Research on Overfitting Problem and Correction in Machine Learning

Chao Bu1,2, Zengping Zhang1

1School of Computer & Information Management, Inner Mongolia University of Finance and Economics, Hohhot 010070, China
2State Grid East Inner Mongolia Information & Telecommunication Company, Hohhot 010010, China
buchao_idea@163.com

Abstract: Machine learning is the key technology of artificial intelligence, which uses learning and training data model to find the problem to achieve the law. In practical applications, there is always a difference between the input data of the model and the training data. Based on the biased training data, overfitting will occur, which will cause the machine learning to fail to achieve the expected goals. Generalization is the process of ensuring that the model can fully reflect the characteristics of the actual input data. Therefore, the ability of the machine learning model depends to a large extent on the effectiveness of generalization. In order to fully understand the learning and generalization model, we conducted a study by the polynomial curve fitting. First, based on the training data, we analyzed using matrix theory and analytical solutions derived polynomial fit. Then, numerical and maximum likelihood theoretical analysis was carried out for the overfitting problem. Finally, combined with the objective function, a typical regularization method to overcome overfitting and improve generalization ability is elaborated.

1. Introduction

Machine learning is a multidisciplinary computer science technology, which belongs to a branch of artificial intelligence research and application. With the development of information technology and microelectronics technology, and more and more powerful of computing power, massive empirical data are used in machine learning. Machine learning is different from ordinary computer programs. It usually solves problems that cannot be done directly using fixed rules or process codes similar to those faced by us, such as the recognition of handwritten numbers, and is easy to recognize from the picture for humans. Machine learning is a modeling technique for the empirical data, to discover knowledge from data and using simple models, where data refers to a variety of physical records that contain information, just as pictures, video, audio, and models are results of machine learning. Meanwhile, in the process of machine learning, we use computer algorithms for sampling data collection, and analysis data independently to training an optimal model, and this model training process is called learning. In short, machine learning achieves a good performance of the model through self-learning ability using existing empirical data in the process of solving task.

As the core technology of artificial intelligence, machine learning has experienced several iconic in history. In 1959, Arthur Samuel of IBM in the United States developed a machine program that could play against human chess players and constantly improve themselves, defeating professional chess players. In 1997, IBM's Deep Blue supercomputer defeated the former Soviet chess master Kasparov.
in an American chess game. The Watson DeepQA system also developed by IBM in 2011, defeated some outstanding human players in the well-known encyclopedia knowledge quiz TV program in the United States and successfully won the championship in one fell swoop, which enabled us to make progress towards achieving the Turing Test. Recently, Google's DeepMind research team used the machine learning AlphaGo developed by deep learning to overwhelmingly defeated the ninth-dan Li Shishi of Go, and pushed artificial intelligence to a new level.

Machine learning to solve problems are often of a smart question, it is a learning model from the data, and then to solve practical problems. The data used in the modeling process is called training data. Due to the existence of noise, there are always differences between training data and actual input data. Therefore, models learned based on training data often cannot obtain satisfactory results in practical applications. This is a structural challenge faced by machine learning. The process of this problem is called generalization, and generalization mainly solves the problem of overfitting. In the learning process of machine learning, if all the conditions in the training data are matched, including noise data, a model with low generalization ability will be obtained, resulting in overfitting. An important method to suppress over-fitting is regularization, which is a numerical method for constructing minimal models. By adding a penalty to the objective function, a simple model that can better reflect the overall characteristics of the training data is obtained, thereby over-fitting can be overcome.

According to different training methods, machine learning is roughly divided into three categories: supervised learning, unsupervised learning, and reinforcement learning. The training data of supervised learning is composed of attribute vectors and standard output labels. Its learning process is to reduce the error between model output and standard output by adjusting model parameters. Different from supervised learning, the training data of unsupervised learning only contains the input attribute vector, which is the natural clustering law of discovery data. In the training data set of enhanced learning, some data have labels, and some data only have input. It uses the scoring mechanism to train and learn the model.

This paper takes the regression analysis of supervised learning as the object, researches the mechanism of overfitting in machine learning modeling through polynomial curve fitting, and uses regularization to solve the problem of overfitting.

2. Model establishment

Supervised learning for the training data set marked, based on the objective function by computer algorithms learning model, and then enter the actual data and outputs the correct classification or prediction values. Figure 1 is a block diagram of supervised learning.

![Fig.1 Block diagram of supervised learning principle](image_url)

Figure 1 shows that if the predicted value of the learned model is a continuous value, this supervised learning is a regression problem. Next, through the linear regression problem, first analyze the labeled training data, and then explain the polynomial curve fitting model to complete the modeling in machine learning.
2.1. Labeled training data set

The problem to be solved by regression prediction is to input the observed value and predict the target value. In order to facilitate the comparison of the accuracy between the predicted value and the target value, as well as the quality of the model, we use the known function and Gaussian white noise to generate the training data set. The generation process is shown in Figure 2.

