Anthony Tummillo

HW Report #11

Problem 1.

Part a.)

```
Dimension #48   Fisher score: 3.191807e-01
Dimension #25   Fisher score: 2.140076e-01
Dimension #21   Fisher score: 1.909700e-01
Dimension #70   Fisher score: 1.892138e-01
Dimension #65   Fisher score: 1.693200e-01
Dimension #40   Fisher score: 1.673448e-01
Dimension #29   Fisher score: 1.650451e-01
Dimension #19   Fisher score: 1.401964e-01
Dimension #57   Fisher score: 1.254536e-01
Dimension #20   Fisher score: 1.212082e-01
Dimension #24   Fisher score: 9.952481e-02
Dimension #30   Fisher score: 9.502421e-02
Dimension #12   Fisher score: 8.581794e-02
Dimension #47   Fisher score: 8.464365e-02
Dimension #61   Fisher score: 6.067734e-02
Dimension #10   Fisher score: 5.791594e-02
Dimension #34   Fisher score: 5.267844e-02
Dimension #27   Fisher score: 4.621594e-02
Dimension #39   Fisher score: 4.611364e-02
Dimension #41   Fisher score: 4.218517e-02
```

Part b.)

```
Dimension #29   Fisher score: 5.966651e-01
Dimension #48   Fisher score: 5.787858e-01
Dimension #25   Fisher score: 5.563977e-01
Dimension #21   Fisher score: 5.496346e-01
Dimension #10   Fisher score: 5.447761e-01
Dimension #19   Fisher score: 5.418221e-01
Dimension #47   Fisher score: 5.418221e-01
Dimension #57   Fisher score: 5.298507e-01
Dimension #61   Fisher score: 5.246424e-01
Dimension #30   Fisher score: 5.216884e-01
Dimension #70   Fisher score: 5.216884e-01
Dimension #2   Fisher score: 5.149254e-01
Dimension #17   Fisher score: 5.149254e-01
Dimension #33   Fisher score: 5.149254e-01
Dimension #65   Fisher score: 5.142257e-01
Dimension #3   Fisher score: 5.123212e-01
Dimension #32   Fisher score: 5.123212e-01
Dimension #60   Fisher score: 5.123212e-01
Dimension #9   Fisher score: 5.097170e-01
Dimension #15   Fisher score: 5.074627e-01
```

The lists of Fisher scores from Part a.) and AUROC scores from Part b.) are not identical, however, they are noticeably similar. The two lists share most of the same dimensions, albeit not in the same order.

These results are what I would expect to find. Given that a Fisher score represents the amount of variance present for a dimension while here the AUROC score measures the predictive power (tested performance) of a dimension when used for classification (here a logistic regression model was used for classification). I assumed that the lists would be similar but not identical. This is because variance and predictive power, despite tending to go hand in hand, are not synonymous, and while they both tend to indicate the relative importance of a dimension for classification they are separate characteristics and thus cannot be assumed to produce identical lists.

Problem 2.

Part a.)

Written and submitted

Part b.)

Written and submitted

The logistic regression model I will be using for this experiment will be implemented using the fitglm function which is provided by Matlab (https://www.mathworks.com/help/stats/fitglm.html). My implementation can be viewed in the log_regression.m function included in my submission. I will be calling this function with the arguments fitglm(X, y, ‘Distribution’, ‘binomial’). This produces a logistic regression model trained using the passed in X and y data which assumes a binomial distribution and will therefore use the logit function to transform the output of the regression to fall within the range [0, 1]. I will take the output of this model and classify the result as 1 if the result is greater than or equal to 0.5, or 0 if the result is less than 0.5.

Part c.)

Wrapper dimensions = [29, 47, 12, 1, 70, 5]

```
wrapper test confusion matrix =

    6    4
   11   65


wrapper test error =

   0.1744


full test confusion matrix =

    7   18
   10   51


full test error =

   0.3256
```

My first observation is obviously the significant difference in the test errors between the two models. I had originally thought that there could be a slight improvement in the test error of the model using the wrapper function dimensions over that of the full model, but I did not expect the error to nearly be cut in half. I must assume that this is because a lot of the dimensions in the full data set are extraneous and lead to overfitting of the data in the full model.

Part d.)

Wrapper dimensions = [29, 47, 12, 1, 70, 5]
AUROC dimensions = [29, 48, 25, 21, 10, 19] (top k=6 from Problem 1 Part b.))

```
wrapper test confusion matrix =

    6    4
   11   65
```

```
wrapper test error =

   0.1744


AUROC test confusion matrix =

    5     1
   12    68


AUROC test error =

   0.1512
```

