
Project-I by Group Rome

Viviana Petrescu
EPFL
vpetresc@epfl.ch

Abstract

We present our on two tasks, classification and regression on data that is not known. We process the data, investigate baseline methods and present our results here. For the regression task, our best model was \hat{y}_i and for classification was \hat{y}_i .

1 Regression

1.1 Data Description

The training data X_{train} contains 1400 observations, each 43 dimensional. One training sample has 36 real valued features and 7 categorical features. Our task is to predict the values for unseen test data, consisting in 600 samples. We measure the accuracy of our estimation using $RMSE$.

1.2 Data visualization and cleaning

Initially, our data was not centered, as seen in *Fig. 1a*. We changed the categorical features into dummy variables, leading to a new vector of size 56. X_{train} was then normalized to have 0 mean and standard deviation 1. We applied the same operations to the test data X_{test} on which we will report the results.

We plotted the correlation of every feature with respect to the output y_{train} and the scatter plots did not look random. We concluded the features explain the output, but we could not tell if one of them is insignificant.

The predicted values were real values in $[1000, 7000]$ and seemed to be grouped into two blobs. We believed initially the smaller blob represented outliers, but since it contained almost 10% of the data we decided not to ignore it.

After looking at every feature individually, we noticed that feature 36 offers a clear separation of the two blobs (see *Fig. 1b*). We therefore chose to fit two models, one in which feature 36 has values > 1.4 (after normalization) and one in which it is smaller, corresponding to smaller predicted values.

We visually observed some linear correlations between certain features such as feature 2 and 24, 13 and 16, 17 and 20, but we decided to keep them since we did not have time to experiment with their removal or to test their significance. This is corroborated by the fact that X_{train} is rank deficient, it has 57 columns after the use of dummy variables but rank 50.

If the input is normally distributed with mean 0 and std 1, then 99.99% of them appear between the values -3.891 and 3.891. We therefore remove any points that are outside this interval, considering them outliers.

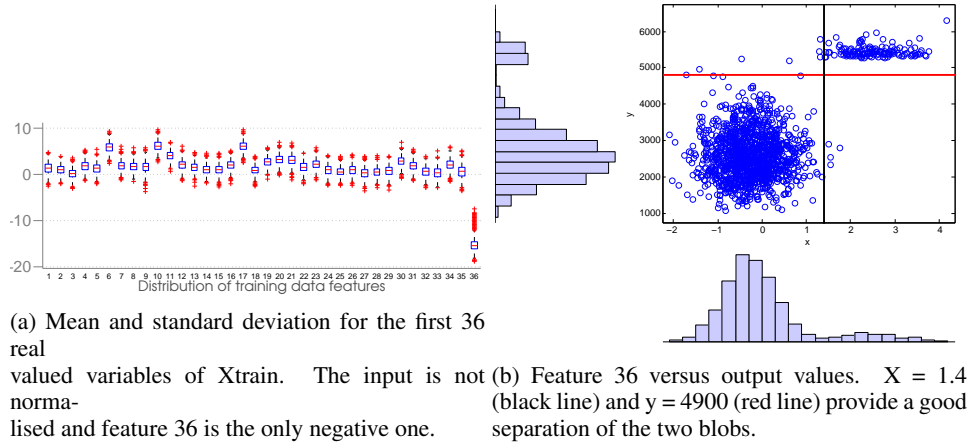


Figure 1: Data visualization.

1.3 Ridge regression Baseline methods

1.4 Feature transformations

We tried polynomial and exponential transformations of the features. The exponential transformation did not prove to work well and we focused on polynomial regression. We only transformed the first 35 variables and kept the categorical variables as they are since polynomial transformation of categorical variables does not have an intuitive interpretation.

The degree of the polynomial was varied from 1 to 10 for both blob-models. In *Fig2a,2b* we plotted the mean training and validation rmse error only the degree varied between 2 - 10 and 2 - 6 respectively for readability (for other values the error was too big).

Using 5-fold cross validation with $\lambda = 1e - 5$, $\alpha = 0.1$ we notice for the first blob, RMSE train and validation is smaller for a polynomial of degree 3. For the second blob, using $\lambda = 0.001$, $\alpha = 0.1$ we obtain a best fit for a polynomial of degree 2.

