CSE847: Homework 4

Shihab Shahriar Khan

March 31, 2022

Code in Jupyter notebook format is available at : https://github.com/Shihab-Shahriar/cse847-hw4.

1 Logistic Regression: Experiment

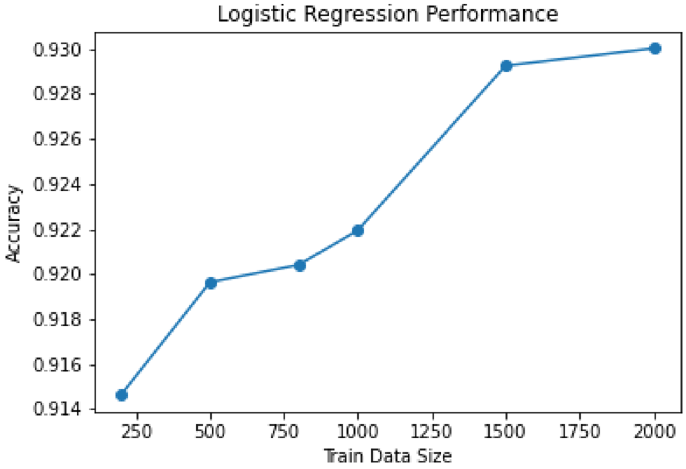

Figure 1: Performance of Logistic Regression model with respect to training data sizes.

In this part, I implemented logistic regression code using first-order gradient descent. Labels were encoded in +1/-1 format. Entire implementation is vectorized, training for all dataset sizes combined take less than half a second.

The performance, as expected, improves as dataset size increases. But even with low number of samples, the implementation results in quite good accuracy of .915.

I also tested the model against scikit-learn's Logistic Regression. When n=2000, my implementation has an accuracy of .930, scikit-learn's has .938.

2 Sparse Logistic Regression: Experiment

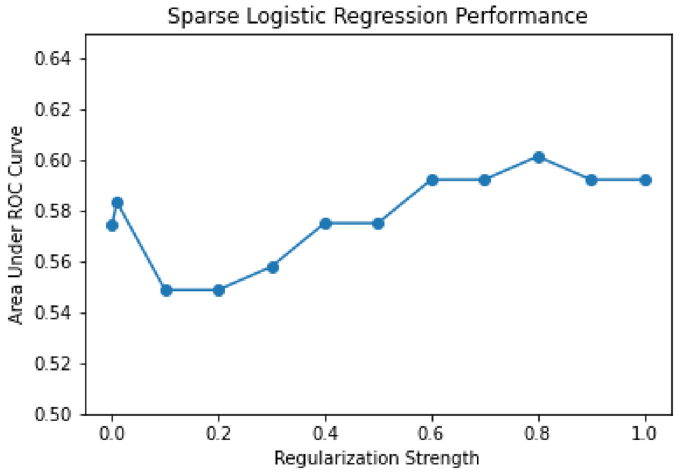

Figure 2: Performance of Sparse Logistic Regression with respect to regularization strength.

The figure above shows how different values of regularization impacts performance of a sparse logistic regression model. Overall there is no clear trend for this dataset. However if we ignore values in smaller extreme, the RoC-AUC values trends higher as regularization strength increases.

(Please note that the Python implementation used (from scikit-learn library) uses a different definition of regularization strength i.e. smaller values indicate higher strength. The Inverse of noted values were therefore used in my implementation.)

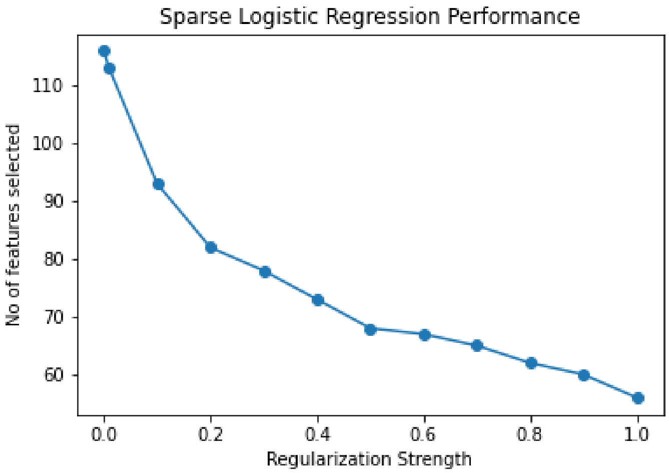

Figure 3: Number of features with non-zero weights with respect to regularization strength.

As expected, higher regularization strength leads to sparser model i.e. number of features which contribute to prediction (i.e. have non-zero weight) trends downwards, a property of L1 regularization.