Assume that the function is a sine function, e.g. \( g(x) = \sin 2\pi x \). Taking into account measurement error is often due to the introduction of error of measurement system, the noise is modeled as additive white Gaussian noise, zero mean, unit variance. Then the probability distribution is \( \phi(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2} \).

If there are \( N \) sampled observation values and the corresponding target values, that is, \( (x_1, x_2, \ldots, x_N) \), \( (y_1, y_2, \ldots, y_N) \). Now let the sampling value \( x \in [0,1] \), sampling at equal intervals, and sampling interval \( \Delta = 0.1 \), the training data set \( D \) shown in Table 1 is obtained, and the size is \( |D| = 11 \).

| N | X   | Y     |
|---|-----|-------|
| 0 | 0   | -0.270|
| 1 | 0.1 | -0.177|
| 2 | 0.2 | 0.967 |
| 3 | 0.3 | 0.755 |
| 4 | 0.4 | 1.276 |
| 5 | 0.5 | 0.412 |
| 6 | 0.6 | 0.300 |
| 7 | 0.7 | -0.808|
| 8 | 0.8 | -0.689|
| 9 | 0.9 | -0.989|
| 10| 1.0 | -0.571|

The images of the training data set \( D \) and function \( g(x) \) are shown in Figure 3. Among \( g(x) \) is the law contained in the data set \( D \), which is the goal of our mining knowledge and exploration. Scattered points are sample points sampled. Due to the interference of system errors and measurement errors, there is always a certain random error between them and the real target value. The random error is introduced by Gaussian white noise simulation.
2.2. Modeling

According to the training data set to explore the laws contained in the data, here we use a polynomial function to fit the training data. The polynomial function is

\[ f(x) = w_0 + w_1x + w_2x^2 + \ldots + w_Mx^M \]

\[ = \sum_{j=0}^{M} w_jx^j \]  \hfill (1)

Where \( M \) is the order of the polynomial function, and the vector \( w = (w_0, w_1, \ldots, w_M)^T \) is the undetermined coefficient of the function.

In order to determine the undetermined coefficients, it is necessary to fit the training data with a polynomial function. That is to say, for each sample point, the error between the output value of the polynomial function and the measured value of the sample point is the smallest. Here we take the sum of squared errors (MSE) of all sample points as the objective function of learning is

\[ E(w) = \frac{1}{2} \sum_{n=1}^{N} \left[ f(x_n) - y_n \right]^2 \]  \hfill (2)

Here, the objective function \( E(w) \) is a function of the undetermined coefficient \( w \), obviously \( E(w) \geq 0 \). If and only if all of the sample points fitted \( f(x) \), the squared error is zero. By selecting different parameters \( w \) to minimize the objective function \( E(w) \), the polynomial curve fitting problem is transformed into the following optimal problem for solving the objective function

\[ w = \text{arg min } E(w) \]  \hfill (3)

Next, we use the matrix method to solve equation (3). For any sample \( \forall x \in \{x_0, x_1, \ldots, x_N\} \), there are

\[ f(x) - y = \sum_{j=0}^{M} W_jx^j - y = \varepsilon \]  \hfill (4)

Where \( \varepsilon \) is a fitting error, the linear equation is

\[ \sum_{j=0}^{M} W_jx^j - y_i = \varepsilon_i, \quad (i = 0, 1, \ldots, 10) \]  \hfill (5)

Let

\[ A = \begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^M \\ 1 & x_1 & x_1^2 & \ldots & x_1^M \\ & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \ldots & x_N^M \end{pmatrix} \]  \hfill (6)
\[ w = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{pmatrix}, \quad y = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_M \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_M \end{pmatrix} \]  

Then the variable matrix of formula (5) is

\[ Aw - y = \epsilon \]  

According to formula (2), (3), (8), we know

\[ \epsilon = \arg \min \frac{1}{2} \| H \epsilon \|_2^2, w = \arg \min \frac{1}{2} \| H \epsilon \|_2^2. \]

Where the square of the \( \epsilon \) 2-norm is

\[ \| \epsilon \|_2^2 = \epsilon^T \epsilon = (Aw - y)^T(Aw - y) \]

Let the gradient of equation (9) with respect to \( w \) be zero, that is \( \nabla \| \epsilon \|_2^2 = 0 \), so the optimal solution of \( w \) is

\[ A^T A w + w^T A^T \epsilon - 2 A^T y = 0 \]

\[ 2 A^T A w - 2 A^T y = 0 \]

\[ w = (A^T A)^{-1} A^T y \]

