Project-I by Group Ha-Noi

Frederic Ouwehand **Marco Antognini**

Fall 2015

Abstract

This report states our findings and methodology for the regression and classification problems in Machine Learning, such as identifying different sources used to produce data, which feature transformations should be used depending on the context and how to compare different models. We conclude with two prediction models and their respective expected error when predicting response for unseen data.

1 Regression

In this section we first present the data for which we need to predict the output before dwelling into the details of the different regression models we experimented with. Finally, we present the result of our prediction for unknown data.

1.1 Data Description

Training and testing data is provided. The goal is to learn from training data, predict response for the test dataset and estimate expected error that we do on unseen data. In our case, the response is approximatively between 0 and 12000. The training and testing datasets are made of respectively 2800 and 1200 data points, each with 67 features. From those features, a total of 12 are categorical: 3 are binary, 6 are ternary and 3 have quaternary.

1.2 Data Exploration

The histogram of $\mathbf{y}$ shows three Gaussian curves that are likely to be generated by three different input sources. Therefore, we might want to use a different learner for each source.

Two interesting features were identified as allowing us to classify the data points efficiently: a) if $x_n^{(62)} < 15.75$ then $\mathbf{x}_n$ belongs to the first source, b) if $x_n^{(62)} >= 15.75$ and $x_n^{(25)} >= 15.25$ then $\mathbf{x}_n$ belongs to the second source, c) otherwise $\mathbf{x}_n$ belongs to the third source. Linear boundaries 15.75 and 15.25 were chosen by visual inspection to minimize the number of misclassified data points. The clustering produced is shown in Figure 1. Despite our carefully chosen classification the histogram shows that some data points are misclassified, raising the problem of outliers handling. These outliers are circled in red.

Other clustering methods such as 3-Means and Gaussian Mixture Model were investigated but yielded poor results compared to our manual method.

1.3 Model Explorations

Several methods were evaluated on the dataset using the training/validation methodology with a $0.7/0.3$ ratio. The evaluation is repeated 20 times with a different random permutation of the data points and produces an approximation of the expected Root-Mean-Square Error (RMSE) and a variance for each method. A full comparison is shown in Figure 2, in which each method is shown with

