma 4.4, and permutation σ chosen uniformly at random from Γ_m, if $m \geq 400(VCdim(H)+1)/\epsilon^2$, then

$$\max_{z \in Z^{2m}} Pr(\sigma z \in R) \leq 4 \cdot 41^{VCdim(H)} exp\left(-\frac{\epsilon^2 m}{576}\right).$$

Introduction : Goals of This chapter

- Provide lower bounds on the estimation error and sample complexity of any learning algorithm in terms of the VC-dimension of the class.
- These lower bounds are not vastly different from the upper bounds of the previous chapter.
- A function class is learnable if and only if it has finite VC-dimension.

A technical lemma

- **Lemma 5.1** Suppose that α is a random variable uniformly distributed on $\{\alpha_{\text{-}}, \alpha_+\}$, where $\alpha_{\text{-}} = 1/2 - \epsilon/2$ and $\alpha_+ = 1/2 + \epsilon/2$, with $0 < \epsilon < 1$. Suppose that $\xi_1, \ldots, \xi_m$ be i.i.d. $\{0,1\}$-valued random variables with $Pr(\xi_i = 1) = \alpha$ for all i. Let f be a function from $\{0,1\}^m$ to $\{\alpha_{\text{-}}, \alpha_+\}$. Then

$$P(f(\xi_1, \ldots, \xi_m) \neq \alpha) > \frac{1}{4}\left(1 - \sqrt{1 - exp\left(\frac{-2\lceil m/2\rceil \epsilon^2}{1-\epsilon^2}\right)}\right).$$

Hence, if this probability is no more than δ, where $0 < \delta < 1/4$, then

$$m \geq 2\left\lfloor \frac{1-\epsilon^2}{2\epsilon^2} \ln\left(\frac{1}{8\delta(1-2\delta)}\right)\right\rfloor.$$

The general lower bound

- **Theorem 5.2** Suppose that H is a class of $\{0,1\}$-valued functions and that H has VC dimension d. For any learning algorithm L for H, the sample complexity $m_L(\epsilon,\delta)$ of L satisfies
$$m_L(\epsilon,\delta) \geq \frac{d}{320\epsilon^2}$$
for all $0 < \epsilon, \delta < 1/64$. Furthermore, if H contains at least two functions, we have
$$m_L(\epsilon,\delta) \geq 2\left\lfloor \frac{1-\epsilon^2}{2\epsilon^2} \ln\left(\frac{1}{8\delta(1-2\delta)}\right)\right\rfloor$$
for all $0 < \epsilon < 1$ and $0 < \delta < 1/4$.

The Restricted Model

- **Theorem 5.3** Suppose that H is a class of $\{0,1\}$-valued functions and that H has VC dimension d. For any learning algorithm L for H in restricted model, the sample complexity $m_L(\epsilon, \delta)$ of L satisfies

$$m_L(\epsilon, \delta) \geq \frac{d-1}{32\epsilon}$$

for all $0 < \epsilon < 1/8$ and $0 < \delta < 1/100$. Furthermore, if H contains at least two functions, we have

$$m_L(\epsilon, \delta) > \frac{1}{2\epsilon} \ln\left(\frac{1}{\delta}\right)$$

for all $0 < \epsilon < 3/4$ and $0 < \delta < 1$.

VC-Dimension Quantifies Sample Complexity and Estimation Error

- inherent sample complexity is $m_H(\epsilon, \delta) = \min_L m_L(\epsilon, \delta)$.

- **Theorem 5.4** Suppose that H is a set of functions that map from a set X to $\{0, 1\}$. Then H is learnable if and only if it has finite VC dimension. Furthermore, there are constants $c_1, c_2 > 0$ such that the inherent sample complexity of the learning problem for H satisfies

$$\frac{c_1}{\epsilon^2}\left(VCdim(H) + \ln\left(\frac{1}{\delta}\right)\right) \leq m_H(\epsilon, \delta) \leq \frac{c_2}{\epsilon^2}\left(VCdim(H) + \ln\left(\frac{1}{\delta}\right)\right)$$

  for all $0 < \epsilon < 1/40$ and $0 < \delta < 1/20$.

- proof) Combine theorem 5.2 and 4.10.

- if L is a SEM algorithm for H, then its sample complexity satisfies these inequalities, and so its estimation error grows as $\sqrt{VCdim(H) + \ln(1/\delta)/m}$.

VC-Dimension Quantifies Sample Complexity and Estimation Error

- **Theorem 5.5** For a class H of functions mapping from a set X to $\{0, 1\}$, the following statements are equivalent.
  (1) H is learnable.
  (2) The inherent sample complexity of H, $m_H(\epsilon, \delta)$, satisfies
  $$m_H(\epsilon, \delta) = \Theta\left(\frac{1}{\epsilon^2} \ln\left(\frac{1}{\delta}\right)\right).$$
  (3) The inherent estimation error of H, $\epsilon_H(m, \delta)$, satisfies
  $$\epsilon_H(m, \delta) = \Theta\left(\sqrt{\frac{1}{m} \ln\left(\frac{1}{\delta}\right)}\right).$$
  (4) $VCdim(H) < \infty$.

VC-Dimension Quantifies Sample Complexity and Estimation Error

- **Theorem 5.5**(*continued*)
  - (5) The growth function of H, $\Pi_H(m)$, is bounded by a polynomial in m.
  - (6) H has the following uniform convergence property: There is a function $\epsilon_0(m, \delta)$ satisfying
    - for every probability distribution P on $X \times \{0, 1\}$,

$$P^m \left\{ \sup_{h \in H} |er_P(h) - \hat{er}_z(h)| > \epsilon_0(m, \delta) \right\} < \delta,$$

    - $\epsilon_0(m, \delta) = \Theta\left(\sqrt{(1/m)\ln(1/\delta)}\right).$

- $\Theta(\cdot)$ notation indicates the functions are asymptotically withen a constant factor of each other.

VC-Dimension Quantifies Sample Complexity and Estimation Error

- H is learnable in the restricted model iff H has finite VC dimension.
- And the inherent sample complexity of the restricted learning problem for H satisfies

$$\frac{c_1}{\epsilon}\left(VCdim(H) + \ln\left(\frac{1}{\delta}\right)\right) \leq m_H(\epsilon, \delta) \leq \frac{c_2}{\epsilon}\left(VCdim(H) + \ln\left(\frac{1}{\delta}\right)\right)$$

for some constants $c_1, c_2 > 0$.