3. Polynomial model learning

For the above training data set \( A \), we take the order \( B \) of the polynomial model function. According to formula (10), the training of the model is performed respectively. The training results are shown in Table 2.

| \( M \) | \( w_0 \) | \( w_1 \) | \( w_2 \) | \( w_3 \) | \( w_4 \) | \( w_5 \) | \( w_6 \) | \( w_7 \) | \( w_8 \) | \( w_9 \) | \( w_{10} \) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | -0.056 | -1.256 | 15.912 | 1280460.13 | | | | | | | |
| 1 | 0.572 | -27.325 | -6408442.22 | | | | | | | | |
| 3 | 1.1302 | 13746035.75 | | | | | | | | | |
| 10 | -0.519 | -16512332.70 | | | | | | | | | |

3.1. Model selection

According to the model learning results (Table 2), a polynomial curve fitting graph is drawn as shown in Figure 4.

Figure 4 (a) and (b) show that the polynomial curve cannot fit the training sample data points well. This underfitting phenomenon is caused by the small order of the polynomial.
Figure (c) shows that when the order \( A = 3 \). The polynomial curve fits the sample points well. Figure (d) shows that the polynomial curve has passed through each sample data point and perfectly tuned the training data. However, the polynomial curve has severe oscillations. This phenomenon is called overfitting.

In order to test the generalization performance of the model, we often use the data set to test the performance of the model. Overfitting can well tune the training set. Due to the difference between the training set and the test set, the performance of the model on the test set is poor and cannot be used in practical applications, which means that the generalization ability of the model is poor.

Here we introduce the root mean square error to specifically evaluate the generalization performance of the model, and its mathematical expression is

\[
E_{\text{RMS}} = \sqrt{\frac{1}{N} \sum (w_{\text{ideal}} - w_{\text{model}})^2}
\]

Where, \( N \) is the number of data points, to compare data sets of different sizes, we use \( N \) divided by the residuals of the model.

Use the same method to generate the test set. Then, take the order \( M \in [0,10] \) of the model, the step size is 1, and draw the root mean square error curves of the training set and the test set, as shown in Figure 5.

Figure 5 shows that the root mean square error curve has an obvious inflection point at \( M = 3 \). Before \( M = 3 \), as the order increases, the root mean square error decreases linearly. When \( 3 \leq M \leq 9 \), root mean square error remained basically stable. At \( M = 10 \), for the training data set, the error is zero.
due to over-fitting. At this moment, for the test set, the error increases in vain, which indicates that the generalization performance of the model at \( M = 10 \) is poor.

3.2. **Analysis of overfitting mechanism**

The predicted value of the model is uncertain, so we can use a random variable to represent it, which obeys the Gaussian distribution, and the probability density is

\[
P(y) = \frac{1}{(2\pi \sigma^2)^{\frac{1}{2}}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}
\]

(12)

Where, \( \mu \) is the mean value of the Gaussian distribution, which is equal to the output value \( f(x) \) of the fitted model. \( \sigma^2 \) is the variance of the Gaussian distribution.

Given the training set \( D \), the likelihood function is

\[
L(w, \sigma^2) = \prod_{n=0}^{N} \frac{1}{(2\pi \sigma^2)^{\frac{1}{2}}} e^{-\frac{(y-f(x_n))^2}{2\sigma^2}}
\]

(13)

Equation (13) takes natural logarithms on both sides, then

\[
\ln L(w, \sigma^2) = -\frac{1}{\sigma^2} \sum_{n=0}^{N} (y-f(x_n))^2 - \frac{N+1}{2} \ln(2\pi \sigma^2)
\]

(14)

Find the maximum value of equation (14) with respect to \( w \). Since the second term on the right side of equation (14) does not depend on the parameter \( w \), this problem is transformed into finding the minimum value of \( \sum_{n=0}^{N} (y-f(x_n))^2 \). This result is consistent with the above objective function. This means that minimizing the objective error function is essentially to obtain the parameter estimates \( \mu_{\text{ML}} \) and \( \sigma^2_{\text{ML}} \) of the maximum likelihood function, and because the maximum likelihood estimates of the mean and variance of the Gaussian distribution are respectively,

\[
\begin{align*}
\mu &= \frac{1}{N+1} \sum_{n=0}^{N} y_n \\
\sigma^2 &= \frac{1}{N+1} \sum_{n=0}^{N} (y_n - \mu)^2
\end{align*}
\]

(15)

Their respective mathematical expectations are

\[
\begin{align*}
E[\mu] &= \mu \\
E[\sigma^2] &= \frac{N}{N+1} \sigma^2
\end{align*}
\]

(16)

Equation (16) shows that the mean \( \mu \) is unbiased. In other words, the model \( f(x) \) of maximum likelihood estimation is unbiased. The variance \( \sigma^2 \) is biased and reduced, which will lead to overfitting.

