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.

4   Best method : Ridge Regression

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

5   Feature transformations

Different transformation (myPoly, sqrt, etc)

6   Summary

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

Acknowledgments

References