MiniProject 2: Optimization and Text Classification

Benslimane, Saad; Stappas, Oliver; Yahyaei, Mohsen

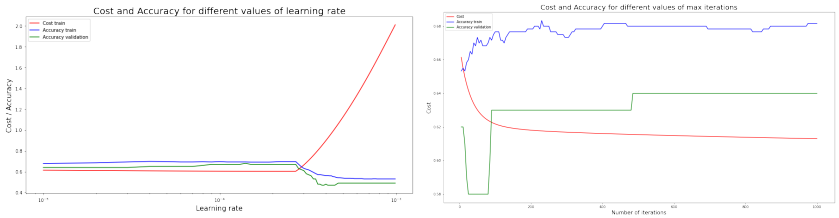
October 27, 2021

Abstract

Logistic regression (LR) is a widely used statistical approach that fits best in binary classification. In this project, we implement LR in the context of "health care" and "text classification". In the first case, the LR use Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and Age of almost 768 persons and their correspond information about diabetic disorder (1 if he/she has the disorder, 0 otherwise) and tested for training dataset. The results show, by reducing learning rate, we can obtain more accurate model. Moreover, by increasing the iteration, we obtain more accuracy at the cost of increasing computational time. Then, we applied mini-batch stochastic gradient descent with different size of mini-batch. The result shows that by increasing the size, the convergence speed increased. Moreover, we analyze effect of momentum coefficient on the performance of LR, the obtained result shows that it can reduce computational cost without significant decreasing of accuracy. In the case of the text classification, the LR uses only the textual contents of articles to determine whether the article was generated by computer or not. By leveraging and integrating various preprocessing techniques like Tokenization and Lemmatization, a highly accurate classifier is developed. Hyperparameter tuning is also crucial in improving the LR's accuracy. The result show that our classifier is able to detect computer-generated articles with 79% accuracy on the test dataset.

1 Introduction

Classification can be performed on a given set of data in numerous contexts. The dependent variable Y is a continuous random variable in linear regression models with one or more independent variables X. The Y, on the other hand, can be qualitative in nature and defined as two (binary) or more categories, implying that it can accept only two or more value (limited to the number of categories). The least square approach does not produce reasonable estimators in this scenario [Atk85]. In such situation, the relationship between a binary Y and X is described and estimated leveraging a logistic regression (LR) model. Logistic regression models are adaptable, have a strong interpretation, and have been used to describe events in a wide range of medical and nonmedical research fields [ZRTP21]. In context of medical application, we consider Diabetes disorder, which is one of the most prevalent human disorders and has become a major public health problem throughout the world. Almost 422 million people worldwide have diabetes and according to WHO it is one of the leading cause of death in the world [Org21]. Scholar developed diabetic disorder prediction model based on machine learning methods and LR [JD21, KA18]. One of major challenge in LR and gradient descent algorithm (GDA) is tune up of parameters such as learning rate and number of training iterations. The present study tries to find effect of these parameters on the performance of GDA based LR and also performance of existing versions of GDA such as stochastic gradient descent algorithm (SGC), mini-batch SGC, momentum based SGC. We also consider fake news classification in the general context of text classification. The binary Y is whether the news is computer generated or not, based solely on the textual contents of the articles (X variable). The present study tries to demonstrate the effect of combining preprocessing techniques with hyperparameter tuning to optimize classification accuracy. The major challenge is to test so many possible combinations of preprocessing techniques and parameter values, without needing large amounts of computational time.



(a) learning rate (b) number of training iterations

Figure 1: Impact of (a) learning rate and (b) number of training iteration on the performance of LR

2 Datasets

In this project, we investigate the application of LR to predict diabetic disorder based on eight attributes including Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and Age. The dataset includes totally 768 observations (600 of them as training data, 100 as validation data and 68 as test data). It is worth noting that this dataset does not require any data cleaning. We also investigate the application of LR to predict whether an article is computer generated or not based solely on its textual contents. The dataset includes 25 000 observations in total (20 000 of them as training data, 2000 as validation data and 3000 as test data). This dataset required preprocessing.