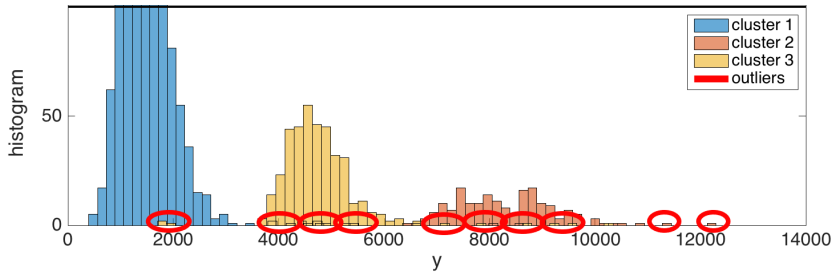


Figure 1: Histogram of training response classified in three sources

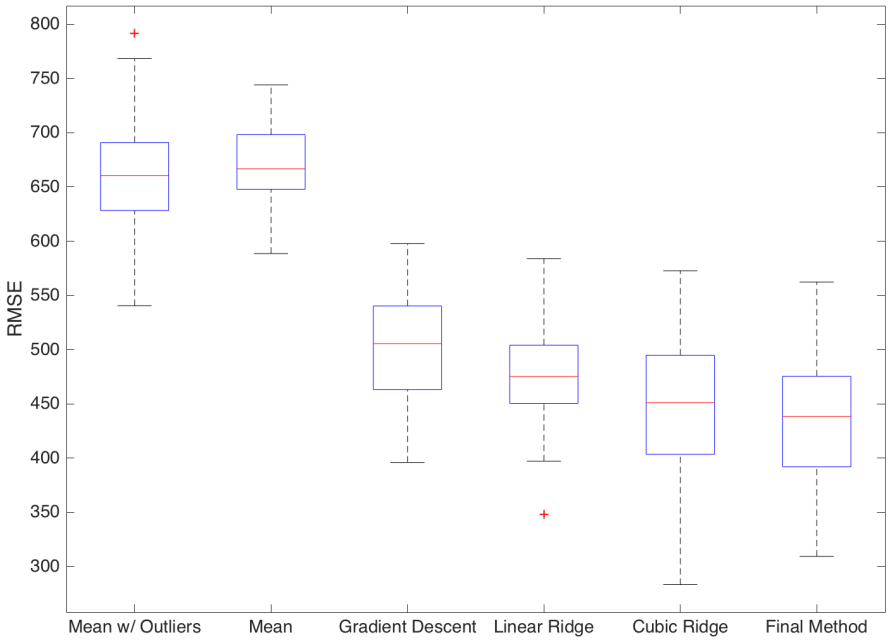


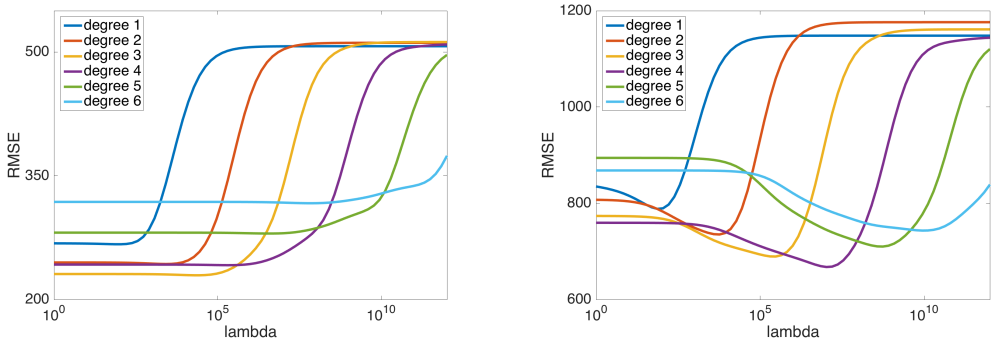
Figure 2: Validation RMSE per method

its error on the validation set. Note that methods use exclusively the training set to compute their models. Hence, hyper parameters such as lambdas for penalized regressions are computed using training data which may be split into training/testing pieces in the case of K-Fold. The intuition is that our final error estimate should not be computed using data that has been used to tune our models.

1.3.1 Path to the Best Model

Naively predicting the mean of the training responses yielded a median RMSE of 2250. This is 3 times more than our baseline which matches each data point to its source and predicts the *mean* of the training points in this source. With a median RMSE of 650, this baseline validates our intuitions about building three separate models for each identified source. It is also another concrete case were building many models is better than just one (see *Domingos(2012)*). This method is the left most one in the Figure.

Despite our careful classification there exists a tiny percentage of misclassified points that have a strong impact on learning our models. For instance, one point belonging to the third source but



(a) source 1 (b) source 2

Figure 3: λ vs Validation RMSE using 10-Fold for different degrees

classified in the first source will be an extreme point since the mean between these sources differ by a ratio of 4. Methods such as Ridge Regression or Gradient Descent are based on RMSE and does not handle well such extreme outliers. However, since our sources appear to be Gaussian, discarding points farther than 3 standard deviations is a good compromised between not having outliers and keeping most of our data points and the information they represent to build our models. Removing outliers and predicting means is shown in the second column of the Figure. This method decreases the variance of our models while keeping approximatively the same median RMSE. All further methods were computed on the training data with outliers removed.

After normalizing our features ($\mu = 0, \sigma = 1$), *Gradient Descent* was successful in computing models with lower biases and consequently reducing median RMSE by 20%. For the sake of efficiency, the alpha step size was computed using Line Search making the method usually converging in very few steps.

We might benefit from models with a bit more bias but less variance. *Linear Ridge* expresses well this bias-variance trade-off in Figure 2. Both variance and median RMSE is improved using a penalization term computed by a 10-Fold on the training data, choosing the lambda which corresponds to the minimal mean RMSE on the testing set among the 10-Folds. Setting $K = 10$ is known to predict well the expected test error since it has less tendency to overestimate it (see *HTF* Section 7.10).

Polynomial basis functions can reduce further model bias. Figure 3(a) and 3(b) compares the performance of several different degrees feature transformations on the first and the second source. Degree 3 performs the best on both of these sources. Figure 4 is useful to determine how much the testing error variates for each degree and for each lambda on the third source. Variance problems occurs for the third source because of its low number of data points. The figure shows that degree 1 has less variance than degree 2 or degree 6. Taking into consideration this overfitting problem, the *Cubic Ridge* method makes some improvement on median validation RMSE while keeping the same number of features.

To benefit from the number of data points for the first source, the *Final Method* increases the model searching space for the first model by extending its features with $\Phi = \left[x_n^2, x_n^3\right]$. It has the effect of reducing both variance and bias.

Categorical features were left aside because they were not relevant in improving our models.

1.3.2 Learning Curve

We ran our experiment several times with different ratios for splitting the given training data into training and validation sets. This resulted in a relatively stable training and validation RMSE at 70%.

