A Review of Statistical Learning Theory

Wei HUANG, Gao-ming LI and Wen-wu CHEN
Armed Police Engineering University, Xi’an, Shaanxi, China.
*Corresponding author

Keywords: Machine learning, Artificial intelligence, Experience risk, Consistency, Generalization bound.

Abstract. Statistical learning theory is developed on the basis of machine learning, and it provides the theoretical basis under the condition of small samples of pattern recognition, function fitting and the density estimation. In recent years, the research on statistical learning theory has made more and more important contribution to the research on artificial intelligence. In view of its importance, this paper has sorted out the basic problems of statistical learning theory: (1) The conditions of statistical learning consistency under the rule of minimum empirical risk minimization; (2) The conclusions about the generalization of statistical learning methods in these conditions; (3) Small sample induction and inference criteria established on the basis of these boundaries; (4) At last, the current research status and the further development of statistical learning theory are discussed.

Introduction

Many things in the objective world can only be understood indirectly through observation, and statistics can provide an effective analytical method for solving such problems. However, the traditional statistical research is based on the asymptotic theory in the case of large samples. But there is far less demand for the infinite samples in reality. Though people know this, they still hope that all kinds of algorithms under the premise that the assumption of the infinite number of samples also has good performance in the small sample, Obviously, in many application of this reasoning will perform very poor. In order to solve this problem, people have made many improvements, but most of them focused on the correction of traditional statistical principle and it difficult to break through the old idea of framework.

Statistical learning theory is a theory which is specific to small samples. It provides a strong theoretical foundation for the research of three problems in machine learning under the limited sample. And the support vector machine (SVM) method developed on the basis of showed the excellent characteristics.

Statistical Learning Consistency Condition under Empirical Risk Minimization Criteria

The Conditions for the Consistency of Empirical Risk Minimization Principles

For the function set Q(z,a) and the probability distribution function F(z), if you have any non-empty subsets Λ(c), c ∈ (−∞, ∞) of this set of functions which made the convergent established:

\[ \inf_{a \in \Lambda(c)} R_{\text{emp}}(a) \rightarrow_{\text{as} \to \infty} \inf_{a \in \Lambda(c)} R(a) \]  

And then we call the empirical risk minimization method is strictly (non-trivial) consistent.

The Necessary and Sufficient Conditions for the Strict Consistency of Minimize Experience Risk

The consistency analysis of the empirical risk minimization method is essentially associated with the convergence analysis of the two empirical processes. The bilateral empirical process refers to the sequence of random variables dependent on probability measure F(z) and function set Q(z,a):
The formula of the concept of unilateral empirical process is given with the bilateral experience process:

\[ \xi^i = \sup_{a \in \Lambda} \left| \sum_{i=1}^{n} Q(z, a) \right| \]  

The problem of the empirical process is to describe a set of conditions, which make the downward convergence established for any positive \( \varepsilon \). On this basis, the expression of uniform convergence and uniform unilateral convergence is given as follows:

\[ P[\sup_{a \in \Lambda} \left| \sum_{i=1}^{n} Q(z, a) \right| > 0] \xrightarrow{r \to \infty} 0 \]  

\[ P[\sup_{a \in \Lambda} \left| \sum_{i=1}^{n} Q(z, a) \right| > 0] \xrightarrow{r \to \infty} 0 \]  

It is interesting to note that what we have been looking for the necessary and sufficient condition of strictly consistency of empirical risk minimization is unilateral uniform convergence established in a given set of functions, which is the famous Key Theorem. It should be noted that the necessary and sufficient condition of strict consistency of empirical risk minimization method of learning process is given by consistent unilateral convergence rather than bilateral convergence, which reason is that we are faced with a situation that is asymmetric: what we are looking for is the results consistency of minimization of the empirical risk, but we do not consider the results consistency of maximization of the empirical risk.

The Conclusion of the Generalization Bound of Statistical Learning Method

Three Milestones of Statistical Learning Theory

On the basis of the previous section, we summarized three milestones of statistical learning theory. Consider two new concepts based on \( N^\lambda(z_1, ..., z_i) \) values:

1. the VC entropy of annealing:

\[ H^\lambda_{\text{ann}} = \ln N^\lambda(z_1, ..., z_i) \]  

2. Growth function:

\[ G^\lambda(l) = \ln \sup_{z_1, ..., z_l} N^\lambda(z_1, ..., z_l) \]  

The major milestones of learning theory are established based on these functions. The first milestone:

\[ \lim_{l \to \infty} \frac{H^\lambda_{\text{ann}}(l)}{l} = 0 \]  

It describes sufficient conditions for the consistency of empirical risk minimization principles. Consider the question raised in the previous section on this basis: under what conditions is the convergence rate the fastest? We concluded that the equation:

\[ \lim_{l \to \infty} \frac{H^\lambda_{\text{ann}}(l)}{l} = 0 \]  

This equation is the second milestone in learning theory, which guarantees the fast convergence speed. The above two milestones are effective for a given probability measure, but the goal of learning theory is under certain conditions, which is not dependent on the probability measure, ERM principle
is consistent and has the fast convergence speed at the same time. So we have the third milestone of learning theory:

\[
\lim_{l \to \infty} \frac{G^l(l)}{l} = 0 \tag{3.5}
\]

which gives the sufficient and necessary conditions for the consistency of the ERM principle, and the convergence rate is the fastest when this condition is established. These three milestones constitute the basis for establishing the convergence speed of learning machines, which includes the distribution independent boundary and the strict dependence on the distributed boundary.