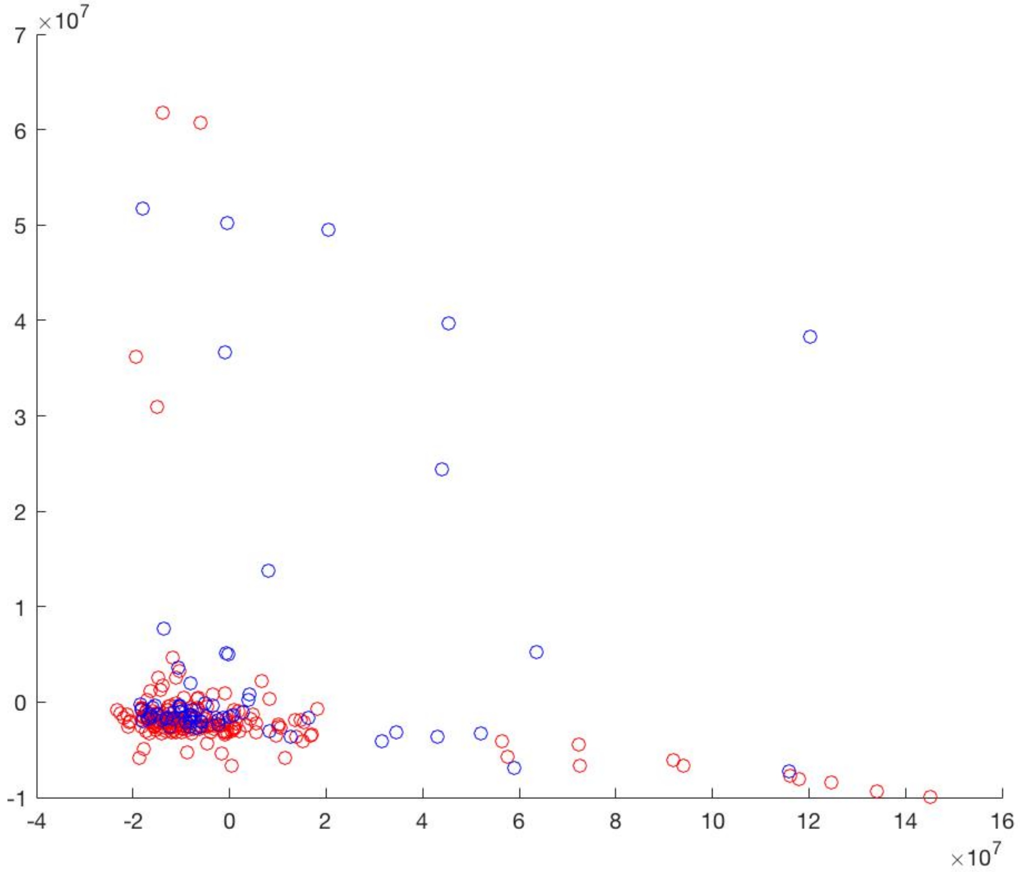
Again, I am quite surprised at the improvement in performance in terms of test error for both the wrapper model and the AUROC model. I am even more surprised that the AUROC model performed better than the wrapper model. I had assumed the AUROC model's test error would have been lower than the test error for the full model, but would still be higher than that of the wrapper model. This is simply because of the high amount of computational work that had to be done to run the wrapper_function and determine the set of dimensions to include. I had assumed that something as quick and simple as testing the AUROC score of each dimension individually and taking the top k dimensions was too simple an approach and would introduce too much bias to outperform the wrapper function, but clearly in this instance I was wrong.

Problem 3.

Part a.)

```
Principal Component #1   Eigenvalue:8.286545e+14
Principal Component #2   Eigenvalue:9.526692e+13
Principal Component #3   Eigenvalue:3.377121e+13
Principal Component #4   Eigenvalue:1.089127e+12
Principal Component #5   Eigenvalue:9.967471e+10
Principal Component #6   Eigenvalue:3.751389e+10
Principal Component #7   Eigenvalue:2.212180e+10
Principal Component #8   Eigenvalue:9.089330e+09
Principal Component #9   Eigenvalue:7.511962e+09
Principal Component #10   Eigenvalue:3.342770e+09
```

Part b.)



Based on my visual analysis I would argue that this new two-dimensional representation does not help us to discriminate the two classes well. This is because it can be very clearly seen in the figure above that most points in both classes occupy the same small cluster of the two-dimensional input space (the lower left corner of the figure above). I would assume that if this two-dimensional representation helped us to discriminate the two classes well that most of the points in each class would occupy visually distinct regions of the two-dimensional input space.

Part c.)

```
PCA test confusion matrix =

    5    11
   12    58

PCA test error =

   0.2674
```

Once again, these results mostly surprised me in comparison to the wrapper and AUROC results, but not the full dimension results. It was reasonable to assume the PCA model would perform better than the model using the full dimensions, and this was indeed the case.

What surprised me is the comparison of the results of the PCA model to those obtained from the wrapper model and AUROC model. I assumed that because transforming the data based solely on the most influential principal components seemed to be such a sound idea theoretically that this method would yield results fairly like those produced by the wrapper and AUROC models. What I found though was that the PCA model performed considerably worse than the wrapper or AUROC models (about 0.1 higher test error than wrapper and AUROC results).

I thought that the differences between the wrapper and AUROC models and the PCA model could be because the wrapper and AUROC models had one additional input dimension than the PCA model did, so I reran this part using the top 6 principal components. While the results for the PCA model improved, they were still far worse than those obtained by the wrapper and AUROC models.