Project-I by Group MUMBAI

Audrey Loeffel
EPFL
audrey.loeffel@epfl.ch

Meryem M'hamdi
EPFL
meryem.mhamdi@epfl.ch

Abstract

In this report, we summarize our findings for project I applied on two Mumbai data sets. We started by analyzing the characteristics of our data, cleaning it using normalization and dummy encoding and investigating the nature of correlations between input and output data. In our regression data, we observed that three clusters exist; therefore, we needed to use logistic regression to classify our data and choose ridge regression for the best model. We also opted for logistic regression for classification data set and used cross validation with different values of alpha to minimize 0-1 loss error.

1 Data Description

Our regression data contains input variables $\mathbf{X\_train}$ and output variables $\mathbf{y\_train}$ for training. We have $N = 2800$ data examples of dimensionality $D = 73$ and type double. Our data consists of 61 real valued variables and 12 categorical variables. Out of these 12 variables, 4 are binary, 5 have three categories and 3 have four categories.

Test-data is provided as $\mathbf{X\_test}$ but its corresponding output is not given. Only $N = 1200$ testing data examples are provided.

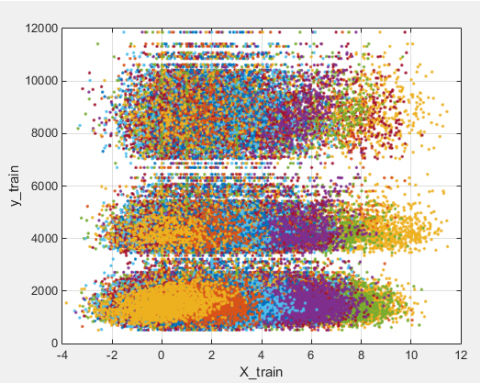
Our classification data consists of input $\mathbf{X\_train}$ of size $N = 1500$ and dimensionality $D = 28$ and its corresponding output $\mathbf{y\_train}$ provided for training purposes. Our classification input consists of 23 real valued variables and 5 categorical variables. Out of these 5 variables, two have 2 categories, one has 3 categories and two have 5 categories. The values of $\mathbf{y\_train}$ are either -1 or 1 making our data fall into 2 categories.

2 Exploratory data analysis, visualization and cleaning

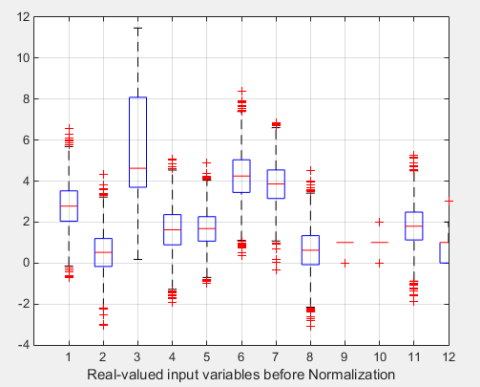
To learn about the characteristics of our input data, we examined closely how its variables are distributed. Figure 1(a) shows the scatter plot of our data before normalization. 1(b) shows that the data is not centred around 0, which means we need to normalize it. Figure 1(d) displays our data distribution after transforming it to have zero mean and unit variance.

From the scatter plot of y with respect to X shown in figure 1(c), we observed that our data contains three distinct clouds. The histogram of the output y in figure 1(e) confirms the presence of three clusters. The distribution is clearly not Gaussian overall but the distributions of the three clusters are definitely Gaussian. This implies that our input data is not fully correlated with our output therefore instead of finding one model to fit the whole training set, we need to find three models to fit each cluster separately. For any new testing point, we choose the best model to be used by determining the cluster to which the point belongs to using logistic regression.

