ML HW 2 (Yaffa Atkins, Rubiya Tasnim)

Pocket Algorithm for Breast cancer datasets:

Mathematical Representation:

$$\hat{y} = 1 \text{ if } \sum_{i=1}^{n} w_i x_i \geq b$$

$$\hat{y} = 0 \text{ otherwise}$$

1) For t= 0,1,2…… t (iterations number)

2) **a) Method 1: Initialize W** using the np.ones and assign it to w_initial
   b) Method 2: Initialize W using the linear regression, finding WLin
   W_initial = W_Lin = pseudo inverse of X.Y
3) Initialize b=0
4)Calculate y_hat/ y_ predict by multiplying w and x.
   a) Assign 1 if y_predict is greater than b or threshold value
   b) Else 0.
5) check y_predict and target value is same for X, Y. If not the same, assign error values and update w and b using learning rate.

6) continue iterating through steps 4-5 until t times.

Datasets:

Breast cancer datasets with 569 data points and 30 features.

5 sample datasets:

The first sample dataset uses 90% of the whole data set for training and the remaining 10% for testing. The second sample dataset uses 80% for testing and 20% for testing and so on, with the third using 70% and 30%, the fourth 60% and 50%, and the fifth equally split between training and testing. We used the train_test_split function from the Sklearn to split up the data into testing and training.

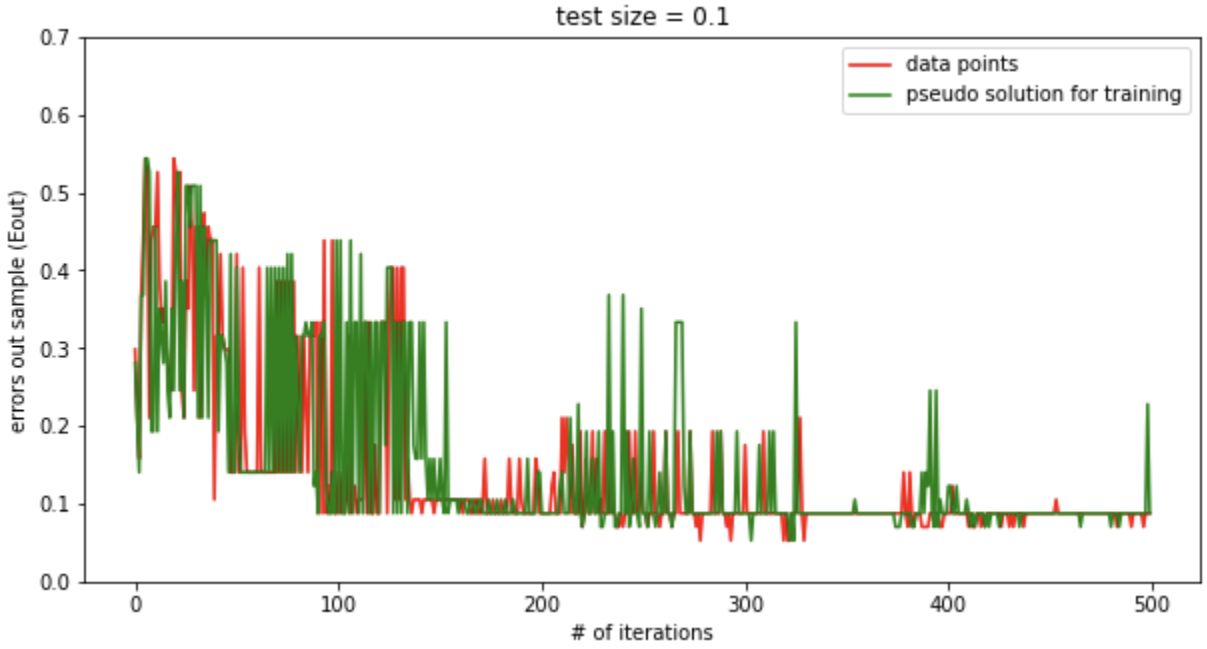
Ein(g) and Eout(g) set up:

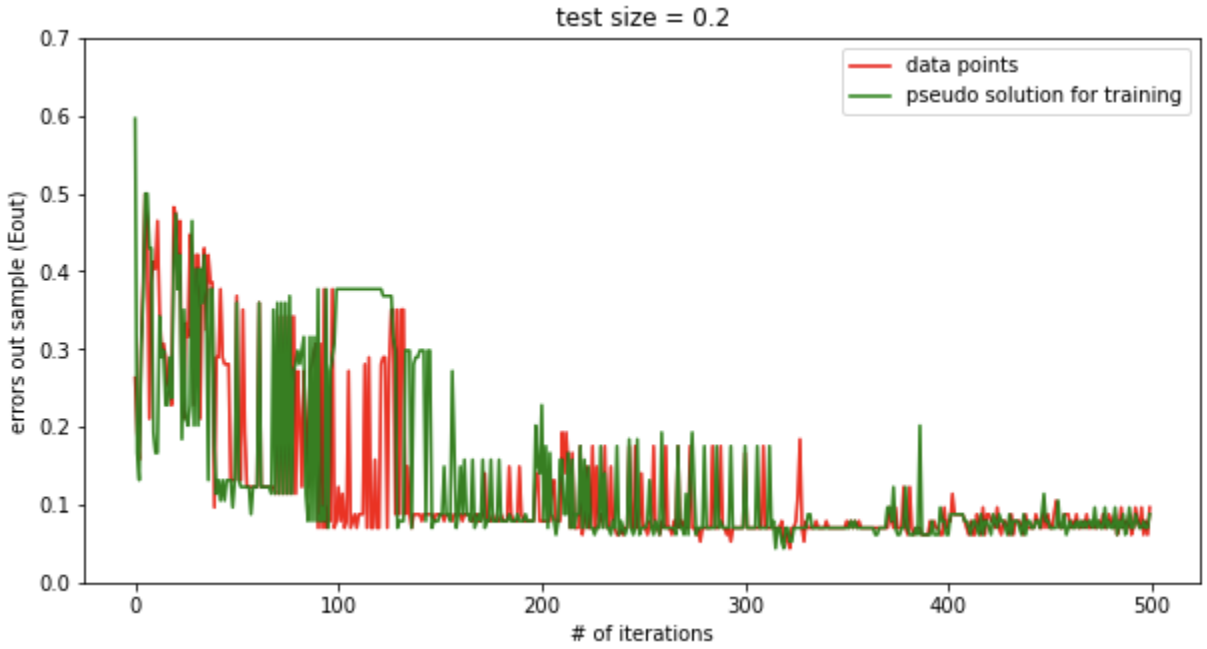
E in (g) is the in-sample error. We set it up using calculating predicted values of y from our model and comparing the target values from the training sets.