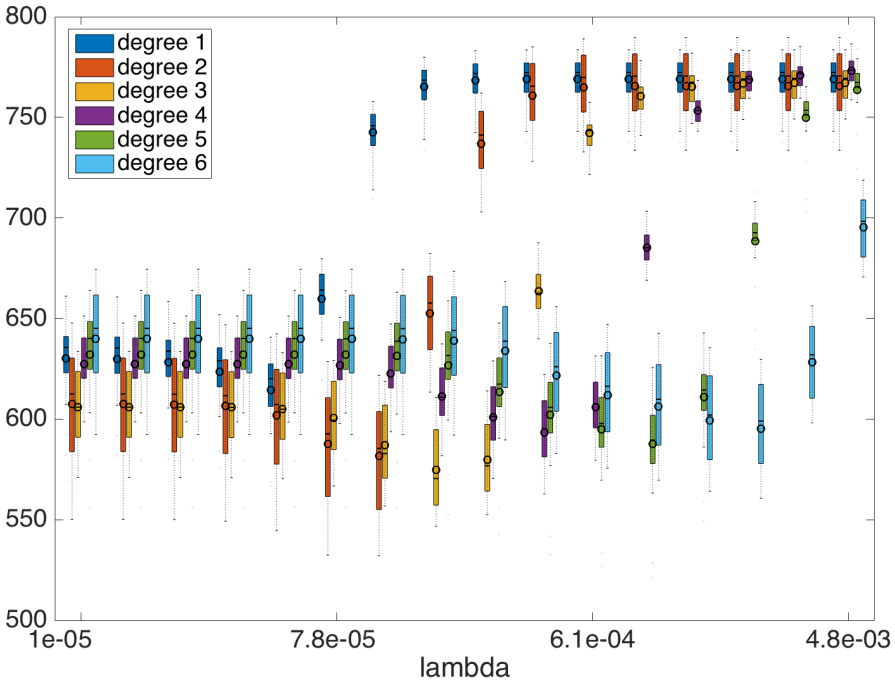


Figure 4: λ vs Validation RMSE using 10-Fold for different degrees for source 3

1.4 Results

Our baseline has a median RMSE of 660 that is close to the mean, a 25th percentile of 628 and a 75th percentile of 691. Compared to it, our final method offers a 30% improvement for the median, which is around 440, while keeping a similar precision. Finally, we expect the RMSE error to be around 463 for unseen data.

2 Classification

In this section we discuss how we applied different classification methods on a different dataset and conclude with the estimation of our error for the classification of unknown data.

2.1 Data Description

Training and testing data is provided. Using the training data, which has in our case a response of either -1 or 1, our goal is to predict the probability that a data point from the testing set is classified as 1. The training and testing datasets are made of respectively 1500 data points, each with 40 features. From those 5 are categorical: 3 are binary and 2 quaternary.

2.2 Path to the Best Model

Following the same methodology as in the regression problem, several methods are evaluated and their validation error averaged on 20 seeds. These methods are shown in Figure 5.

The training response is composed by a majority of 1. Predicting only ones is correct in more than 60% of cases and is a good baseline to compare more complex methods against.

The 11th feature is decisive for predicting the model since nearly all data points having their value on this feature above -10 belong to the class 1. A *naive*, but informed, method will predict ones for

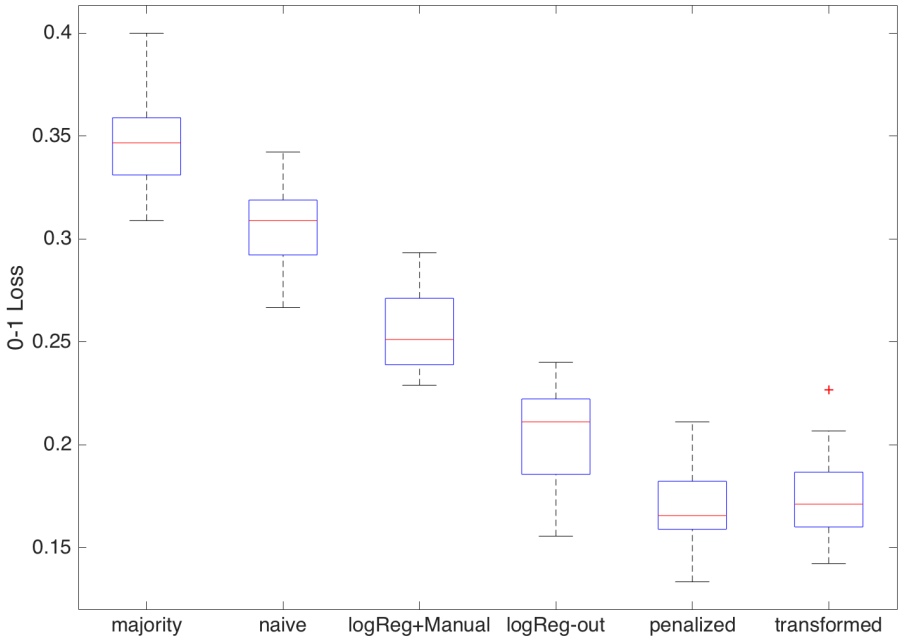


Figure 5: 0-1 Loss of Validation data per methods

such data points and flip a fair coin otherwise. Such strategy already make 10% improvements over the baseline.

