Newton's Method

Dataset:

We are using the Breast Cancer Wisconsin dataset from UCI machine learning repository to use Newton's method with the help of a quadratic approximation of our function to identify what would be the next best course of action in terms of cancer research.

The dataset used is **Breast Cancer Wisconsin dataset**.

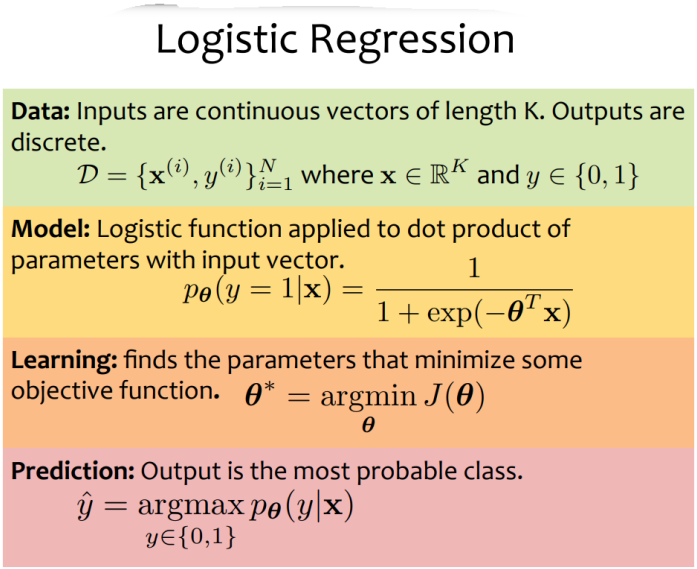
Project Design:

1)We are fetching the data from the given dataset and then we are identifying the data using various columns, we have re classified the class as malignant and benign as (4=1 and2=0) and checked for any null values in the data.

2)when we are identifying the data type in the dataset, we find that there is an "Object" type data that is present in the dataset (? And we are converting that into 0) for calculation and converting the Bare Nuclei into int64 type for identifying the data pattern.

3) we are representing the data on the graph for how many instances are showing the data for malignant and benign cancers and get the counts(458, 241) and based on that we move on to newtons method.

4)Newton's method: in this method we mainly focus our motivation on using a quadratic approximation of a function to make a good guess where we should step next.



Here in the logistic regression approach we first start with the regularization step and then followed by we check for convergence.

It was given in the question to make 10 splits and the average of those 10 splits are to be used to identify the generalization performance.

So I have written functions to check convergence, and function for Test model where we are converting the probability to prediction and the next step would be to hyperparameter tuning and validating the data of the data set.

Here in the test model function, I have considered the tolerance is 0.01 and splitting the data in to 80% for training and 20% and here the max iterations are 20 but we can see that the convergence is happening at 4^{th} iteration and the data is showing a varied accuracy each time the functions are run.

After we run the functions to identify the accuracies of the model we are printing it in the Jupiter notebook file and the value keeps changing each time.

Testing:

We are testing the algorithm with beta values close to the average of the accuracies. Then we get the accuracy of the model. Which happens to be 87.14285714285714.

As shown below:-

The accuracy of the model is: 87.14285714285714

We can further test the data by following the hint given in the question by using the sklearn logistic regression and again splitting the data in to 80% for training and 20% for the data and we consider the same tolerance of 0.01 and the accuracy comes down to 88.23529411764706 ,We then test the data with the logistic regression and we get an accuracy of 94.28571428571428 which is close to newton's method.

Result:

As the validation set is randomized, the accuracies of the model keep changing. But the overall performance depends on the train set and the accuracy for both newton's and sklearn method is very close and they are as follows for this particular instance,

The accuracy of the Newtons model is: **87.14285714285714**

The accuracy using sklearn logistic regression is: **88.23529411764706**