Statistics Learning Theory:VC-Dimension and Linear Predictor

2019.02.16

Introduction

In today's slide, we aim to introduce the followings:

1. VC-dimension and the fundamental theorem of PAC learning(The proof will be introduced in the nex week)
2. Linear predictor and their VC-dimension

Motivation:Infinite-Size Classes

In the last week, we prove that finite classes are learnable. However, the infinite size classes may be learnable. Consider the following example.

Motivation:Example

Let $\mathcal{H}$ be the set of threshold functions over the real line, namely, $\{h_a : a \in R\}$, where $h_a(x) = I_{[x<a]}$. Then $\mathcal{H}$ is infinite, however it can be proved is learnable in the PAC model using the ERM algorithm.

VC-Dimension

The natural question arises: what is the sufficient conditions for learnability?
Answer: VC-dimension

VC-dimension

Definition (Restriction of $\mathcal{H}$ to C)

Let $\mathcal{H}$ be a class of functions from X to $\{0,1\}$ and let $C = \{c_1, \cdots, c_m\} \subset X$. The restriction of $\mathcal{H}$ to C is the set of functions from C to $\{0,1\}$ that can be derived from $\mathcal{H}$. That is,

$$\mathcal{H}_C = \{h(c_1), \cdots, h(c_m)) : h \in \mathcal{H}\}$$

Definition (Shattering)

A hypothesis class $\mathcal{H}$ shatters a finite set $C \subset X$ if the restriction of $\mathcal{H}$ to C is the set of all functions from C to $\{0,1\}$. That is, $|\mathcal{H}_C| = 2^{|C|}$

VC-dimension

Definition (VC-dimension)

The VC-dimension of a hypothesis class $\mathcal{H}$, denoted VCdim($\mathcal{H}$), is the maximalsize of a set $C \subset X$ that can be shattered by $\mathcal{H}$. If $\mathcal{H}$ can shatter sets of arbitraily large size we say that $\mathcal{H}$ has infinite VC-dimension.

VC-dimension:No Free Lunch Theorm

By No Free Lunch Theorem, given a training sample S with size m, if there exists a shattered set of size $2m$, then we can find a distribution $\mathcal{D}$ such that $L_{\mathcal{D}}(A(S)) \geq 1/8$ with probability at least $1/7$

Therefore, we have the following theorem

Theorem

Let $\mathcal{H}$ be a class of infinite VC-dimension. Then, $\mathcal{H}$ is not PAC learnable

VC-dimension:Examples

To show that $VCdim(\mathcal{H}) = d$ we need to show that

1. There exists a set C of size d that is shattered by $\mathcal{H}$
2. Every set C of size $d+1$ is not shattered by $\mathcal{H}$

Examples:Threshold Functions

$C = \{c_1\}$, $\mathcal{H}$ shatters C, therefore, $VCdim(\mathcal{H}) \geq 1$. If an arbitrary set $C = \{c_1, c_2\}$ where $c_1 \leq c_2$, $\mathcal{H}$ does not shatter C.
$VCdim(\mathcal{H}) = 1$

Examples:Intervals

Let $\mathcal{H}$ be the class of intervals over R, namely, $\mathcal{H} = \{h_{a,b} : a, b \in R, a < b\}$, where $h_{a,b} : R \to \{0, 1\}$ is a function such that $h_{a,b}(x) = 1_{x \in (a,b)}$. If $C = \{1, 2\}$. Then $\mathcal{H}$ shatters C and therefore $VCdim(\mathcal{H}) \geq 2$. However $C = \{c_1, c_2, c_3\}$. The labeling $(1, 0, 1)$ cannot be obtained. $VCdim(\mathcal{H}) = 2$.