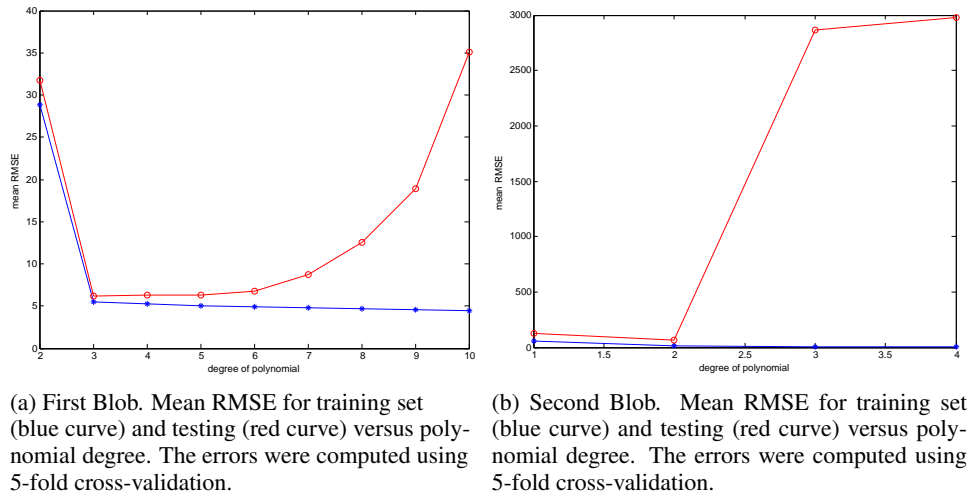


Figure 2: Selection of polynomial degree

We increased lambda to prevent overfitting, since we have 10 times less samples for the second blob. However, we still noticed that our RMSE for the validation set was bigger for the second blob. This might be due to the fact that we have a very small set for both training and testing of the second model.

Both polynomial regression models significantly outperform the normal ridge regression, which gave a RMSE ≈ 100 .

Note also that our final RMSE (reported on the whole training data) is smaller than the sum of the two RMSE.

2 Classification

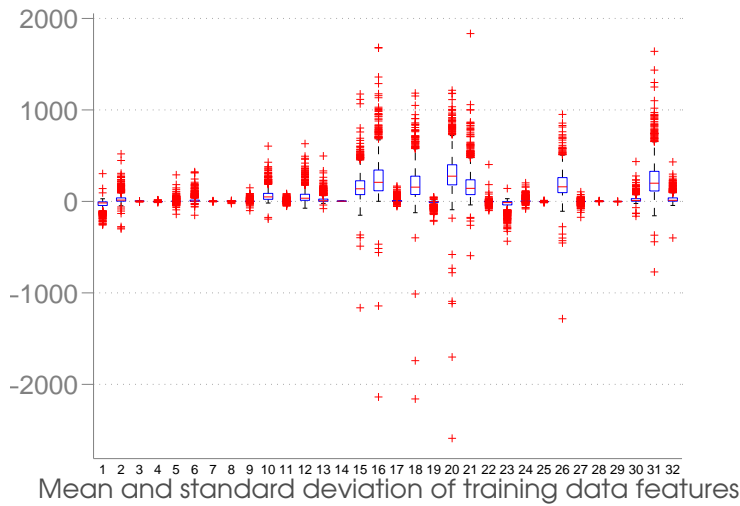
2.1 Data description

We have a two-class classification problem. The training data X_{train} contains 1500 observations, each 32 dimensional. One training sample has 31 real valued features and 1 categorical feature, the 14th feature. Our task is to predict the category for unseen test data, consisting in 1500 samples. We measure the accuracy of our estimation using $RMSE$, $0 - 1loss$ and $log - loss$.

2.2 Data visualization and cleaning

We noticed that the samples were not equally distributed among classes, about 35% of the samples were coming from one class and 65% from the other. We changed the labeling from -1/1 to 0/1.

The training data was again not centered as seen from *Fig.3a*. We noticed more outliers in the data of classification than in the one from regression. After removing the outliers and changing feature 14 to a dummy variable we normalized the data to have 0 mean and std 1.



(a) Mean and standard deviation for the training data features. The input is not normalized and contains outliers.

Figure 3: Data visualization.

2.3 Logistic regression

2.4 Feature transformation

3 Summary

Acknowledgments

References