EPFL

Machine Learning Project 1 Report

K. Kappel & I. Leimgruber

October 30, 2017

Abstract

In this project, the goal was to implement a machine learning algorithm in order to obtain weighs able to discriminate particle events as precisely as possible to deduce whether these events were caused by a Higgs boson or not.

1 Least Squares

Using least squares on the training Data, we obtained a discrimination correctness rate of 74.4%

2 Ridge Regression

In this section we look over the ridge regression method

2.1 Least squares vs Ridge

With a good choice of lambda, penalizing heavy models reduces overfitting. First attempt gave around 76%

2.2 Ridge cross-validation

A basic grid-search with mean of losses over each fold for lambda, degree. We realised too late that Figure 1 was showing a high variance.

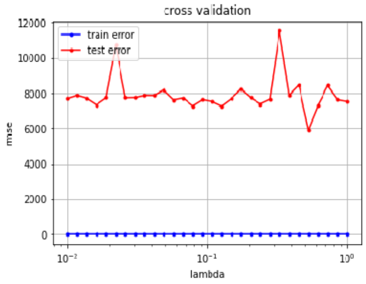

Figure 1: Plot of rmse(lambda) with fixed < degree

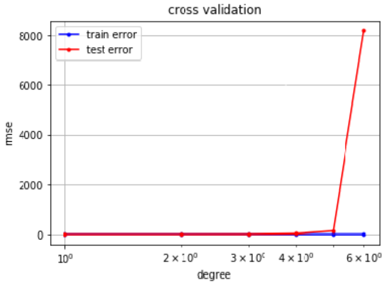

Figure 2: Plot of rmse(degree) with fixed lambda

3 Logistic Regression

Logistic regression is meant to be better to discriminate amongst a limited number of cases

3.1 Logistic cross-validation

We did a cross-validation for 3 hyper parameters: lambda, gamma and the degree of the polynomial model. We then plotted each hyperparameter with the rmse, each time taking the best values returned by the cross-validation for fixing the two other hyperparameters

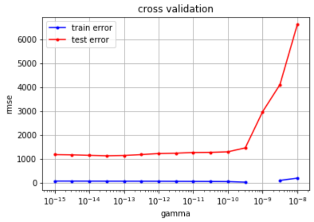

Figure 3: Plot of rmse(gamma) with best < degree and best lambda

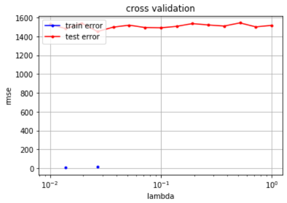

Figure 4: Plot of rmse(lambda) with best degree and best gamma

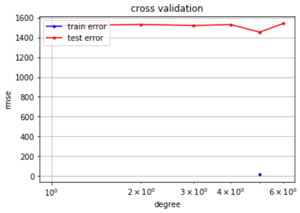

Figure 5: Plot of rmse(degree) with best lambda and gamma

We can see that the plots with the best hyperparameters didn't give good results as well as the Ridge Regression. We think that this is due to the negative loss on logistic and a high variance with the data.

3.2 Logistic vs Ridge

Our implementation does not confirm our hypothesis that Logistic regression would give better results than ridge, which is probably due to a mistake in our implementation of the logistic regression.

3.3 Ridge Improvement

As Ridge gave the best approximation, we thought about some improvements. First we saw that some values were -999 and that could penalise our final ratio. So we replace them (first by zero and then by the mean). It did not change our result drastically. We saw that one column called PRI-JET-NUM and the values of this column were 0,1,2,3. We realized that these values indicate different kinds of data and using this information could help us better our approximation. We considered each group individually and cross-validated them, which gave various results between 0.79 to 0.83. After testing our model, we understood that the weighted mean was lower than our ridge original result, probably caused by the sparsity or the easier-discriminated groups.

4 Conclusion

Our implementation of logistic regression does not work. Our ridge regression gives the best result so far. It seems that we could remove two features to better the approximation