Eamples: Finite case

Let $\mathcal{H}$ be finite case, then for any set C, we have $|\mathcal{H}_C| \leq |\mathcal{H}|$
Thus, C cannot be shattered if $|\mathcal{H}| < 2^{|C|}$, hence,
$VCdim(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$

The Fundamental Theorem of PAC learning

Theorem (The Fundamental Theorem of Statistical Learning)

Let $\mathcal{H}$ be a hypothesis class of functions from a domain X to $\{0, 1\}$ and let the loss function be the 0-1 loss. Then, the following are equivalent:

1. *$\mathcal{H}$ has the uniform convergence property*
2. *Any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$*
3. *$\mathcal{H}$ is agnostic PAC learnable*
4. *$\mathcal{H}$ is PAC learnable*
5. *Any ERM rule is a successful PAC learner for $\mathcal{H}$*
6. *$\mathcal{H}$ has a finite VC-dimension*

Remark:Reminding the quantitative Version, the proof idea:the growth rate of $\mathcal{H}_C$ is ploynomial, by Hoeffding's inequality, the converge rate is exponentially with $|C|$

Linear predictor

The basic idea of linear predictor is to use the linear function to predict the target.

1. Classification: Logistic Regression
2. Regression: Linear regression

Linear predictor:hypothesis classes

$$L_d\{h_{w,b} : w \in R^d, b \in R\}$$

where

$$h_{w,b}(x) = \langle w, x\rangle + b = (\sum_{i=1}^{d} w_i x_i) + b$$

$$HS_d = sign \circ L_d = \{sign(h_{w,b}(x)) : h_{w,b} \in L_d\}$$

Note that: we can embedding nonhomogenous linear function in R^d into homogenous inear function in R^{d+1}

Linear predictor:implementation of ERM rule

1. In the realizable case (PAC case), ERM rule can be solved efficient
2. In the agnostic case, implementing the ERM rule is computationally hard(Ben-David & Simon 2001)
3. Due to the computation difficulty, the Logistic regression uses the surrogate loss function to learn a halfspace that does not necessarily minimize the empirical risk with the 0-1 loss.
4. In the following, we will introduce the two way to implement the ERM rule(relizable case), and prove the learnability of the algorithm.

Implementation of ERM rule: Linear Programming

Linear programs(LP) are problems that can be expressed as maximizing a linear function subject to linear inequalities. That is,

$$\max_{w \in R^d} \langle u, w \rangle$$

subject to

$$Aw \geq v$$

Linear programs can be solved efficiently, we will show that the ERM problem for halfspace in the relizable case can be expressed as a linear program.

Implementation of ERM rule: Linear Programming

Let $S = \{(x_i, y_i)\}_{i=1}^{m}$ be a training set of size m.
Since we assume the realizable case, an ERM predictor should have zero errors on the training set. That is we are looking some vector $w \in R^d$ for which

$$sign(\langle w, x \rangle) = y_i$$

Equivalently

$$y_i \langle w, x \rangle > 0$$

Implementation of ERM rule: Linear Programming

Let w^* be a vector that satisfies this condition(existence is ensured by realizability assumption). Define $\gamma = \min_i(y_i\langle w^*, x_i\rangle)$ and $\bar{w} = \frac{w^*}{\gamma}$, then we have

$$y_i\langle \bar{w}, x_i\rangle = \frac{1}{\gamma} y_i\langle w^*, x_i\rangle \geq 1$$

Therefore, there exists a vector that satisfies

$$y_i\langle w, x_i\rangle \geq 1 \tag{1}$$

and clearly, such a vector is an ERM predictor.

Implementation of ERM rule: Linear Programming

To find a vector that satisfies equation 1 we can rely on an LP solver as follows. Set A to be the $m \times d$ matrix whose rows are the instances multiplied by y_i. That is, $A_{i,j} = y_i x_{i,j}$. Let v be the vector $(1, \cdots, 1) in R^m$. Then the equation 1 can be rewritten as

$$Aw = \begin{bmatrix} y_1 x_{1,1} & y_1 x_{1,2} & \cdots & y_1 x_{1,d} \\ \vdots, & \vdots & \vdots & \vdots \\ y_m x_{m,1} & y_m x_{m,2} & \cdots & y_m x_{m,d} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix} \geq \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = v$$

