Neural Network Learning: Theoretical Foundation
Chap. 2-3

Gyuseung Baek

July 8, 2017

Supervised learning issues

- Approximation
- Estimation
- Computational efficiency

Terminology

- $x \in X$: input, $y \in Y = \{0, 1\}$: output
- P : probability distribution on $Z = X \times Y$ (fixed, unknown)
- $z = ((x_1, y_1), \cdots, (x_m, y_m)) = (z_1, \cdots, z_m) \in Z^m$

- Aim : Finding the most appropriate function $h \in H : X \to Y$

- error of h w.r.t. P

$$er_P(h) = P\{(x, y) \in Z : h(x) \neq y\}$$

- *Approximation error* of the class H

$$opt_P(H) = \inf_{g \in H} er_P(g)$$

Learning algorithm

- H : class of functions from X to Y
- A *learning algorithm* L for H is a function

$$L : \bigcup_{m=1}^{\infty} Z^m \to H$$

with the following property :
for any $\epsilon \in (0,1)$ and $\delta \in (0,1)$, $\exists m_0(\epsilon, \delta)$ s.t.
if $m \geq m_0(\epsilon, \delta)$, then for any probability distribution P on Z,

$$P^m\{er_P(L(z)) < opt_P(H) + \epsilon\} \geq 1 - \delta$$

- $L(z)$ is called a *hypoothesis* of the learning algorithm L

Learning algorithm(Continued)

- L is a learning algorithm if $\exists \epsilon_0(m, \delta)$ s.t., for all m, δ and P, w.p. at least $1 - \delta$ over $z \in Z^m$ chosen according to P^m,

$$er_P(L(z)) < opt_P(H) + \epsilon_0(m, \delta)$$

- $\epsilon_0(m, \delta)$: *estimation error bound* for the algorithm L
- $m_0(\epsilon, \delta)$: *sufficient sample size* for (ϵ, δ)-learning H by L

Measures of Learning algorithm

- *sample complexity* function $m_L(\epsilon, \delta)$ of L:

  $m_L(\epsilon, \delta) = \min\{m : m \text{ is a sufficient sample size for } (\epsilon, \delta)\text{-learning } H \text{ by } L\}$

- *estimation error* $\epsilon_L(m, \delta)$ of L : the smallest possible estimation error bound

- *inherent sample complexity* $m_H(\epsilon, \delta)$ of H:

$$m_H(\epsilon, \delta) = \min_L m_L(\epsilon, \delta)$$

  provies an absolute lower bound on the size of sample needed to (ϵ, δ)-learn
- The inherent estimation error $e_H(m, \delta)$ may be defined similary

Learning finite function classes

- Goal : If H is finite, then

$$L(z) = \underset{h \in H}{\operatorname{argmin}}\, \hat{er}_z(h)$$

  is a learning algorithm

- ex). binary output defined on a finite input, function having weights that can take only a finite number of values

Learning finite function classes

- **Thm** Suppose that h is a function from X to $\{0,1\}$. Then

$$P^m\{|\hat{er}_z(h) - er_P(h)| \geq \epsilon\} \leq 2\exp(-2\epsilon^2 m)$$

for any probability distribution P, any ϵ, m

- *proof* : Use Hoeffding's inequality

- **Corr** Suppose that H is a finite set of functions from X to $\{0,1\}$. Then

$$P^m\{\max_{h\in H}|\hat{er}_z(h) - er_P(h)| \geq \epsilon\} \leq 2|H|\exp(-2\epsilon^2 m)$$

for any probability distribution P, any ϵ, m

Learning finite function classes

- **Thm** Suppose that H is a finite set of functions from X to $\{0,1\}$. Let $L : \bigcup_{m=1}^{\infty} Z^m \to H$ be such that for any m and any $z \in Z^m$,

$$L(z) = \operatorname*{argmin}_{h \in H} \hat{er}_z(h)$$

Then L is a learning algorithm for H, with estimation error

$$\epsilon_L(m, \delta) \leq \left(\frac{2}{m} \log \left(\frac{2|H|}{\delta} \right) \right)^{1/2}$$

and sample complexity

$$m_L(\epsilon, \delta) \leq \frac{2}{\epsilon^2} \log \left(\frac{2|H|}{\delta} \right)$$

Restricted Model

- t is called the *target function* if $P\{(x, t(x)) : x \in X\} = 1$
- **Thm** If the target function $t \in H$, then $L(z) = \text{argmin}_{h \in H} \hat{er}_z(h)$ is a learning algorithm for H, with sample complexity

$$m_L(\epsilon, \delta) \leq \frac{1}{\epsilon} \log \left(\frac{|H|}{\delta} \right)$$

Introduction

- Unfortunately, many interesting function classes are not finite
- We shall see that learning is possible for many function classes if the function class is not too *complex*
- How do we measure the complexity of function spaces?