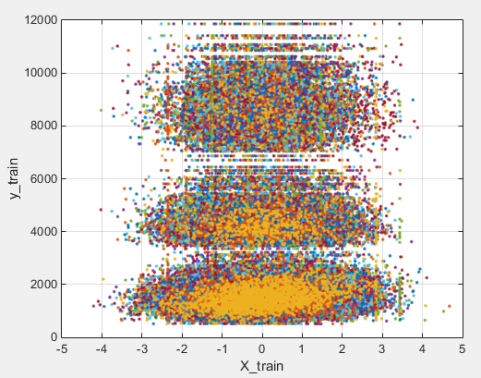
We computed the correlation of each feature's outputs with the input for each cluster as shown in figure 2(a). We observed that some features seem more correlated and can be interesting to use



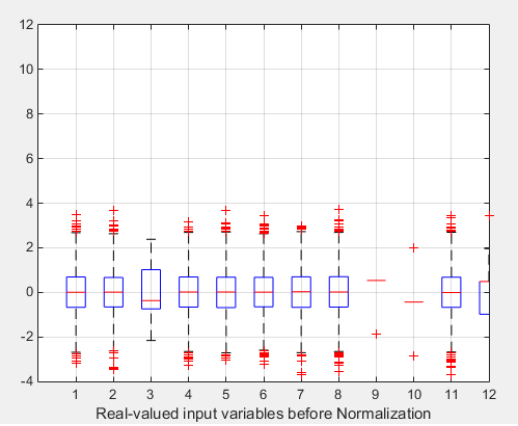
(a) Scatter Plot of Regression Output $\mathbf{y\_train}$ Before Normalization



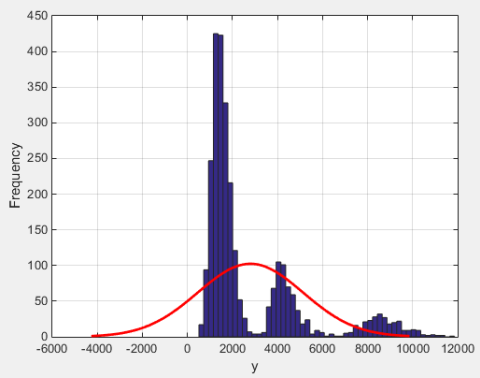
(b) Boxplot of real-valued $\mathbf{X\_train}$ before normalization.



(c) Scatter Plot of Regression Output $\mathbf{y\_train}$ After Normalization



(d) Boxplot of data after normalization

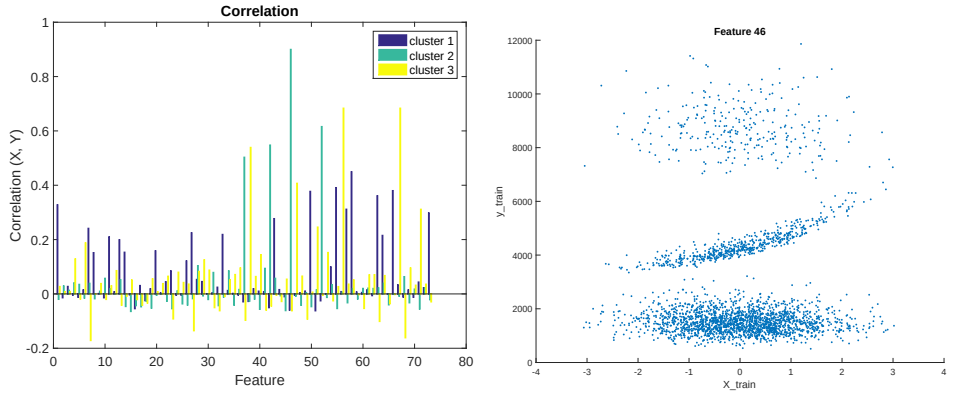


(e) Histogram of regression output data $\mathbf{y\_train}$

Figure 1: Exploratory Data Analysis and Normalization

for feature transformation. Figure 2(b) shows a plot of a feature with a higher correlation between cluster 2[1] and the outputs $\mathbf{y\_train}$.

[1]We named the clusters as cluster 1 has the lower outputs, cluster 3 has the higher outputs ans cluster 2is between 1 and 3.



(a) Bar chart of correlation between each cluster of each feature and $\mathbf{y\_train}$

(b) Feature with higher correlation in the cluster 2

Figure 2: Correlations between Input and Output data

We also used dummy encoding for all categorical variables in both regression and classification datasets.