The LP form requires a maximization objective yet all the w that satisfy the constraints are equal candidates as output hypothesis. Thus, we set a dummy objective, $u = (0, \cdots, 0) \in R^d$

Implementation of ERM rule: Batch Preception

Algorithm 1: Batch Perceptron

Input :← A training set$(x_1, y_1), \cdots (x_m, y_m)$;

Initialize :← $w^{(1)} = (0, 0, \cdots, 0)$;

while $\exists i \;\; s.t. \;\; y_i \langle w^{(t)}, x_i \rangle \leq 0$ **do**

$\quad w^{(t+1)} = w^{(t)} + y_i x_i$;

end while

Output : $w^{(t)}$

Implementation of ERM rule: Batch Preception

Theorem

Assume that $(x_1, y_1), \cdots, (x_m, y_m)$ *is separable, let* $B = \min\{|w| : \forall i \in [m], \quad y_i \langle w, x_i \rangle \geq 1\}$ *and let* $R = \max_i |x_i|$. *Then, the Preceptron algorithm stops after at most* $(RB)^2$ *iterations, and when it stops it holds that* $\forall i \in [m], y_i \langle w^{(t)}, x_i \rangle > 0$

Batch Preception:proof

By the definition of the stopping condition, if the Perceptron stops it must have separated all the examples. We will show that if the Perceptron runs for T iterations, then we must have $T \leq (RB)^2$. Let w^* be a vector that achieves the minimum in the definition of B. That is,$y_i \langle w^*, x_i \rangle \geq 1$ for all i, and among all vectors that satisfy these constraints, w^* is of minimal norm.

Batch Preception:proof

We claim that

$$\frac{\langle w^*, w^{(T+1)} \rangle}{|w^*||w^{(T+1)}|} \geq \frac{\sqrt{T}}{RB}$$

by Cauchy-Schwartz inequality, it will imply that

$$1 \geq \frac{\sqrt{T}}{RB}$$

hence

$$T \leq (RB)^2$$

Therefore, we focus on proving the above inequality.

Batch Preception:proof

We first show that $\langle w^*, w^{(T+1)} \rangle \geq T$.
For $t = 1$, $w^{(1)} = (0, \cdots, 0)$ holds, suppose that on iteration t, we have that

$$\langle w^*, w^{(t+1)} \rangle - \langle w^*, w^{(t)} \rangle = \langle w^*, w^{(t+1)} - w^{(t)} \rangle$$

$$= \langle w^*, y_i x_i \rangle = y_i \langle w^*, x_i \rangle$$

$$\geq 1$$

Therefore, after T iterations, we get

$$\langle w^*, w^{(T+1)} \rangle = \sum_{t=1}^{T} (\langle w^*, w^{(t+1)} \rangle - \langle w^*, w^{(t)} \rangle) \geq T \qquad (2)$$

Batch Preception:proof

Next, we upper bound $|w^{(T+1)}|$

$$|w^{(T+1)}|^2 = |w^{(t)} + y_i x_i|^2$$

$$|w^{(t)}|^2 + 2y_i\langle w^{(t)}, x_i\rangle + y_i^2|x_i|^2$$

$$|w^{(t)}|^2 + R^2$$

Using above recursively for T iterations, we obtain that

$$|w^{(T+1)}|^2 \leq TR^2 \rightarrow |w^{(T+1)}| \leq \sqrt{T}R \tag{3}$$

Batch Preception:proof

Combining equation 2 with equation 3, we have that

$$\frac{\langle w^*, w^{(T+1)} \rangle}{|w^*||w^{(T+1)}|} \geq \frac{T}{B\sqrt{T}R} = \frac{\sqrt{T}}{BR}$$

Halfspace:VC-dimension

Theorem

The VC-dimension of the class of homogenous halfspace in R^d is d

Theorem

The VC-dimension of the class of nonhomogenous halfspace in R^d is $d+1$