The Growth Function

- S : finite subset of the input space X
- $H_{|S}$: restriction of H to the set S

- The growth function of H, $\Pi_H : \mathbb{N} \to \mathbb{N}$, is defined as

$$\Pi_H(m) = \max\{|H_{|S}| : S \subseteq X \text{ and } |S| = m\}$$

- $\Pi_H(m) \leq 2^m$ for all m
- if H is finite, then $\Pi_H(m) \leq |H|$ for all m, and $\Pi_H(m) = |H|$ for sufficiently large m

The Growth Function of simple perceptron

- A simple perceptron computes a function f of the form

$$f(x) = \text{sgn}(w \cdot x - \theta)$$

- **Thm** Let N be the real-weight simple perceptron with $n \in \mathbb{N}$ real inputs and H the set of functions it computes. Then

$$\Pi_H(m) = 2\sum_{k=0}^{n} \binom{m-1}{k}$$

The Growth Function of simple perceptron

- For a subset $S \subseteq \mathbb{R}^n$, $CC(S)$ denote the number of path-connercted components of S

- **lemma1** For a set $S = \{x_1, \cdots, x_m\} \subseteq \mathbb{R}^n$, let $P_1, \cdots, P_m$ be the hyperplanes given by

$$P_i = \{(w, \theta) \in \mathbb{R}^{n+1} : w^\top x_i - \theta = 0\}$$

Then

$$|H_{|S}| = CC\left(\mathbb{R}^{n+1} - \bigcup_{i=1}^{m} P_i\right)$$

The Growth Function of simple perceptron

- **lemma2** For $m, n \in \mathbb{N}$, suppose $T = \{z_1, \cdots, z_m\} \subseteq \mathbb{R}^{n+1}$ has every subset of no more than $n+1$ points linearly independenty. Let $P_i = \{v \in \mathbb{R}^{n+1} : v^\top z_i = 0\}$ for $i = 1, \cdots, m$ and define

$$C(T) = CC\left(\mathbb{R}^{n+1} - \bigcup_{i=1}^{m} P_i\right)$$

Then $C(T)$ depends only on m and n, so we can write $C(T) = C(m, n)$, and for all m and n, we have

$$C(m, n) = 2\sum_{k=0}^{n} \binom{m-1}{k}$$

The Vapnik-Chervonenkis Dimension

- If H can compute all dichotomies of S, i.e. $|H_{|S}| = 2^m$, we say that H *shatters* S
- The Vapnik-Chervonenkis dimension (VC-dimension) of H is the size of the largest shattered subset of X (or infinity, if it does not exist)
- Equivalently, the VC-dimension of H is the largest value of m for which the growth function $\Pi_H(m)$ equals 2^m
- For the perceptron, $\text{VCdim}(H) = n + 1$

The Vapnik-Chervonenkis Dimension

- **Thm** Let N be the real-weight simple perceptron with $n \in \mathbb{N}$ real inputs. Then a set $S = \{x_1, \cdots, x_m\}$ is shatterd by $H(= H_N)$ iff S is affinely independent, i.e. $\{(x_1^\top, -1), \cdots, (x_m^\top, -1)\}$ is lenearly independent.

- **Thm** Suppose F is a vector space of real-valued functions, g is a real-valued function, and $H = \{\text{sgn}(f + g) : f \in F\}$. Then VCdim$(H)$=dim$(F)$.

The Vapnik-Chervonenkis Dimension

- **Thm** For a function class H with VCdim(H)=d,

$$\Pi_H(m) \leq \sum_{i=0}^{d} \binom{m}{k}$$

for all positive intergers m

- **Thm** For $m \geq d \geq 1$,

$$\sum_{i=0}^{d} \binom{m}{i} < \left(\frac{em}{d}\right)^d$$

Hence, for a function class H with VCdim(H)=d,

$$\Pi_H(m) \begin{cases} = 2^m & \text{if } m \leq d \\ < \left(\frac{en}{d}\right)^d & \text{if } m > d \end{cases}$$

and, for $m \geq 1$, $\Pi_H(m) \leq m^d + 1$