3 Regression

3.1 Logistic Regression

First we classified the data into three clusters. We used two relevant features shown in figure 3 to separate the inputs in the first and the third clusters respectively. Once we have determined which inputs of the training set belong to which cluster, we applied logistic regression in the three clusters. The average errors for each regression are shown in the table 1.

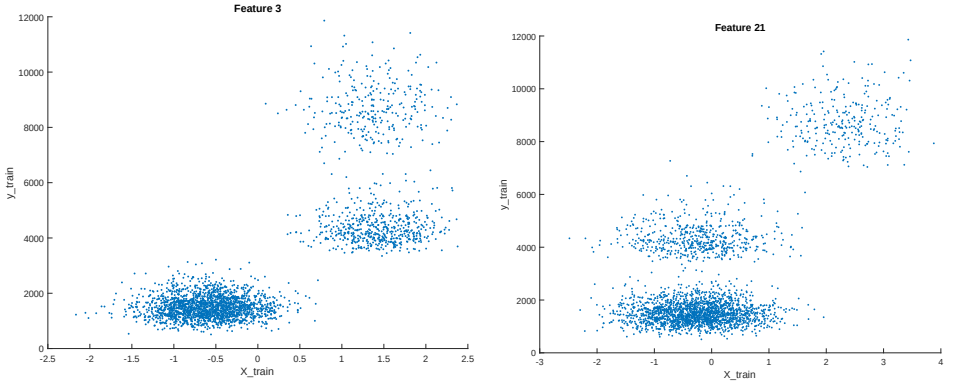


Figure 3: Scatter Plots of Regression Output $\mathbf{y\_train}$

Clusters	Training error	Test error
Cluster 1	12.650	12.624
Cluster 2	12.565	12.567
Cluster 3	10.614	10.554

Table 1: RMSE with Logistic Regression

For each testing point, we computed the probability that it belongs to a cluster with the betas obtained from logistic regression. We therefore classified each point in the cluster with the higher probability.

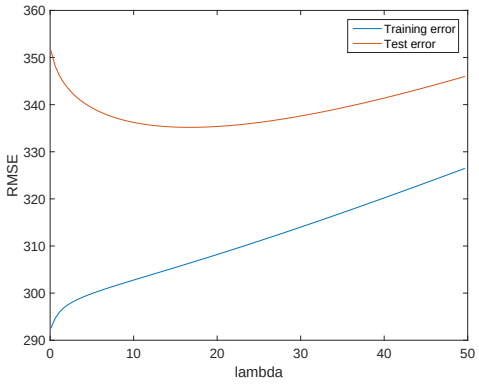
3.2 Ridge Regression and Feature Transformation

The next step was to apply a different model of regression depending on the cluster to which the input belongs to. For all three models, we chose ridge regression but we tried different models (with all features or with a subset of features and with or without feature transformation). All categorical variables were dummy encoded.

For the first cluster, we found that ridge regression with all features and without feature transformation gave the best result. We thought that choosing features with a high correlation would fit better to the model but the difference wasn't significant. For the second cluster, we found that ridge regression with all features gave the best result. We improved the performance by taking the square of the 46^{th} feature $(X_{46})^2$.

For the third cluster, we found that ridge regression with all features gave the best result. We improved the performance by taking the square of the 38^{th} feature $(X_{38})^2$ and of the 56^{th} feature $(X_{56})^2$.

Figure 4(a) shows that we optimized the error with lambda equals to 15. The table 2 shows the average error for training and test set of the best model fitting each cluster with $lambda = 2$



(a) Correlation

Figure 4:

Best Model of cluster	Training error	Test error
Model of cluster 1	144.294	151.978
Model of cluster 2	72.059	82.874
Model of cluster 3	127.720	184.616