**The Boundary of Convergence Velocity Structure**

In this case, we only give the structural boundary which has nothing to do with distribution based on the VC dimension of the function set, and the VC dimension of the function set can be used to find the structural boundary, because the following inequality is established:

\[
\frac{H^l(l)}{l} \leq \frac{G^l(l)}{l} \leq \frac{h(\ln \frac{l}{h} + 1)}{l}, (l > h) \tag{3.6}
\]

Therefore, the limited VC dimension is a sufficient condition for the consistency of the ERM method, and a limited VC dimension means the fast convergence speed. In the limited case of VC dimension, the following structural boundary is established:

A. The following inequality is true for all unbounded nonnegative functions:

\[
R(a) \leq \frac{R_{emp}(a)}{(1 - a(p)\tau \sqrt{|\phi|})}. \tag{3.7}
\]

B. For the function \(Q(z,a)\) that minimizes the risk of experience, the following inequality is true at least \(1 - 2\eta\) :

\[
\frac{R(a) - \inf_{a \in \Lambda} R(a)}{\inf_{a \in \Lambda} R(a)} \leq \frac{\tau a(p)\sqrt{|\phi|}}{(1 - a(p)\tau \sqrt{|\phi|})} + O(\frac{1}{l}) \tag{3.8}
\]

**Small Sample Inductive Inference Criterion Is Established Based on These Boundaries**

Using the previous relevant conclusion to discuss promotion ability that control learning machine. It mainly in order to construct a inductive principle to minimize risk functional by using of small sample training instances, which names The Structural Risk Minimization Inductive Principle. First of all, consider the boundary of the learning machine generalization ability of the complete bounded non-negative function set.

\[
R(a) \leq R_{emp}(a) + \frac{B\phi}{2} \left[1 + \sqrt{1 + \frac{4R_{emp}(a)}{B\phi}}\right] \tag{4.1}
\]

and the field of learning machine generalization ability of unbounded function sets:

\[
R(a) \leq \frac{R_{emp}(a)}{(1 - a(p)\tau \sqrt{|\phi|})}, \tag{4.2}
\]

where, if the function set \(Q(z,a)\) contains an infinite number of elements and the VC dimension \(h\) is finite, then there is:

\[
\phi = 4 \frac{h(\ln \frac{2l}{h} + 1) - \ln \left(\frac{2}{4} \right)}{l} \tag{4.3}
\]

And if the function set contains a finite number of \(N\) elements, then:
\[
\phi = 2 \ln \frac{N - \ln \eta}{l}
\] (4.4)

Structural risk minimization principle: for a given set of observations \(Z_1, ..., Z_l\), the principle of structural risk minimization is to choose a function that minimizes the risk of experience in a subset of the minimum risk upper bound. The structural risk minimization principle actually defines a tradeoff between the precision of a given data approximation and the complexity of the approximation function.

**Research Status and Prospect**

Although statistical learning theory is currently regarded as the best theory to solve the problem of small sample machine learning, the theory also shows its own limitations with the deepening of the research application:

1. The statistical learning theory is based on a probability space, and the probability measure is quite harsh conditions of additivity, in practice, the conditions are often not satisfied, so how the statistical learning theory to a large number of existing in the actual non-additive on the set of functions is a difficult problem to be solved.

2. The statistical learning theory is based on the real random variables, in real life application there are a lot of the real random variables, so the real random variables based on statistical learning theory is difficult to solve the problem in real life the real random variables.

In order to make the statistical learning theory more widely used, a large number of scholars have conducted in-depth studies on it. The research status is as follows: Ha Minghu et al. extended the statistical learning theory from the probability measurement space to the Sugeno measurement space, and proposed the principle of minimizing the non-trivial uniform convergence of empirical risk in this space, and also gave the key theorems of learning theories in this space and the bounds of uniform convergence speed in the learning process; Bai et al. obtained the sub-key theorem and key theorem of the learning theory in the credibility measurement space and the bound of uniform convergence speed; Zhang Xiankun et al. gave the key theorem of learning theory in uncertain space; Tian et al. gave the key theorems of learning theory based on fuzzy samples and the bounds of uniform convergence rate of learning process; Liu et al. presented the key theorems of learning theories based on rough samples and the bounds of uniform convergence rates of the learning process; Zhang Qiming et al. proposed the key theorems based on the learning theory of complex random samples, the bounds of the uniform convergence rate of the learning process, the VC dimension of the function set, and the principle of minimizing the structural risk; Wang Chao et al. presented the key theorems of fuzzy sample-based learning theories in Sugeno measure space and the bounds of uniform convergence rates in the learning process; Li Kunlun and others constructed a fuzzy multi-class support vector machine and used in intrusion detection; Yang Zhimin et al. proposed an uncertain support vector machine.

Finally, I think that the current work on statistical learning theory needs to be further: (1) To further improve the key theorems of non-real random samples in the generalized uncertainty measure space (probability measure and non-probability measure space), the bounds of the uniform convergence rate of the learning process, and the principle of minimizing structural risk; (2) Fusion of uncertain support vector machines based on theoretical constructs such as fuzzy, rough, unascertained theory and likelihood measures, and uncertainty support vector machines constructed based on the principle of uncertainty structural risk minimization in correspondingly uncertain statistics learning theory, to construct a generalized theory of uncertain support vector machines; (3) Selection and construction of extended kernel functions on uncertain support vector machines.

**References**


[7] Li Jia. Key theorem of statistical learning theory in the measurement space of Sugeno measurement [D]. Hebei University, 2005.