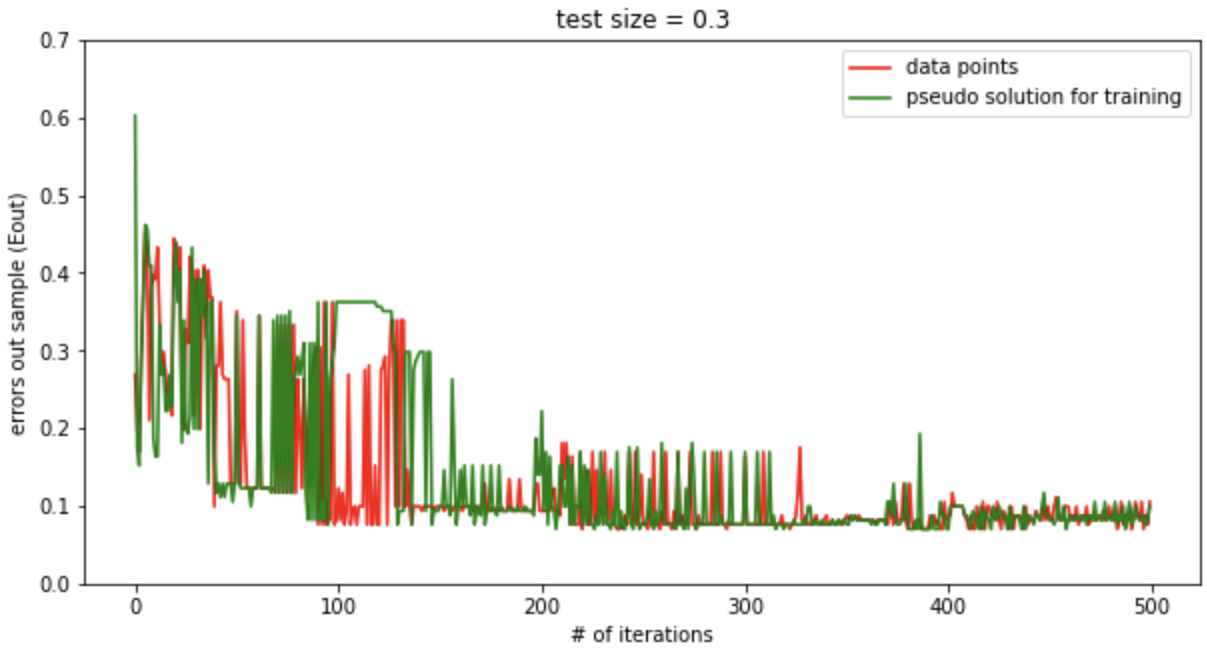
E out (g) is the out-of-sample error. We set it up by coming the best weight from the trained algorithm and using it to predict y. We then compared predicted y with target y to check error.

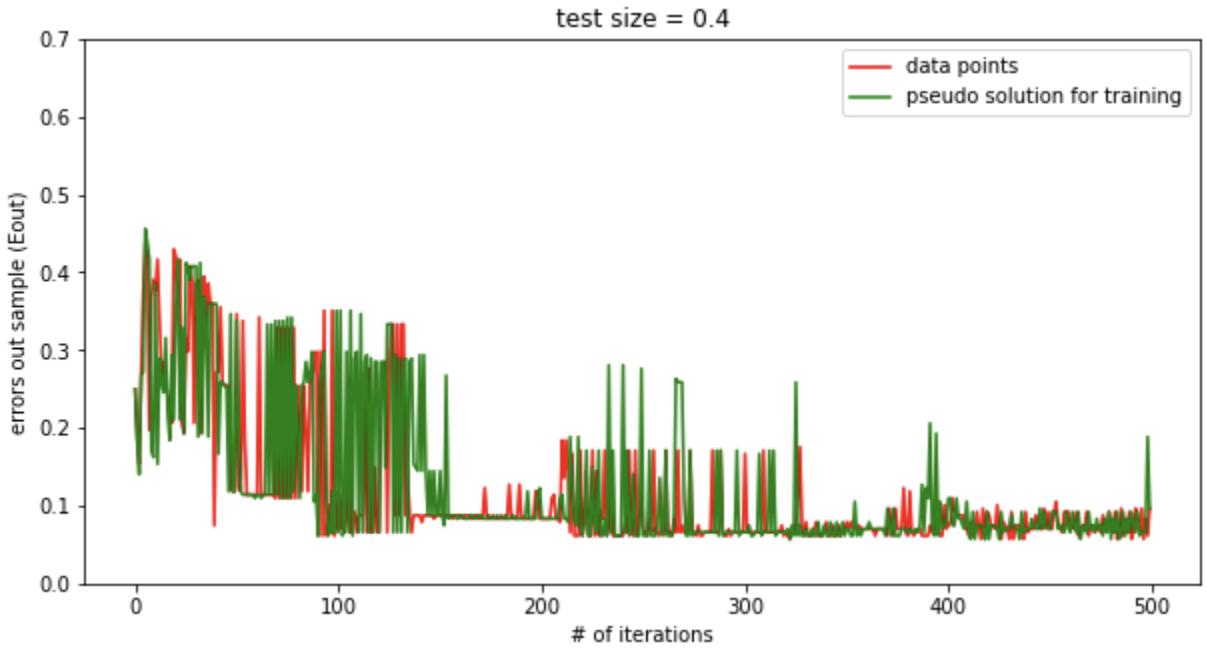
We compute Eout (g) by dividing all misclassifications by total number of classifications.

