Project 2

PART A regression

1.

In this project, the data that is used is based on many attributes in order to achieve the prediction of heart disease to patients. Specifically, the attribute that is used are ‘’oldpeak. This ‘’oldpeak’’ attribute is continuous, so with linear regression we will visualize some results. Also, there is no need to make feature transformation in our attributes such as one-of-K coding. From project 1, the data matrix X of our data are transformed in a way that mean is 0 and standard deviation 1, so a better visualization of our data is succeed. In the code the file ‘’Heart\_2.mat’’ contains all the data for the regression. In the matrix X, there are 12 attributes and 302 observations for each of them. The matrix Y contains continuous data of the ‘’oldpeak’’ attribute.

|  |  |
| --- | --- |
| **Attribute** | **Trestbps** |
| **Count** | 303 |
| **Mean** | 1.039604 |
| **Std** | 1.161075 |
| **Min** | 0 |
| **25%** | 0 |
| **50%** | 0.8 |
| **75%** | 1.6 |
| **Max** | 6.2 |

2. For this question a parameter λ is introduced in order to make regularization in our data and with linear regression to predict what will happen. First, we make a selection of our data, some of them will be TRAINING data and the rest TESTING data. We separate our data in blocks. For example, the first blocks i.e. 75% of our data will be training data and the 25% of the rest will be testing data. With the method of K=20-fold cross-validation, we can complete a whole examination of all the blocks and select which block should be training data and which will be testing data. It summarizes the results at the end, as cross validation uses them all and presents the best model. As it is shown in the lectures algorithm 5, some values of the parameter λ will show the amount of generalization error.

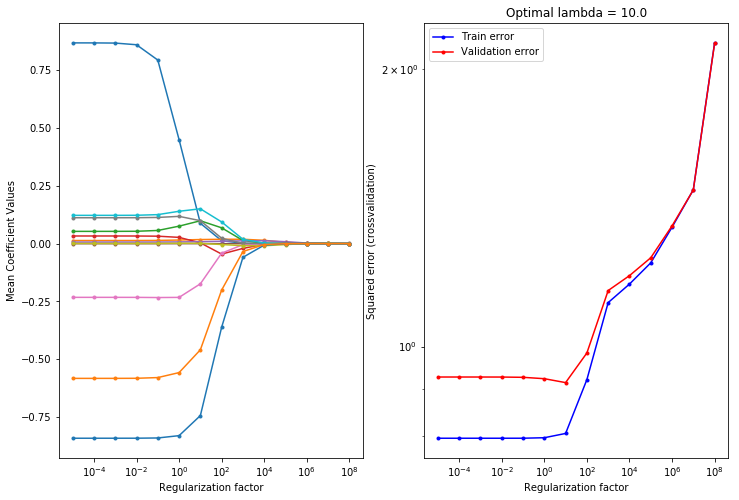
Trying different λ values, the optimal value will be selected in order to get the best model for our predictions.

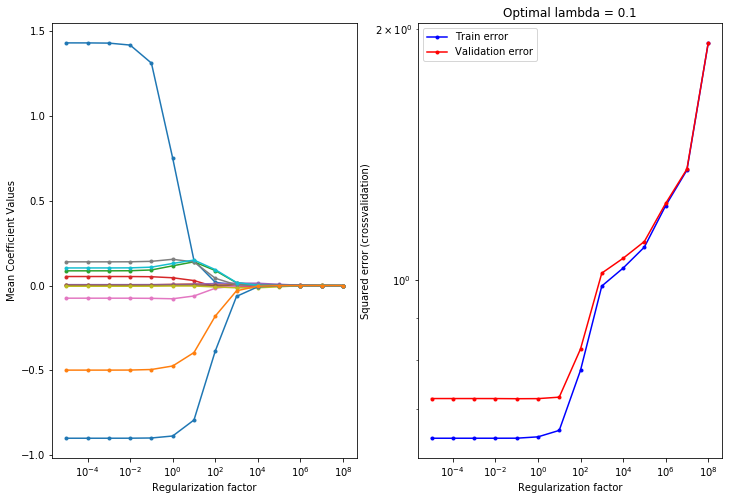
We can also solve for the optimal weights to get:

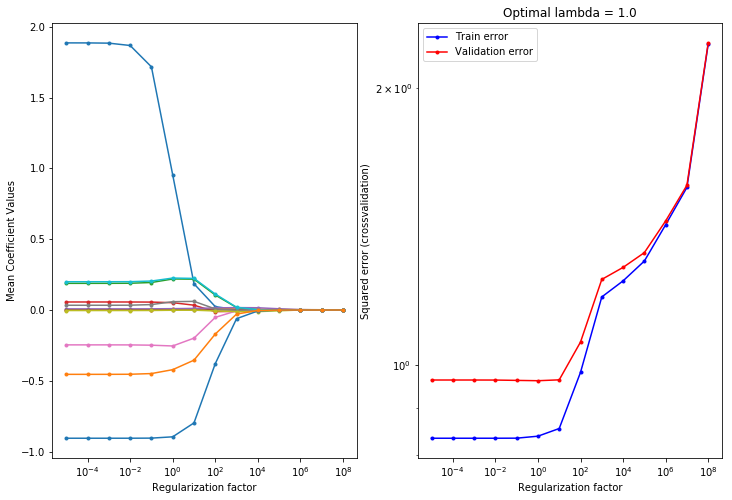


The Ridge and Lasso Regression are used together in order to get the best model and best fitting line. Overfitting will be done, as with Ridge regression maybe we get low bias and high variance and with Lasso there is high bias and low variance.

In this way, some results will be extracted also for the generalization error of our potential models.







3.

There are the results for our training and testing data for the errors.

Linear regression without feature selection:

**- Training error: 0.7775218435536388**

**- Test error: 0.8548581351366968**

**- R^2 train: 0.4206286197938574**

**- R^2 test: 0.351242487008715**

With selection feature.

Regularized Linear regression:

**- Training error: 0.784719590722744**

**- Test error: 0.8629690201700881**

**- R^2 train: 0.41526520943270323**

**- R^2 test: 0.34508708251978215**

In the previous graphs, we can see our attributes how they are contributing to our model in terms of training and testing data. Our code computes and generates the best case for parameter λ and generalization error.

*When λ is small, the weights are large indicating high variance but low bias. When λ is larger, the weights become smaller indicating lower variance.*

The optimal lamda is 0.1 and in the second figure we can see that the validation error is near the train error. Also, λ=0.1, λ=1 and λ=10 is tested as it can be seen in the python file.