A Deeper Understanding Of Growth Function & VC Dimension And Their Functions In Learning

Zhiyuan Zhang 15220162202517

March 13, 2019

1 Introduction[1]

As we know, according to Hoeffding's inequality, in-sample error converges to out-of sample error as sample size N increases.

$$Pr\left(|E_{in}(h) - E_{out}(h)| > \epsilon\right) \leq 2e^{-2\epsilon^2 N}$$

And if a hypothesis set contains more than one candidates functions, it allows a higher probability of difference in two kinds of errors.

$g \in H = \{h_1, h_2, \cdots h_M\}$

$$Pr\left(|E_{in}(g) - E_{out}(g)| > \epsilon\right) \leq 2|H|e^{-2\epsilon^2 N}$$

where $|H| = M$ is the size of H.

So we can see that, this inequality is suitable for finite learning only. But in practice, hypothesis sets typically contains infinite candidate functions, like a liner fitting model contains infinite number of points. When H converges to infinity, this inequality will lose all its instructive feature in learning. So obviously, when H actually goes to infinity, we need an alternative way to generalize $|H|$, to make further progress in learning. And this leads to the idea of growth function and VC dimension.

2 Example

Let's consider a hypothesis set $H = \{h_1, h_2, \cdots h_n\}$, where n is some finite number, and h_i are mutually different.Using this set we can generate our Hoeffding's inequality:

$g \in H = \{h_1, h_2, \cdots h_n\}$

$$Pr\left(|E_{in}(g) - E_{out}(g)| > \epsilon\right) \leq 2ne^{-2\epsilon^2 N}$$

[1] part of the content comes from Jiaming's slides Brief introduction of Jiaming slides resources

Now, if we add another function h_{n+1} into the set, and let $h_{n+1} = h_n$, so we have a new set H'. And now there is a question, for $g \in H = \{h_1, h_2, \cdots h_n\}$, what is the new $Pr\left(|E_{in}(g) - E_{out}(g)| > \epsilon\right)$?

According the inequality:

$$Pr\left(|E_{in}(g) - E_{out}(g)| > \epsilon\right) \leq 2(n+1)e^{-2\epsilon^2 N}$$

But as we know, the newly added h_{n+1} is exactly the same as h_n, and H' in essence is no different from H, so the probability should also be the same.

So here is the key idea: the number of functions of the same kind under sample data does not matter, only disparate candidate functions counts. Based on this idea, we can have the definition of growth function.

3 Growth function

3.1 definition[2]

let H be a hypothesis-class (a set of binary functions) and C a set with m elements. The restriction of H to C is the set of binary functions on C that can be derived from H:

$$H_C = \{h_{(x_1)}, h_{(x_2)}, \cdots h_{(x_m)} | h \in H, x_i \in C\}$$

The growth function measures the size of H_C as a function of $|C|$

$$growth(H, m) = \underset{C:|C|=m}{max} |H \cap C| = \underset{C:|C|=m}{max} |H_c|$$

3.2 interpretation

As we can see, the growth function describes the unique inclusiveness of the hypothesis set. It can be considered as the maximum number of disparate situations, or to be more professional, dichotomies, that H can generate based on N observations. This way, all the useless, repetitive parts of H will be eliminated, leaving only the distinguished ones, which means that H can contain infinite number of candidate functions yet has a finite growth function outcome. And the pure increase in function's number does not always give us more information. And this is the key feature of growth function, it gives us the description of truly distinguished information of a hypothesis set, rather than blindly counting numbers. Extra functions are only rewarded in growth function when they are the first of a new kind.

[2] this definition is from Wikipedia wiki growth function

4 VC dimension

4.1 definition[3]

Let H be a set family(of h)and C be a set, define their intersection as:

$H \cap C = \{h \cap C | h \in H\}$

C is said to be shattered by H if $H \cap C$ contains all the subsets of C, ie:

$|H \cap C| = 2^{|C|}$

the VC dimension of H is the largest integer D that there exists a set C with cardinality D that is sgattered by H. And if arbitrarily large subsets can be shattered, then the VC dimension is infinity.

We can also draw the main idea of VC dimension from the definition of growth function above. We can see apparently that $m_H(N)$, ie $growth(H,m) \leq 2^N$, since one extra binary point can only produce up to two different kinds of outcomes. And when the equation holds, ie, $growth(H,m) = 2^N$, then H is said to shatter $x_1, x_2 \cdots x_N$, and the VC dimension of H is the largest N of all values that satisfies $growth(H,m) = 2^N$.

4.2 interpretation

VC dimension is typically a way to describe the complexity of a hypothesis set, the larger its VC dimension is, the more complex it is, and the more information it can convey. And it can also partly reflect the value of growth function. Generally with other things being equal, hypothesis set with higher VC dimension has larger growth function value.

5 Connection between GF & VC

5.1 in definition[4]

We can see apparently that $m_H(N)$, ie $growth(H,m) \leq 2^N$, since one extra binary point can only produce up to two different kinds of outcomes. And when the equation holds, ie, $growth(H,m) = 2^N$, then H is said to shatter $x_1, x_2 \cdots x_N$, and the VC dimension of H is the largest N of all values that satisfies $growth(H,m) = 2^N$.

5.2 Sauer–Shelah lemma[5]

If $VCdim(H) = d$, then,

for all m: $Growth(H,m) \leq \sum_{i=0}^{d} (\substack{m \\ i})$

in particular, for all $m > d+1$:$Growth(H,m) \leq \left(e\frac{m}{d}\right)^d$

[3] this definition is from Wikipedia wiki VC dimension

[4] part of the content comes from Jiaming's slides slides resources

[5] part of the content is from Wikipedia wiki growth function

so when the VC dimension is finite, the growth function grows polynomially with m, and this upper bound is tight, ie:

for all $m > d$, there exists H with VC dimension d such that : $Growth(H, m) = \left(e\frac{m}{d}\right)^d$

6 Applications in learning

6.1 growth function

As mentioned before, Hoeffding's inequality loses its instructive function when Hypothesis set contains infinite functions, even though not all functions are useful. So in learning we replace $|H|$ with its growth function value$m_H\left(N\right)$. We use $m_H\left(N\right)$ to represent the truly essential, distinguished number of functions in the whole hypothesis set so that Hoeffding's inequality continues to provide constructive information in learning.

6.2 VC dimension[6]

The VC dimension can predict a probabilistic upper bound on the test error of a classification model:

$Pr(|E_{in}(g) - E_{out}(g)| \leq \sqrt{\frac{1}{N}[D(\log(\frac{2N}{D}) + 1) - \log(\frac{\varrho}{1})]}) = 1 - \varrho$

where D is the VC dimension, $0 \leq \varrho \leq 1$, and N is the sample size(only valid when $D < N$, or overfitting can cause problems otherwise)

6.3 analysis

From the contents above we can know that a hypothesis set H can reduce the difference between in-sample error and out-of-sample error by inducing more candidate functions, ie increasing its complexity, but at the same time losen the restriction bound of the Hoeffding's inequality, due to increased $m_H(N)$. And VC dimension indicates that there exists an optimal balance of the trade-off between complexity and error difference. So in practice we will need to take both these two offsetting forces into consideration.

[6] part of the content is from Wikipedia wiki VC dimension