Class Project 1

Kanbak Can, Kubik Anna, Vostriakov Alexander
Ecole Polytechnique Federale de Lausanne

***Abstract*—In this project, various basic machine learning methods, namely least squares, ridge regression and logistic regression are used for classifying a given data set. Data is preprocessed to decrease the classification error percentage by cleaning the missing values and using a polynomial basis. The methods are implemented using the formulas given in the lecture notes except for logistic regression which are changed to better fit the given data. In the end, the best performing method is found as regularized logistic regression with more than 82% correct classification.**

I. Introduction

Currently, machine learning is one of the most popular tools and it is used in a wide variety of areas from experimental physics to social networks. One of the simpler methods in this is to use various regression schemes to get a dependence function between the given features and the output value and use this function for classification. Another simple method, called logistic regression improves the simple regression by considering the classification usage in the end. In this project, we have implemented various requested functions to get a good classifier for a given data set and this report presents how we implemented these functions as well as the methods we used to improve the results of the classification of the test set.

II. Implementation

A. Data sets

The train data used for this project is composed of the output variables $\mathbf{y}$ and of a set of input variables $\mathbf{tX}$. The input variables $\mathbf{tX}$ have a dimensionality $D = 30$ and a cardinality of $N = 250000$ - same cardinality as the output variables $\mathbf{y}$. The 30 variables of $\mathbf{tX}$ contains 29 real variables and one categorical ones containing 4 categories.

The test data used to generate prediction of possible outputs $\mathbf{y_{pred}}$ consists of a set of input variables $\mathbf{tX_{test}}$ of the same dimensionality as $\mathbf{tX}$ and a cardinality of $N = 568238$.

B. Data pre-processing

After a visual inspection of the data, it appeared that most the input variables contained the value -999. We supposed that those values corresponded to an error or a missing value.

To treat this issue, we replaced those values by the average value of the corresponding input if that input had more good values that errors. Otherwise, we tried to not take in account that input, but it appeared less efficient.

Furthermore, it seemed that in the last input, that error value was 0 instead of -999. Therefore we applied the same method as previously. But after testing with the same code, the treatment of the last output showed slightly lower performances, as shown in Figure Fig.1.

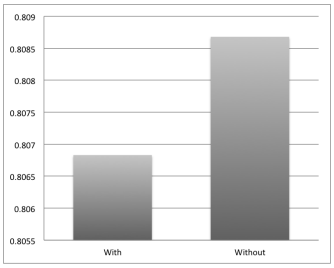

Fig. 1. Kaggle results with and without the treatment of the 0 values in the 30th column.

After the treatment of missing values, a polynomial basis is constructed to increased the feature space. It consists of square of each feature, multiplication of all features with another and third to tenth powers of each feature as well as the features themselves. In total this means we have a 736 dimension feature space to work with. Finally, after this basis is constructed, the features are divided by their standard deviation to get the standard deviation of all to 1. This is required as high valued features(e.g. powers of 10) can mess up with the numerical computation.

C. Regression

This section treats about the implementation of the six requested regression methods.

1) Least Squares: For all the Least Squares methods – using normal equations, gradient descent and stochastic gradient descent – we applied the formula from the lecture notes.

For the two gradient descents, the parameter were optimized and were set as: $\gamma = 1/L$ which is a common heuristic where L is the Lipschitz constant of the objective function($L = 2\|X^T X\|$ for least squares) and $max\_iter = 100000$.

max_iter could have been set much higher for better results but would have had resulted in much higher computational resources consumption.

2) Ridge Regression: The ridge regression method is done using the analytical solution from the lecture notes. The regularization term λ is chosen using 10-fold cross-validation. In each fold, the error is calculated for every λ value and then the λ which gave the least test error in average is used for the final training using the whole data. This way, it was intended to get a λ value which would make the weights connected to the data but also not over-fit.

3) Logistic Regression: In the data that was given, the output values y was from the set {-1,1}, while the formulas

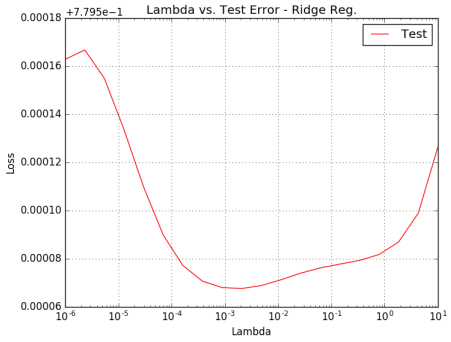

Fig. 2. Test error for ridge regression for different lambdas.

for logistic regression on the lecture notes was for output values {0,1}. So, by using the fact that for sigmoid function $\sigma(-x) = 1 - \sigma(x)$, the conditional probability of y_n was set as $Pr(y_n|x_n, w) = \sigma(y_n w^T x_n)$. So the loss function became:

$$L(w) = -\sum_{n=1}^{N} \log(\sigma(y_n w^T x_n)) = \sum_{n=1}^{N} \log(1 + e^{y_n w^T x_n})$$

and the gradient for the descent became:

$$\nabla L(w) = -\sum_{n=1}^{N} \frac{y_n x_n}{1 + e^{y_n w^T x_n}}$$

With these in hand, gradient descent was used to find the weights for classification and again $\gamma = 1/L$ was used where this time $L = \frac{1}{4}\|X^T X\|_2$.

max_iter is chosen as 2000 to limit the running time of the algorithm. Similar to least squares, a higher max_iter can result in better solutions.

4) Regularized Logistic Regression: In this part, the same loss function as the previous part, along with a regularization term $\lambda\|x\|^2$. γ was again set to $1/L$ where L was accordingly increased by 2λ. Similar to the ridge regression, 4-fold cross validation was used to find a good value for λ. However, the number of folds and the number of tested λs was decreased to reduce the complexity of the implementation. Similar to the logistic regression, max_iter was chosen 2000.

D. Results

As expected, the Regularized Logistic Regression method performs better that the other methods, and thus it was the method used for the final submission. It can handle the outliers better than least square type solutions, and it is better than logistic regression as the regularization term lowers the over-fitting of simple logistic regression.

III. Summary

For this project, we mainly focused of improving our results by selecting the best regression and optimizing the parameters and the treatment of the information. By doing so, we managed to reach a decent result.

As a potential improvement that we talked about but did not have the occasion to implement was to use the average or a weighted average of the classifications from the different regression methods.

Other possible improvement could have been to treat the missing values not by simply averaging over the whole input but by analyzing the test data to find similar entries and use them to guess a value for the missing value.