3 Results

3.1 Baseline analysis: impact of learning rate and iterations

Before performing the baseline experiments, we need an accuracy measure or cost function for a given value of weights w and bias b. By using the accuracy or cost function, we can evaluate the LR model's performance. In other words, we determine how accurate the model's predictions are compared to the actual outputs. we consider the cross-entropy cost function as follows:

$$J(w) = \frac{1}{N}\sum_{n=1}^{N} y_n \log\left(1+\exp(-w^T x)\right) + (1-y_n)\log\left(1+\exp(w^T x)\right)$$

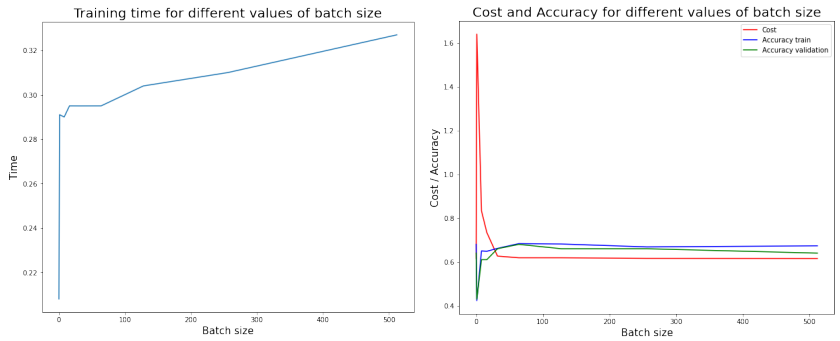
Moreover, we consider *ScikitLearn* package for accuracy evaluation, and it is formalized as below.

$$L_{0/1}(\hat{y}, y) = I(\hat{y} = y)$$

. By defining the cost function, our aim is to minimize it by using gradient descent by discover the best weights and bias. To obtain the impact of learning rate on the cost function and accuracy, we consider interval $[10^{-5}, 10^{-3}]$ as the range of experiments on learning rate. The obtained results have been depicted in Figure 1a. Considering the cost function indicates that the best learning rate is 0.00027. However, using accuracy as a measure, recommend us to select the learning rate as 0.00014. It is notable that there is no significant variation in both cost and accuracy before learning rate equals to 0.00027. This means that learning rate less than this value indicates the convergence of the algorithm. The impact of number of training iterations has been investigated and the obtained result has been reported in Figure 1b. Generally speaking, both accuracy and cost function takes better value by increasing the number of iteration. However, by setting iteration as 200, there would be no significant improvement in accuracy or cost function, and we can conclude that the algorithm is converged.

3.2 SGD and mini-batch

In this section, To analyze impact of batch-size on SGD, different size of batch from 0 to 512 has been considered (batch size equals 0 means fully batched baseline). The depicted results in Figure 2a show that by increasing the batch size, the algorithm runtime increases. However, as depicted in Figure 2b, after initial improvement on accuracy and cost, no significant improvement has been observed (for batch size more than 64).



(a) Time (b) Accuracy and cost

Figure 2: Impact of batch size on (a) algorithm Run-time and (b) cost/accuracy function

(a) Accuracy and Cost (b) Runtime

Figure 3: Impact of momentum coefficient on (a)cost/accuracy function and (b) the algorithm Run-time

3.3 Adding momentum to GDA

To analyze impact of momentum on LR cost and accuracy, the value of momentum coefficient (β) has been changed from 0 to 1. The results are illustrated in Figure 3. The obtained results show that by increasing β, the there exist fluctuation on both cost and accuracy in interval between [0,0.8]. However, after 0.8, both accuracy and cost worsen. From computational time standpoint, the runtime of the algorithm reduces sharply in $\beta = 0.2$.

3.4 Impact of batch size and momentum

Finally, we investigate the effectiveness of LR by considering different batch size under three scenarios of momentum coefficients (0.2,0.5 and 0.8).

The illustrated result in Figure 4 shows that, similar to the result obtained in section 3.2, the cost function and the precision show stability for cases where the batch size is equal to or greater than 64 and the momentum coefficient is small (0.2 and 0.5). Otherwise, the model performance in respect to time and accuracy (or cost function) is not poor on small size of batch (less than 64). In higher value of momentum coefficient (0.8) models start to converge on batch size equals to 256 which obviously increase needs more computational time.

3.5 Techniques Used For Classifying Fake News

Because of there being only one feature for the fake news dataset, and because we were dealing with textual data that had not been preprocessed at all, preprocessing would be a major factor in creating an accurate classifier. A pipeline was used with three major components; a CountVectorizer vectorizer to count the occurrences of words; a TfidfTransformer transformer to get the frequency of these words; a Logistic Regression classifier LogisticRegressionCV, which automatically uses 5-fold

