PAC Learning Theory

Kangcheng Hou

Zhejiang University

May 30, 2018

Introduction

We will introduce the basic of computational learning theory.
There are quite a lot concepts, so hold on tight.

Definition I

Genralization error

$$E(h;\mathcal{D}) = P_{\mathbf{x}\sim\mathcal{D}}(h(\mathbf{x}) \neq y)$$

Empirical Risk

$$\hat{E}(h;\mathcal{D}) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(h(\mathbf{x}_i) \neq y_i)$$

Useful Tools I

Jensen inequality: for convex function f

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$$

Hoeffding inequality: if $x_1, \ldots, x_m$ are m independent random variable and $0 \leq x_i \leq 1$, then

$$P(\frac{1}{m}\sum_{i=1}^{m} x_i - \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[x_i] \geq \epsilon) \leq \exp(-2m\epsilon^2)$$

McDiarmid inequality: if $x_1, \ldots, x_m$ are m independent random variable, and for $1 \leq i \leq m$, the function f satisfies

$$\sup_{x_1,\ldots,x_m,x_i'} |f(x_1, \ldots, x_m) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_m)| \leq c_i$$

Useful Tools II

Then we have

$$P(f(x_1, \ldots, x_m) - \mathbb{E}[f(x_1, \ldots, x_m)] \geq \epsilon) \leq \exp(\frac{-2\epsilon^2}{\sum_i c_i^2})$$

Note that Hoeffding inequation can be seen as a special case of McDiarmid inequality. If we assign

$$f(x_1, \ldots, x_m) = \frac{1}{m} \sum_{i=1}^{m} x_i$$

then we have $c_i = \frac{1}{m}$, thus the Hoeffding inequality holds.

PAC Learning I

The concept c, is the target of our learning, it can be seen as ground truth. And the set of all concepts form a concept class $\mathcal{C}$. And we also have some learning algorithm $\mathcal{L}$ that assumes some range of possible hypothesis $\mathcal{H}$. For example, we may want to predict tomorrow's temperature based on the temperature of the past week. We may consider two models, the first one is Gaussian process, the second one is based on some if-else and output three discrete values, say, 15, 20, 25 degree. We may say the hypothesis space $\mathcal{H}_1$ of the first model is too large, and the hypothesis space $\mathcal{H}_2$ is too small because we know the concept $\mathcal{C}$ should be some continuous value neither too small nor too large.

PAC Learning II

PAC identify: There exists learning algorithm $\mathcal{L}$ which can produce final hypothesis $h \in \mathcal{H}$ that satisfies

$$P(E(h) \leq \epsilon) \geq 1 - \delta$$

This means the learning algorithm $\mathcal{L}$ find the right hypothesis h from $\mathcal{H}$ at error ϵ with probability $1 - \delta$.
PAC learnable: the sample number m needed to learn in the hypothesis space is a polynomial w.r.t $\frac{1}{\epsilon}, \frac{1}{\delta}, \text{size}(\mathbf{x}), \text{size}(c)$.
efficiently PAC learnable: the running time is also polynomial.

Finite hypothesis space I

We will now discuss

VC dimension I

The VC dimension can be defined as

$$VC(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\}$$

which is the max sample size m that can be scattered by model $\mathcal{H}$. This is to say, if there **exists** dataset of size d that can be **scattered** by $\mathcal{H}$, but there does not exist any dataset of size $d+1$ that can be **scattered** by $\mathcal{H}$, then we say the VC dimension of model $\mathcal{H}$ is d.
Theorem tells us that

$$P(|E(h) - \hat{E}(h)| > \epsilon) \leq 4\Pi_{\mathcal{H}}(2m)\exp(-\frac{m\epsilon^2}{8})$$

VC dimension II

There are close relation between the growth function and the VC dimension. If the VC dimension of the hypothesis space $\mathcal{H}$ is d, then for any $m \in \mathbb{N}$, we have

$$\Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$$

From this conclusion, we can get a bound For a hypothesis space $\mathcal{H}$ of VC dimension d,

$$\Pi_{\mathcal{H}}(m) \leq (\frac{e \cdot m}{d})^d$$

This means the number of different configuration of a hypothesis space $\mathcal{H}$ of VC dimension d is bounded by $(\frac{e \cdot m}{d})^d$. With this, we can further have

agnostic PAC learnable

When the target concept $c \notin \mathcal{H}$, then it is imposssible for learning algorithm $\mathcal{L}$ to learn the ϵ approximation of the target concept c. But there exists a hypothesis that minimizes the generalization error in the hypothesis space $\mathcal{H}$. The goal of **agnostic learning** is to learning a hypothesis h in $\mathcal{H}$ such that

$$P(E(h) - \min_{h' \in \mathcal{H}} E(h') \leq \epsilon) \geq 1 - \delta$$

Knowing the concept of agnostic PAC learnable, we have the following statement, every hypothesis space $\mathcal{H}$ with finite VC dimension is agnostic PAC learnable. The target of learning is

$$E(g) = \min_{h \in \mathcal{H}} E(h)$$

We want to prove that the output h of the learning algorithm is close to the g which has the minimal generalization error in the hypothesis space $\mathcal{H}$.

Rademacher Complexity