The effect of regularization coefficient on polynomial regression

Xing Wan
Leshan Vocational and Technical College, Leshan 614000, China
krantson@163.com

Abstract. Polynomial regression can be used for nonlinear models. When solving polynomial regression problems, polynomial regression can be transformed into linear regression to solve. In order to avoid over-fitting in polynomial regression, a regularization method can be used to suppress the coefficients of higher-order polynomial, and the article evaluates the influence of regularization coefficients on polynomial regression.

1. Introduction
Polynomial regression\[^1\] can be used to fit nonlinear models. Many of the models in the actual problem are inappropriate to linear models, and if a linear model is used, the fitting effect is poor. Although the polynomial model has a good training error for the training data, the disadvantage is that it is easy to produce over-fitting. Especially as the model order increases, the curve will fit the training data better. However, for test data other than training data, high-order models often have no good fitting effect on low-order models. Therefore, it is necessary to suppress the coefficient of the high-order term by the regularization method. Through the regular coefficient, the influencing factors of the high-order term can be reduced, and the coefficient of the high-order term is reduced to close to 0, so that a certain compromise can be achieved in fitting the training data and maintaining the good generalization ability. Polynomial regression usually has two solutions: the gradient descent method and the normal equation. Therefore, the influence of the regular coefficients on the polynomial regression in the gradient descent method and the normal equation is discussed respectively.

2. Polynomial regression model
For a polynomial regression model, the model can be divided into linear model, quadratic curve model, cubic curve model, etc. according to the number of highest items. The linear model is as follows:

\[ h(x; \theta) = w_0 + w_1x \]  
(2.1)

If using a quadratic model:

\[ h(x; \theta) = w_0 + w_1x + w_2x^2 \]  
(2.2)

Or use a quadratic curve model

\[ h(x; \theta) = w_0 + w_1x + w_2x^2 + w_3x^3 + w_4x^4 \]  
(2.3)

The following figure shows the polynomial regression algorithm to solve the men's 100m freestyle problem in the Olympic Games. The quadratic model, the quadruple model and the eighth model are shown in the following figure:
It can be seen from Fig. 1 that the training data loss decreases as the model order increases, but the test data loss increases as the model order increases. The results show that the more complex the model, the higher the degree of overfitting. If you want to avoid overfitting, collecting more data or using regularization methods.

3. The effect of regularization coefficient on polynomial regression
Assuming that the regression model uses four models, the model can be expressed as:

\[ h(x; \theta) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 \]  \hspace{1cm} (3.1)

Assume \( \theta = (w_0, w_1, w_2, w_3, w_4) \), in the polynomial, the order of polynomial than 2 is generally referred to as a high-order polynomial. It is because the high-order polynomial leads to over-fitting, so as to reduce the influence of the high-order term, the coefficient of the high-order term is reduced to close to 0, so that it can be compromised in both the training data and the generalization ability. The basic method of regularization is to reduce the value of parameters, especially the high order factor. Incorporating these coefficients into the cost function and minimizing the cost function suppresses the high order term coefficients. The cost function after introducing regularization is as follows:

\[ J(\theta) = \frac{1}{2} \left[ \frac{1}{N} \sum_{i=1}^{N} (h(x^{(i)}; \theta) - y^{(i)})^2 + \lambda \sum_{j=1}^{5} w_j^2 \right] \]  \hspace{1cm} (3.2)

\( \lambda \) is called the regularization coefficient, and its value has a great influence on the model. If the gradient descent method is used, then the derivative is required and the cost function is minimized, so the gradient descent algorithm can be changed to:
do:

\[
\begin{align*}
\mathbf{w}_0 &= \mathbf{w}_0 - \alpha \frac{1}{N} \sum_{j=1}^{N} \left( (h(x^{(j)}; \theta) - y^{(j)}) x^{(j)} \right)
\end{align*}
\]

for \(i=1\) to \(D\) do:

\[
\begin{align*}
\mathbf{w}_0 &= \mathbf{w}_0 - \alpha \frac{1}{N} \sum_{j=1}^{N} \left( (h(x^{(j)}; \theta) - y^{(j)}) x^{(j)} - \lambda \mathbf{w}_i \right)
\end{align*}
\]

end for;

until convergence;

A polynomial regression algorithm was used to solve the Olympic men's 100-meter freestyle problem, and a regularized gradient descent algorithm was used here. It can be seen that the curve no longer pursues fitting the whole data, but pays more attention to the inherent law of data. Therefore, the training error is larger than the regularization method, and the generalization ability is enhanced. The obtained performance curve is as follows:

![Fig.2 Influence of regular coefficients on gradient descent](image)

It is also possible to use a matrix form when the normal equation is using a least squares method\(^{[2]}\). The gradient descent method is an algorithm that minimizes the cost function. The normal equation is to derive the parameter and make it equal to 0, and directly solve the parameter set that minimizes \( J \).

\[
J(\theta) = \frac{1}{2} \left[ \frac{1}{N} \sum_{i=1}^{N} (h(x^{(i)}; \theta) - y^{(i)})^2 + \lambda \sum_{j=1}^{D} w_j^2 \right] = \frac{1}{2N} (X\theta - y)^T (X\theta - y) + \frac{1}{2} \lambda \theta^T A \theta \tag{3.4}
\]

Let the cost function gradient be 0:

\[
\frac{\delta J(\theta)}{\delta \theta} = \frac{1}{N} X^T X \theta - \frac{1}{N} X^T y + \lambda A \theta = 0 \tag{3.5}
\]

Parameters is find:

\[
\theta = \left( X^T X + N \lambda A \right)^{-1} X^T y \tag{3.6}
\]

Assume \( \lambda = N \lambda \), then the formula (3.6) can be rewritten as

\[
\theta = \left( X^T X + \lambda A \right)^{-1} X^T y \tag{3.7}
\]

Using the normal equation method, the performance curves obtained are as follows:
4. Conclusion
The regularization coefficient directly affects the training error and test error of polynomial regression. The conclusions are as follows:
(1) When regularization coefficient is larger, the parameters are closer to 0, resulting in underfitting.
(2) If the regularization coefficient is smaller, there is almost no penalty for the parameter, which tends to lead to overfitting.
(3) The regularization coefficient should compromise between training error and generalization performance.

Acknowledgments
This work was financially supported by the Applied Research Project of Education Department of Sichuan Province (18ZB0280).

References
[1] Christopher M. Bishop. Pattern Recognition And Machine Learning[M]. Springer Science+Business Media, LLC, 233 Spring Street, New York., NY 10013, USA, 2006
[2] Tom M. Mitchell. Machine Learning[M]. McGraw-Hill Education – Europe, London. 1997.
[3] Cramer, K. and Y. Singer.. On the algorithmic implementation of multiclass kernel-based vector machines.[J]. Journal of Machine Learning Research, 2001, 2:265-292.
[4] Ian Goodfellow, Yoshua Bengio, Aaron Courville. Deep Learning[M]. MIT Press, London, England. 2017.
[5] Daphne Koller, Nir Friedman. Probabilistic Graphical Models Principles and Techniques[M]. The MIT Press, Cambridge, Massachusetts, London, England, 2009.
[6] A. Peled and A. Ruiz. Frequency domain data transmission using reduced computational complexity algorithms[C]. Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP-80, 1980, vol. 5: 964-967