Mini Project Report

Aims and Objectives:

The aim of this project is to implement a logistic regression classifier using Newton's method and apply it on a real data set. The motivation behind using Newton's method is that it usually converges faster than gradient descent when maximizing logistic regression log likelihood. Also, one of the objectives is to show how faster with few iterations function converges. Finally, we have to show the accuracy score of Newton's logistic regression classifier.

Data Cleaning and Preparation:

In this project we have used the Breast Cancer Wisconsin dataset from UCI machine learning repository.

After downloading the dataset from the provided link, I have converted the datafile into .csv file from the .data file extension. The column names are gathered from breast-cancer-wisconsin.names file from the same link. I have imported required python libraries such as numpy, pandas, sklearn.linear, seaborn, matplotlib and statistics to complete this assignment. Then I have loaded the breast-cancer-wisconsin.csv data to jupyter notebook using pandas dataframe. By doing data exploratory analysis I have discovered there are 16 '?' present in 'Bare Nuclei' column. Then I have considered them as null value and replaced them using the mean value of that column. As we have learned from the data description the column 'Class' is being used as labels in this dataset. Since the Class types are Benign(2) and Malignant(4), I have converted them into 0 and 1 which will help to apply the logistic regression. Then I have made a plot to show the count of Class type Benign(0) and Malignant(1) using seaborn library. I have also dropped the 'Sample Code Number' column from the dataset since that is negatively correlated with all the other columns and seems unnecessary.

Splitting Data into Training and Testing sets:

According to the requirement, I have used sklearn library (model_selection import train_test_split) to split the dataset into 80% training and 20% testing.

Model Preparation for Newton's Logistic Regression:

- I have created a sigmoid function using the mathematical formula which is used for logistic regression classification problems to convert the continuous values into discrete values as output.
- As we know, Newton's method is a second-order optimization algorithm which helps in finding the best weights in our logistic function in fewer iterations compared to batch gradient descent.
- It is implemented by using multiple functions in this project such as sigmoid, newton_method, convergence, log_test_model and CrossEntropy.
- The newton_method uses the inverse of hessian matrix which is a k-dimensional matrix function. It is a second order derivative of loss function.
- The convergence function is used for finding the convergence point by using theta_old and theta_new.
- The log_test_model is used for calculating the accuracy of the model based on its probability.

- The CrossEntropy is used to determine if the change in loss is below a certain threshold as the stopping point.
- pv1 and pv2 are the probability variables in the function which are used to calculate the predictive values of output by using gradient and cross entropy. All are based on the mathematical equation to calculate Newton's method.
- The old_theta and new_theta are keeping the track of improving loss in each iteration while converging.
- Although I have set the Maximum iteration (max_iter) as 1000 and total tolerance very small (0.00000001) the loss function has converged very quickly after 10 iterations shown below in the output. The quicker convergence and providing a faster solution by using Newton's method was the main goal behind this implementation which is shown in the output.
- Finally, I have trained my dataset and after few iterations, when we see the results, it starts from less values and gradually the accuracy increases.
- As per the instructions, I have performed ten random data splits using a for loop and the average over these 10 trials is used to estimate the generalization performance. The average of 10 random trials accuracy score of Newton's logistic regression model is **85.85714285714286%**.

The 10 trials of Newtons model Accuracy:

[88.57142857142857, 91.42857142857143, 84.28571428571429, 86.42857142857143, 85.0, 85.0, 83.57142857142857, 82.85714285714286, 87.14285714285714, 84.28571428571429]

Newton's Accuracy: 85.85714285714286

References:

1) https://medium.com/machine-learning-with-python/logistic-regression-implementation-in-python-74321fafa95c
2) https://thelaziestprogrammer.com/sharrington/math-of-machine-learning/solving-logreg-newtons-method
3) https://github.com/jrios6/Math-of-Intelligence/blob/master/2-Second_Order_Optimization/Logistic%20Regression%20with%20Newton's%20Method.ipynb
4) https://machinelearningmastery.com/cross-entropy-for-machine-learning/