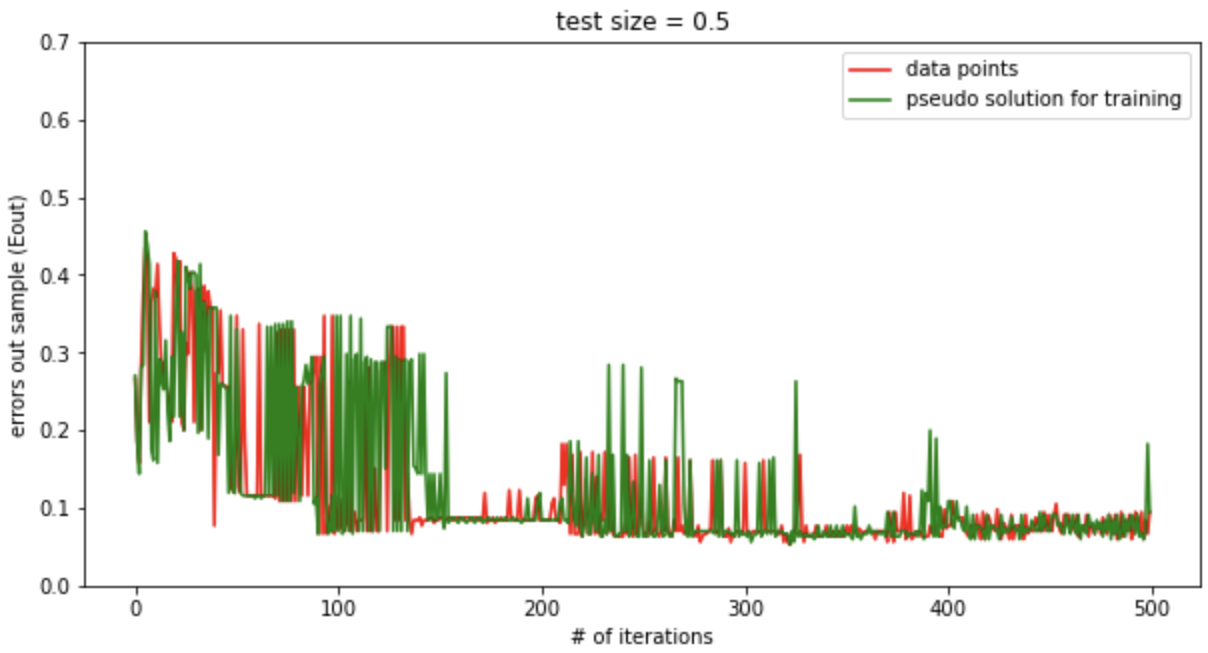
Results:






test size = 0.3
data points
pseudo solution for training
errors out sample (Eout)
# of iterations
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
100
200
300
400
500


test size = 0.4
data points
pseudo solution for training
errors out sample (Eout)
# of iterations
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
100
200
300
400
500



Discussion: from the plots, it seems that the two methods of initializing weight work the same way, one method does not predict less errors than the other method. At the beginning of the iterations, E out decreases over iterations, and plateaus , but eventually with a lot of iterations, there seems to be a lot of noise. This could happen due to overfitting.

Additionally, using 0.1 test size versus using 0.5 test size for E out(g) when splitting the data does not change the amount of errors predicted over the iterations.