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

need to move that somewhereNote that all the gradient descent methods used the eps threshold (machine dependant) in Matlab in the gradient descent function, to allow for a maximum number of iterations. This might make the gradient descent process slow though, so feel free to set another number if needed. We changed it later to 1/1000, so the numbers for the test Errors might be different depending on this parameter when you run the code (the global trends are the same though), since the number of iterations will be different.

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

For regression, $\tilde{\mathbf{X}}$ has rank 49, so no redundancy in variables. For classification, the rank of $\tilde{\mathbf{X}}$ is 35 plus the one vector. this tells us that $\tilde{\mathbf{X}}$ is not ill-conditioned.
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. We used the Pearson correlation.

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.

4 Least Squares

The mean gives a $RMSE$ of 0.99964, over normalized $\mathbf{y}$.

Applying the least squares method on the data without cleaning yields a $RMSE$ of 0.48244. Given that, we see that the $RMSE$ is cut by half so the input data are meaningful.

We got the RMSE down to train error 0.3605 (with estimated test error 0.3805) using direct least Squares, simply by adding the squares of the 18th and 34th variables. When we added the squares of all the variables, we only got a train error of 0.3405 but this is clearly overfit because the corresponding test error estimate is 0.3765.

We tested taking the squares of the 18th and the 34th variable individually and a curious thing happened: when we removed the 18th variable (squared) the test error increased to to 0.3881, but when we removed the 34th variable (squared) the test error increased much more significantly to 0.4987.

Move to transformation ?By isolating the 18th and 34th input variables, and producing polynomial transforms of them using the mypoly function, and then measuring their Pearson correlation coefficient with the output, we could see that the cube and the power 5 of the 18th variable, and powers up to 6 for the 34th variable are correlated with the output. When adding these feature transformations, we could lower the test error to 0.3307, which is quite lower now than the starting test error of 0.48244

5 Ridge Regression

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

By visually inspecting the Pearson correlation graphs, a good strategy to spot well correlated input variables was to take the intersection of the two methods (were only considering). To do that, we tried multiplying the two coefficients (for each variable) and taking the mean. Both strategies clearly showed the 18th and 34th variables are best correlated. We can limit our variable transform to these two (we only have 1400 data points, so with a dimension 48 were already lacking data, by introducing spurious transforms we would be adding fuel to the fire and causing overfitting.)

6 Logistic Regression

We tried to use simple logistic regression using all the variable, the estimated $RMSE$ test error was 0.27099. Also we tried using sqrt on the 11th and 24th variables which actually improves our test error to 0.26779. We can also use squares of the variables instead of the square root and applying this alone the test error decreased to 0.26715. Adding both transformations got us a decrease to a test error of 0.26474. We also tried to remove all categorical variables as we did in regression, but the test error increased to 0.3396, so their information was needed and cannot be ignored.

We then found that, since Pearson correlation predicted better correlation of square roots than squares but squares were actually better able to decrease test error, it might be a good idea to include 3rd degree transforms of the 11th and 24th variables. Added them, the test error increased to 0.26656. Thus taking higher orders doesnt seem to help.

to section transformation?We then tried to include a dummy variable transform, but that increased the test error significantly to 0.32625, therefore were not going to use them.

7 Feature transformations

Different transformation (myPoly, sqrt, dummyvar,etc)

We tried dummy variables for all regression categorical ones. We could only find a significant linear correlation (using Spearman) for the 2nd class of the variable 44, 2nd and 4th class of the 47th.

We tried keeping the polynomial transforms and removed all categorical variables. The test error went down to 0.33043.

Adding back the mentioned dummy variables 44 and 47 actually increased the test error to 0.33105. This lead the matrix $\tilde{\mathbf{X}}$ to become ill conditioned.

We also tried adding square root of the 18th and 34th variable, that means $\sqrt{|X_{n18}|}$ and $\sqrt{|X_{n34}|}$ of $\mathbf{X}$. With the 18th the test error went up to 0.32137 and with the 34th it went up to 0.39349. So this doesn't help and we decided not tu use square root transformation.

We decided that were going to ignore all categorical variables for regression, and also all the uncorrelated variables, giving a test error of 0.33173 while including the significant transformations of the 18th and 34th variables discussed before. We decided to opt for a simpler model because of Occams Razor.

8 Summary

summarize in a few lines and write down all the final results For the regression, we saw that the best method was Ridge Regression. The best parameter was $\lambda = 39.0694$? and we achieved a test error of 0.3073.

For the classification, the best method was penalized logistic regression. We obtained different results depending the type of error we computed.

rmse Best parameter is 0.0621 and we achieved a test error of 0.2663.

01 loss Best parameter is 0.4642 and we achieved a test error of 0.0962.

log loss Best parameter is 3.1623 and we achieved a test error of 0.2812.

Acknowledgments

References