Project-I by Group MexicoCity

Kevin Serrano
EPFL
kevin.serrano@epfl.ch

Youssef El Baba
EPFL
youssef.baba@epfl.ch

Alexandre Helfre
EPFL
alexandre.helfre@epfl.ch

Abstract

In this report, we discuss our implementation and findings for the project-I. Todo at the end

1 Five functions

- *leastSquaresGD(y,tX,alpha)* : Least squares using gradient descent, alpha is the step-size
- *leastSquares(y,tX)* : Least squares using normal equations
- *ridgeRegression(y,tX,lambda)* : Ridge regression using normal equations, lambda is the regularization coefficient
- *logisticRegression(y,tX,alpha)* : Logistic regression using gradient descent or Newton's method
- *penLogisticRegression(y,tX,alpha,lambda)* : Penalized logistic regression

These functions are different machine learning methods.

Formula? Not even sure this section is useful

2 Data observation

We have two data set. One for regression and one for classification.

Regression Consists of output variables $\mathbf{y}$ and input variables $\mathbf{X}$. The number of examples is $N = 1400$ and each input $\mathbf{x_n}$ has dimensionality $D = 48$. The first 34 are real valued and the last 14 are categorical, included 5 that are binaries.

Classification Consists of output variables $\mathbf{y}$ and input variables $\mathbf{X}$. The number of examples is $N = 1500$ and each input $\mathbf{x_n}$ has dimensionality $D = 35$. check hist

3 Data visualization and cleaning

Histogram, correlation, applied methods
Most variables normally distributed with various means. The first 34 input variables are Gaussian and the last 14 are categorical. Given that, we normalized and centred the first 34 variables and let the others untouched. Plotting the histogram showed us that the data are compact and there is no outlier.show figure hist? Or boxplot like in template
WHY WE ARE USING PEARSON: To help us with the task of finding which transforms could

be useful in better predicting the output from the given predictors, we decided to use the help of the pearson linear correlarion coefficient. This coefficient basically gives a number indicating the strength (in its absolute value) of a linear model fit between the input and the data. Obviously, we could visually inspect multiple scatter plots between each variable and the output, and also with the other variables, but this can (for obvious reasons) be somewhat of an overkill and could lead to subjective mistakes. The pearson coefficient allows us to selectively choose wich variables (and which possible transforms accordingly) could be used to fit a good model, in a very fast manner (simply adding columns to tX and checking their corrrelation coefficients). Using this strategy, we could rapidly pinpoint which transforms were probably most useful. The spearman coefficient also performs a similar computation, but tests not only the linear relationship between variables, including any exponential relationship or such. Therefore for some transformations, even though a linearly-fit model would not find them useful, the spearman coefficient would rank them as correlated (to the output and other variables) as the variable they are coming from.

We checked the correlation between output and input variables. The correlation matrix helped us to discover which variables where correlated to no other variables nor the output. We could remove these variables.More details to add here :youssef text

4 Least Squares

Applying the least squares method on the data without cleaning yields a $RMSE$ of 0.48244.

5 Best method : Ridge Regression

explain what is the best method and why for our dataset, add some figures and results

6 Feature transformations

Different transformation (myPoly, sqrt, etc)

7 Summary

summarize in a few lines and write down all the final results

Acknowledgments

References