Table 2: Average RMSE of models

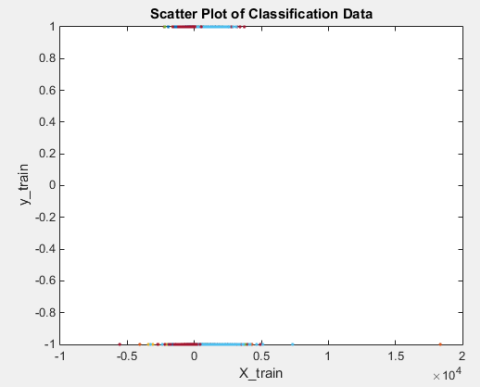
4 Classification

4.1 Logistic Regression

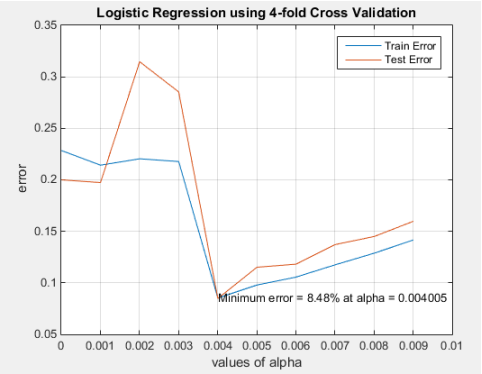
In classifying our data, we think that the application of logistic regression is well-suited in this specific setting. The dimensionality of input matrix $D = 28$ doesn't exceed the number of data points $N = 1500$ $\mathbf{D} \ll \mathbf{N}$. Therefore, we are not in the case where infinitely many solutions exists because the number of unknowns are smaller than the number of equations. We don't need to regularize our problem since there is low variance as shows figure 5(a).

We tried logistic regression with different parameters to find the best model to fit classification data and make predictions with minimal rmse, log loss and 0-1 loss errors. As in regression, we normalized non-categorical variables and dummy encoded categorical variables. We used cross validation

with different values of K and tried different values of alpha to plot learning curve corresponding to the train and test errors versus alpha. We got the best value for 0-1 error (approximately $8.480000e^{-02}$) when using logistic regression with alpha value: $4.005000e^{-03}$ and with no feature transformation.



(a) Scatter Plot of Classification Data



(b) Logistic Regression using 4-fold Cross Validation

Figure 5: Classification of data using logistic regression

4.2 Feature Transformation

We tried using different methods of feature transformation, like taking all the elements of the input matrix $\mathbf{X\_train}$ to power of 1/2, 2, 3 etc or using log. However, we didn't observe any significant improvement. So, we used only normalization and dummy encoding.

5 Summary

In this project, we analyzed regression and classification data set. We used cross-validation with K-Fold to test our models and compute the average error for the training set and the test set.

We tried different methods to find the best model minimizing the error and fitting at best with our data. We thought that selecting a relevant subset of features in our data which appear to be higher correlated with the output values minimizes the error. But we saw that the difference of the errors isn't significant. Finally, our best model with ridge regression has a average RMSE of 714.09.

We used logistic regression to classify our data and got the best value for 0-1 error (approximately $8.480000e^{-02}$) when using alpha value: $4.005000e^{-03}$. The use of feature transformation didn't make any significant improvement.

Acknowledgments

We would like to thank Dr. Mohammed Emtiyaz Khan for giving us the opportunity to work on this project which provided us with a rich learning experience and a great insight of what machine learning entails in the practical world. We also appreciate the help of all TAs and people we talked to during the labs. Without their help, our project could not have be done this way and we couldn't have figured out the correct transitions from theory to code. We collaborated as a team by writing the functions, trying the methods and writing the report together to be able to discuss the results.

References

Advice for Applying Machine Learning http://cs229.stanford.edu/materials/ML-advice.pdf

Feature Engineering: http://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/

A Few Useful Things to Know about Machine Learning: http://homes.cs.washington.edu/ pedrod/papers/cacm12.pdf

http://www.maths.tcd.ie/ dwilkins/LaTeXPrimer/ - tutorial on Latex

http://www.stdout.org/ winston/latex/latexsheet-a4.pdf - cheat sheet with useful commands for Latex

http://mirror.switch.ch/ftp/mirror/tex/info/first-latex-doc/first-latex-doc.pdf - example how to create a document with Latex

http://en.wikibooks.org/wiki/LaTeX - detailed tutorial on Latex