We may wonder how much more complexe models add up to our naive method. After transforming the categorical features with the dummy encoding technique, we used the *naive* binary decision on feature 11th and a *Logistic Regression* on the rest of the datapoints. The predictions were improved with another 15% and is plotted with the label *logReg+Manual*. This proved that investigating more sophisticated approaches such as Logistic Regression is useful.

One may ask if the Logistic Regression alone could get this 11th feature particularity by itself. The reality was even better as this method improved over previous ones. Following the same reasoning as in the Regression problem, outliers are removed and results are shown under label *logReg-out*. Outliers are also eliminated in further methods. Using Line Search with LogLoss function, Logistic Regression was efficient in finding the model under which the training data is more likely to have been generated from.

Adding a penalization term proved useful in reducing both the median Validation Error and its variance as shown in *penalized* method.

Looking at features distribution such as the one shown in Figure 6, we could doubt that Logistic Regression performs at its best since it expects a more Gaussian looking curve. The *transformed* method at the right of Figure 5 first transforms the data points by taking the squared root of their absolute value and then applies Penalized Logistic Regression. Little improvement is made over the last method which was chosen as a final one.

2.3 Results

Compared to our baseline, which had a 0-1 loss median of 0.35, our final method offers a 2x reduction of misclassification on our validation set with a median of 0.17.

3 Summary

Given a regression and a classification dataset, several methods were evaluated against baselines in their ability to produce good models with low variances. More sophisticated methods proved useful under some conditions such as outliers removal and normalization. Performing data exploration along with solving numerical issues were key in applying well known methods *in practice*. Most importantly, our analysis encountered many practical aspects of Machine Learning and caveats such as K-Fold to fight overfitting or the model bias/variance trade-off.

Even though our best models predict outputs reasonably well, we have no insight if the correlation calculated has also a causation validity since we do not know the nature of our data.

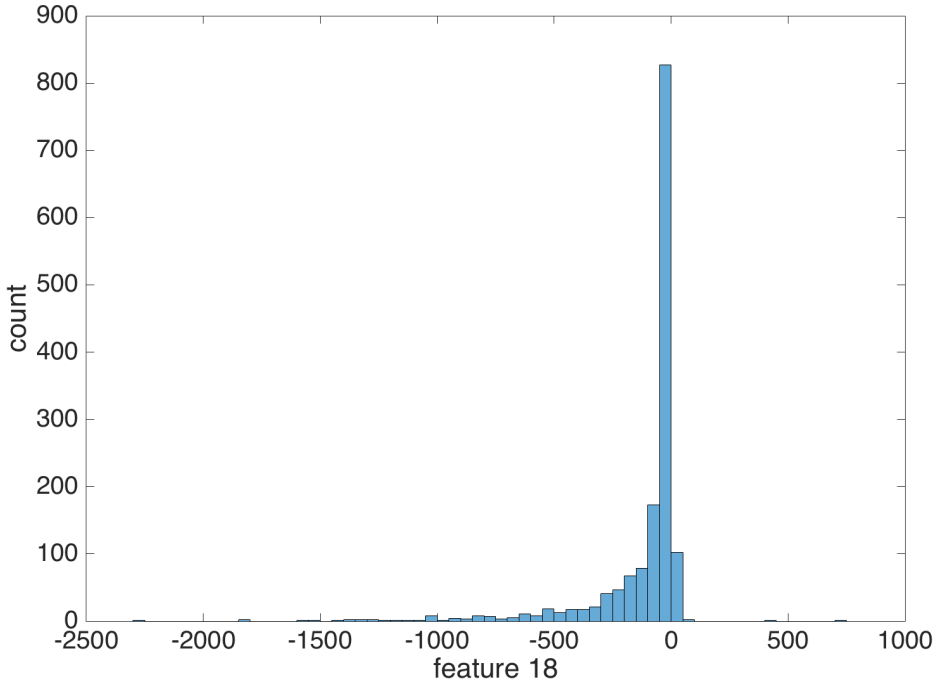


Figure 6: Histogram of feature 18

References

[*HTF*] Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome. The Elements of Statistical Learning : Data Mining, Inference, and Prediction. 2nd ed. New York, N.Y: Springer, 2009. Springer Ser. in Statistics. Web.

[*Domingos*] Domingos, Pedro. "A few useful things to know about machine learning." Communications of the ACM 55.10 (2012): 78-87.