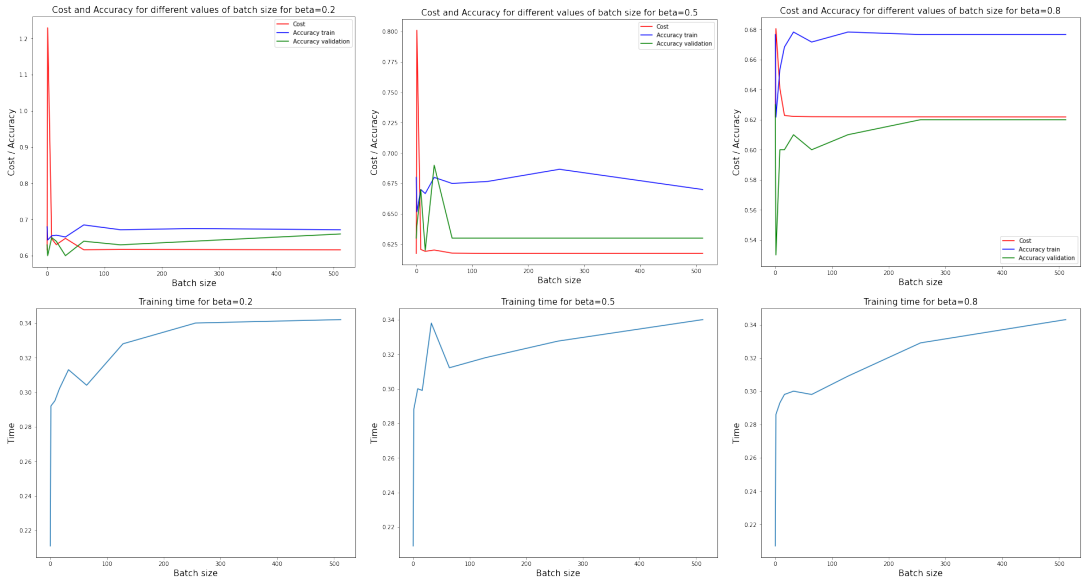


Figure 4: Impact of batch size and momentum coefficient on performance.

cross validation to tune the hyperparameter C (the inverse of regularization strength). To tune the various hyperparameters of our pipeline, we performed a Grid Search (using 5-fold cross validation) of all possible combinations of a few of these parameters to find the choices that maximized accuracy. The data was tokenized to separate different words. Lemmatization was then used, replacing words with a word that is a root of their family of words. We also tried Stemming and removing stop words from the text data, but these did not improve classifier accuracy.

3.6 Fake News Results

The Grid Search gave us optimal hyperparameters. For CountVectorizer, extracting unigrams and bigrams of textual data was optimal. Using no stop words instead of English ones was best, and using a class to lemmatize and tokenize the data was better than not doing so. For the classifier, the l2 penalty was optimal, and then sag solver, due to the large size of our dataset. When then using the classifier to fit on the training data with these parameter values, we observed a training accuracy of 1.0. suggests overfitting, and should ideally be dealt with. After predicting the target values on the test dataset, we got a test accuracy of 0.78733.

4 Discussion and Conclusion

In this project, we present two applications of binary LR in the context of Diabetes prediction and text classification. For the first case, we developed a binary classifier based on 9 attributes. We investigated the impact of several features of sub-gradient based LR on the performance of LR (computational time and accuracy). Based on the case study's experimental findings, it is observed that 1) lower value of learning rate effect can provide more accurate model, 2) batch size can improve the runtime but decrease the accuracy and the best point was 64 where both performance criteria were in acceptable level. Moreover, Momentum coefficient (less than 0.8) can reduce the computational cost without sacrificing accuracy. It is notable that it was difficult to find the best hyperparameters that fit perfectly. For example in the first part, we observed that it was difficult to know the perfect beta value and also sometimes it was not necessary to use a large batch size. So we can conclude that the model converges when it starts to give an acceptable solution for the problem we try to solve, and the most important thing is to maximize the accuracy of the model. For the second case, we developed the LR model to predict news is fake (generated by computer) or not. Our developed model was able to reach at 79% of accuracy level. To improve the accuracy, reducing the overfitting to the training data would likely help.

5 Statement of Contributions

All the group mates contributed equally in the project.

References

[Atk85] Anthony Curtes Atkinson. Plots, transformations and regression; an introduction to graphical methods of diagnostic regression analysis. Technical report, 1985.

[JD21] Ram D Joshi and Chandra K Dhakal. Predicting type 2 diabetes using logistic regression and machine learning approaches. *International Journal of Environmental Research and Public Health*, 18(14):7346, 2021.

[KA18] Pahulpreet Singh Kohli and Shriya Arora. Application of machine learning in disease prediction. In *2018 4th International conference on computing communication and automation (ICCCA)*, pages 1–4. IEEE, 2018.

[Org21] World Health Organization. Diabetes. Technical report, World Health Organization, 2021.

[ZRTP21] Emily C Zabor, Chandana A Reddy, Rahul D Tendulkar, and Sujata Patil. Logistic regression in clinical studies. *International Journal of Radiation Oncology, Biology, Physics*, 2021.