HW 3 - Cancer and Diabeties analisys

ECGR 5105 - Summer 2024

Joshua Ayers

SID: 801083470

Professor: Vinit Katariya

Github: https://github.com/Jayers0/HW3_ECGR5105

Problem 1:

I implmented logistic regression on the diabeties dataset that I had loaded into a pandas dataframe. For this assugnment I shifted from my previous functionall programing paradyme to use classes both because I have not had experience in dedicated OOP in python but also becasue I would like to use the classes created to make a simple ML framework of my oun.

in my log regression class I implemented l2 regularization though I didnt use it for this problem.

Results:
Accuracy: 0.7338
Precision: 0.6250
Recall: 0.6364
F1 Score: 0.6306

Problem 2:

Invoking the class defined in problem 1, I defined 2 models on with L2 regularization and loaded the data. I had issues with getting this to work until I analized the data and found that there 2 issues:
1. The malignat and begine data was defined as a char rather than an int or a bool and thus would cause issues with the classifier
2. There was an extra column defined at the end of the csv that was loading a column of Nan values
To fix these issues I removed the extra column by directly modifying the dataset by removing the extra comma and I used a lambda function to replace M and B with 1 or 0 respectively

Sub-Problem 1 Results:

Accuracy (L2): 0.9912
Precision (L2): 1.0000
Recall (L2): 0.9767
F1 Score (L2): 0.9882

Sub problem 2 Results:
Accuracy (L2): 0.9912
Precision (L2): 1.0000
Recall (L2): 0.9767
F1 Score (L2): 0.9882

In this case I am very suspisious of the performace. In problem 1 the predictive metrics were substantually lower Because there was substantually more data this may allowed closer results. However if this was the case I would be very suprized I belive that there must be some other issue. Reguardless whatever caused this issue obfuscated the performance diffrence between the l2 reularized and the non-regularized models. However I can say that in this case the abuncance of data alowed the model to train to be much more precise.

Problem 3:

In the same way as problem 1 I defined a niave baysian model class this class. Refur to the graphs produced for performance metrics.

Problem 4:

As before I defined a PCA class designed to allow pronciple component analisys. It is notable that this class was much much less performant than the other classes taking over 1 min to train on the data whereas the previous functions took only a second or less to train. This is probably a result of the exponential growth of the permutation of relations as one increases dimentions in PCA.
The perfomrnace results of ths model were quite good. Staying steady at just under 100% percision, accuricy and recall and F1.
Refur to the graphs produced for performance metrics.

Problem 5:

It seems as thought this problem is asking me to repeat problem 2. Log reg beats every other model on the cancer dataset however I think that is an erroniousl result. Otherwise PCA outperforms baysian but not to a degree that would make it more usefull given its high compute load.