Report on project 1

Lamyae OMARI ALAOUI
Nessreddine LOUDIY
School of Computer and Communication Sciences, EPFL Lausanne , Switzerland

I. Introduction

This project is a direct application of the machine learning course.In fact, It's about putting the knowledge and skills acquired into practice on a real world dataset.First of all, we have to explore our data to understand more our features through data visualisations, then we move on to feature processing in order to clean and make our data more meaningful. Once we have prepared our data ,the next step will be choosing a model by using machine learning methods we have seen in labs.

we will use original data from CERN to work on the challenge of finding Higgs boson.To choose our model we will implement linear regression using gradient descent, linear regression using stochastic gradient descent, least squares regression, ridge regression, logistic regression and regularized logistic regression.

II. Models and Methods

In this part, we will detail the steps followed in this project:

A. description of the data set

Our data set contains 30 features and the output is a categorical variable.
-7 features have 177457 missing values
-3 feature have 99913 missing values
-1 feature has 38114 missing values

B. Dealing with Missing values

In our dataset, the value of some variables for some entries is -999,which means that our data set is incomplete and not clean. To handle these missing values , we decided that for each feature ,we have to calculate the median without taking into account the missing values and then replace those missing values by the median .

C. Feature selection

To select features that will give us good or better accuracy whilst requiring less data, we decides to use the filter method:correlation coefficient scores.In fact, if two feature are very correlated we can delete one of them. we took as threshold 0.9.

D. Dealing with outliers

In order to detect outliers, we have used Turkey Method which consists of defining a lower and a higher bound using the first and the third quartiles and consider points outside [low,high] interval as outliers. Once we have detected outliers, we decided to replace those that are smaller than the lower bound by the lower bound value and those that are greater than the higher bound by this latter's value.

$$low = q1 - 1.5 * (q3 - q1)$$

$$high = q3 + 1.5 * (q3 - q1)$$

E. handle feature with low variance

when we apply the method described above , we can observe that some of our features have a very small variance. In fact, when we replaced missing values with the median, the first and the third quartiles of features that had high numbers of missing values are the same and are equal to the median, so when we handle the outliers, all entries for that feature take the same value which leads to a very small variance. So , we have decided to delete those features. we took as threshold 1e-4.

F. Feature scaling

In order to standardize our dataset, we have used Min-Max scaling.

G. Methods

linear regression using gradient descent
: for this function, we have chosen a zero initial vector, a maximum number of iterations equal to 10000 and a step size of 0.01 which gives us an MSE of 0.3364 and a categorical accuracy of 0.757. To save time, we can use stochastic gradient descent.

least squares
: we can save time and compute directly the optimum of the cost function bu using least squares methods, we decided to add a column of ones as an additional offset term. We obtain this time an MSE of 0.3348 and a categorical accuracy of 0.76.

Ridge regression
: To mitigate the problem of overfitting, we will use ridge regression by using cross validation to determine the hyper-parameter λ.we used 6 folds

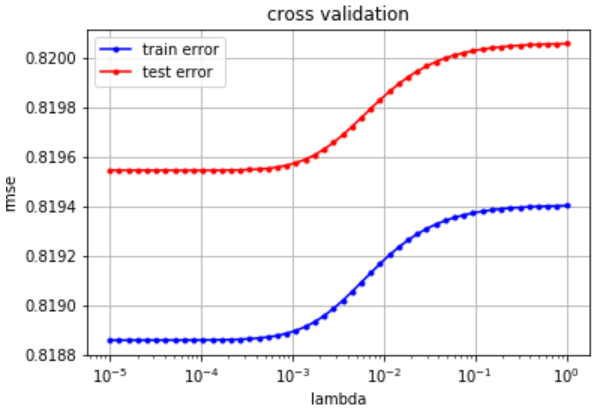

Figure 1. Figure 1: Effect of λ on training and test errors, calculated using 6-fold cross-validation

for cross validation ,and we have chosen λ between 1e-5 and 1. Since we have an important number of features, We decided not to lift the feature vector into a higher dimension. The best value for λ is 1.05e-4. We obtained a loss of 0.3353 and a categorical accuracy of 0.759.

Logistic regression

In our case , y can only take on discrete values (categorical variable), so it's more interesting to use a logistic regression. In order to use the expression of the cost function seen in lecture, we decided to transform our output to take values in 0,1 instead of -1,1, the problem is that we have 250'000 event in our training data set, so a single iteration will take a lot of time, we can think about using stochastic gradient descent to save time. The value of gamma we took is 1.2e-3 with 5000 iteration and with a batch size equal to 1000. we obtain a categorical accuracy of 0.76 and a loss of 124600.

Regularized Logistic regression

We could get an issue if our data is linearly separable, so we can add a penalty term to avoid this problem , we took λ equal to 0.2.

III. Results

from the results obtained, we can conclude that logistic regression works better but takes much more time, to remedy this problem we can work on a less number of points or take a smaller batch-size.

IV. Discussion

Normally, the logistic regression is supposed to give a better result than other methods. Nevertheless, the result obtained is very close to the method of the least squares, this can be due to the fact that we have not well constructed our features.

V. Summary

This project was very interesting and fun, because we applied our knowledge to a concrete application on a real dataset. We were able to implement the different methods, see the difference between them and analyze the results. Moreover this project gave us the opportunity to work on raw dataset, which means that we have to preprocess it before feeding it to our model.

We can conclude that to have a model that works, and to be able to compare different models, our data set has to be well prepared. In fact, data preprocessing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; Therefore, it is extremely important that we preprocess our data before feeding it into our model.