Regularization in Data Mining

Tutorial by Liang Ma

Abstract:

When the training data is not enough, or over training, it often leads to overfitting. At this point, additional information is introduced to the original model to prevent overfitting and improve model generalization performance. The higher order term of the function is adjusted by adding additional penalty terms to the error function. The additional term controls the function of excessive fluctuations such that the coefficients do not use extreme values.

Over-fitting problems usually occur when there are too many variables (features). The equations trained in this case always fit the training data very well. However, such an equation does everything possible to fit the training data, which causes it to not be generalized into new data samples. Regularization is one of the solutions to this problem. We can keep all the feature variables but reduce the magnitude of the feature variables. This method is very effective. Each variable can have a little impact on the forecast, but it won't affect the model too much.

This tutorial will introduce the basic concept of regularization, overfitting and some other concepts we need to know in order to better understand regularization. This tutorial will then use Ridge Regression as an example of Regularization and Least squares as an example of Linear Regression to introduce how regularization deal with the overfitting. Followed by talking about the difference between Ridge Regression, Lasso Regression and Elastic-Net Regression. The tutorial will come with easy-to-understand examples and code sample section.