4. **Regularization of the model**

From the analysis of the over-fitting mechanism, it can be seen that over-fitting is a general feature of maximum likelihood function estimation. Only in the case of big data, the problem of over-fitting can be alleviated.

If the training data set is given, the size \( |D| \) is certain. In order to obtain a certain degree of complexity of the model, while avoiding overfitting, the method we often use is regularization.
The regularization method is to add a penalty term to the objective function to suppress the excessive increase of model parameters. In the polynomial curve fitting process, after the error square and the objective function, add an inner product of the polynomial model coefficient vector, which is
\[
\|w\|^2 = w_0^2 + w_1^2 + \ldots + w_M^2
\] (17)

Then the error objective function becomes
\[
E(w^*) = \frac{1}{2}\sum_{n=1}^{N} [f(x) - y_n]^2 + \frac{\lambda}{2}\|w\|^2
\] (18)

Among them, \(\lambda\) is the introduced inhibitory factor. The larger the \(\lambda\), the stronger the effect of the penalty term.

Using the matrix method, the optimal solution for solving equation (18) is
\[
w^* = (A^TA + \lambda I)^{-1} A^Ty
\] (19)

In the above formula, \(I\) is the identity matrix.

In view of the over-fitting polynomial order \(M = 10\), we then use equation (19) for model learning, and get the fitting curve as shown in Figure 6. Figure 6 shows that the original violent shock has been suppressed.

In order to further illustrate the effect of regularization in suppressing over-fitting, for the training set and test set, we draw the relationship curve between the root mean square error and the suppression factor, as shown in Figure 7.

Figure 7 shows that as the suppression factor decreases, the error of the training set tends to zero, while the error of the test set increases, which indicates the occurrence of overfitting. Conversely, with the increase of the suppression factor, the role of the regularization term increases, and the error of the test set decreases and is controlled, thereby suppressing the occurrence of overfitting.
5. Conclusion
This article elaborates on the overfitting problem and correction methods in machine learning. First, based on the analysis of the training data set, a modeling method is proposed using the objective function. Then we use the training data set to learn the model. Combining the polynomial curve fitting model and applying the probability analysis method, it is concluded that the model learning mechanism is the maximum likelihood estimation of the predicted value probability. Then the overfitting problem of the learning model is analyzed in detail, and the essential principle of overfitting is explained at the same time. It is caused by the bias of the maximum likelihood estimation variance. Finally, for over-fitting, one of the commonly used methods of regularization is analyzed, and the comparison of the root mean square error between the test set and the training set shows that regularization can effectively suppress over-fitting. In this paper, we use polynomial curve fitting to solve the problem of regression in machine learning, and help researchers solve the problem of over-fitting the general rules in machine learning, to deepen understanding of the regularization method to solve the over-fitting problems and improve the analysis of the problem and problem solving ability in machine learning.

Acknowledgments
This work was supported by the Talent Development Fundation of Inner Mongolia Autonomous Region and Natural Science Foundation of Inner Mongolia Autonomous Region (Grant No.2020MS06021).

References
[1] Breiman, L.(1996).Bias, variance, and arcing classifiers. Technical Report 460, Department, University of Californin, Berkeley, CA.
[2] Domingos, P.(2000).A unified bias-variance decomposition. In Proceedings of the 17th International Conference on Machine Learning (ICML), 231-238, Stanford, CA.
[3] Efron, B. and R. Tibshirani (1993). An Introduction to the Bootstrap. Chapman & Hall, New York, NY
[4] Friedman, J. H.(1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. Data Mining and Knowledge Discovery, 1(1): 55-77
[5] Geman, S, E. Bienenstock, and R. Doursat (1992). Neural networks and the bias/variance dilemma. Neural Computation, 4(1): 1-58
[6] Kohavi, R. and D. H. Wolpert.(1996). Bias plus variance decomposition for zero-one loss functions. In Proceeding of the 13th International Conference on Machine Learning(ICML), 275-283, Bari, Italy.
[7] Kong, E. B. and T. G. Dietterich (1995). Error-correcting output coding corrects bias and variance.In Proceedings of the 12th International Conference on Machine Learning (ICML), 313-321, Tahoe City, CA.
[8] Mitchell, T(1997) Machine Learning. McGraw Hill, New York, NY
[9] Wellek, S. (2010). Testing Statistical Hypotheses of Equivalence and Noninferiority, 2nd edition. Chapman & Hall/CRC, Boca Raton, FL.