Linear Regression Classification Project

CSE-6363-005-Machine Learning 2020 Spring

Project author: Jay Denton

Abstract

The topic of this paper is the classification of data in Machine Learning using linear regression. Although, the linear regression is not the best method in classification problems. It has a lot of disadvantages, for instance, predicted values are continuous instead of probabilistic. However, in this paper, it mainly focused on how to train the linear regression classification model to predict as good as possible, with avoidance of overfitting or underfitting.

Introduction

The dataset is given in a .data file with 150 substances. The data will be shuffled and separate to 105 training samples and 45 testing samples. In the training sample, the linear regression method will be used to generate parameters of the model. As addition, cross-validation will be applied to the training samples for tuning the parameters. The cross-validated models can be prepared by their overall cv-error. From the cross-validated model, the lowest cv-error set among all the others will be chosen, and tested in the testing sample to see its performance. All the procedures are coded using Matlab.

Problems

1. How to numerically substitute the class variables? Does it make a difference from 1,2,3 to 1,10,100 or even bigger gap between classes? And how to verify the double output numbers and classify them into a positive integer class.

2. With K-fold cross-validaion method, what is the difference between K=5 and K=1(which is leave-one-out)? How to use the cross-validation to tune the parameters that makes the model fits better?

Data

The first Problem needs to be solved before Data Processing procedure. The solution to Problem1 is that, the classes, Iris-setosa, Iris-versicolor, and Iris-virginica, has to be replaced by numbers 1,3 and 5. And according to the linear regression equation calculate out the parameter β. 1,2,3 or 1,10,100 for classes not really matters, because the final result is always normally distributed. Although, the predicted output, might not be specific 1,3 or 5. So my

solution is the output falls in the interval (negative-infinity, 2) will be considered as Iris-setosa, interval (2,4) will be considered as Iris-versicolor, and interval (4, positive-infinity) will be considered as Iris-virginica.

After replacing the classes with number, then, their indices will be shuffled, with the range from 1 to 10. Indices 1-7 will be the training set, and the indices 8-10 will be the testing set. However, in either dataset, the ratio between three classes needs to be 1:1:1.

Then, the training and testing data will not be changed again, and are used in different cross validation models.

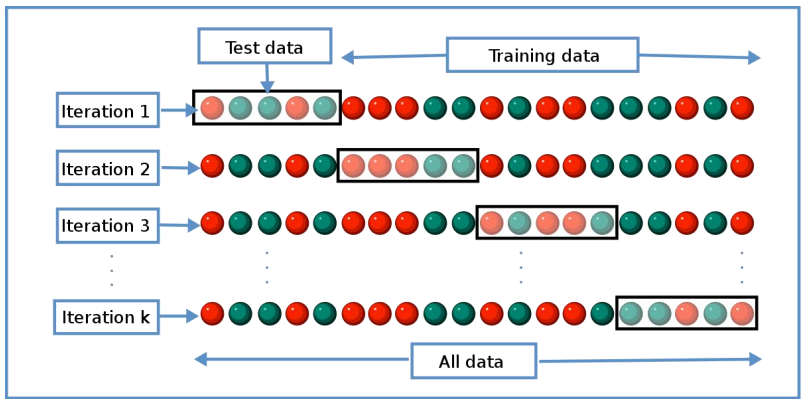
Method

The linear regression method is applied to the training with the equation below:

$$\hat{\beta} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y}$$

And the parameter vector β can be calculated easily.

To apply the cross validation on the training dataset, K-fold cross validation is used with different K values, 1 and 5 in this case. The process is shown below:



The cross validation error for each iteration is using the function below:

$$\mathrm{CV}_k(\hat{r}^{-(k)}) = \frac{1}{n_k}\sum_{i\in F_k}\left(y_i - \hat{r}^{-(k)}(x_i)\right)^2.$$

And the overall cross validation error is the average of cv-errors calculates after K iterations:

$$\mathrm{CVErr}(\hat{r}) = \frac{1}{K}\sum_{k=1}^{K}\mathrm{CV}_k(\hat{r}^{-(k)}).$$

After the model is generated, the final parameter β with minimum cv-error from the K-fold cross validation procedure will be used to test the performance in the testing data set. The performance is calculated by:

Pcorrect = (amount of correctly predicted sample) / (total testing data sample)

Results

With the K-fold, K=5, cross validation model on the training dataset:

```
>> k_5_cv
CVErr of the model
    0.0548

Chosen B parameter
   -0.0678
    0.0798
    0.4880
    1.1091

Lowest CV-error in the folds
    0.1758
```

With the Leave-on-out cross validation model on the training dataset:

```
>> leave_one_out
CVErr of the model
    0.2407

Chosen B parameter
   -0.0529
    0.0702
    0.4841
    1.0795

Lowest CV-error in the folds
   8.4360e-06
```

The Performance of each:

```
>> performance
K-fold with K=5 performance
    0.9778

Leave-one-out performance
    0.9778
```