Programming Assignment 1

Rachit Garg CS14B050

1 Synthetic Data Set Creation

- Data sets to be tested were created and named as DS1.
- Centroids of the classes were close enough so that there is overlap in the classes.
- Care was taken to ensure covariance matrix is not a diagonal matrix.
- Data set is stored in "DS1.csv", target values are stored in "DS1_labels.csv".

2 Linear Classification

The following observations were made

- As the centroids of the data go farther, linear regression performs better.
- Coefficients, accuracy, precision, recall, f-measure were noted.

```
Coefficients for linear regression are:
[ 0.05464495  0.2800748   0.28556224  0.12381125  0.25607031  0.08711873
  0.08981731 -0.29246815 -0.02688647 -0.12842453  0.09580529  0.25726691
 -0.35102378  0.20864582  0.15465231  0.00576786 -0.13859495 -0.31422881
 -0.12984605 -0.08715036]
accuracy precision recall  f-measure
0.818333 0.830450 0.800000 0.814941
```

Figure 1: Value of vector β and other measures

3 k-NN classifier

The following observations were made

- As the centroids of the data go farther, k-NN Classifier performs better.
- Linear regression on indicator variables gives a better accuracy, precision, recall and f-measure than the k-NN classifier.
- General trend observed in k values for this data set is that it has local peaks of max value and as k increases more and more the values of all the measures start repeating over a small range and also are worse off from the max peak.
- Coefficients, accuracy, precision, recall, f-measure were noted.

```
accuracy precision recall  f-measure
0.818333 0.830450 0.800000 0.814941
with k = 1, 0.662500 0.669565 0.641667 0.655319
with k = 2, 0.652500 0.737662 0.473333 0.576650
with k = 3, 0.711667 0.712375 0.710000 0.711185
with k = 4, 0.704167 0.758985 0.598333 0.669152
with k = 5, 0.717500 0.714992 0.723333 0.719138
with k = 6, 0.712500 0.745665 0.645000 0.691689
with k = 7, 0.725000 0.723510 0.728333 0.725914
with k = 8, 0.720000 0.745353 0.668333 0.704745
with k = 9, 0.727500 0.726368 0.730000 0.728180
with k = 10, 0.729167 0.755102 0.678333 0.714662
with k = 11, 0.737500 0.738693 0.735000 0.736842
with k = 12, 0.730833 0.746004 0.700000 0.722270
with k = 13, 0.740833 0.741235 0.740000 0.740617
with k = 14, 0.737500 0.753108 0.706667 0.729149
with k = 15, 0.738333 0.735197 0.745000 0.740066
with k = 16, 0.734167 0.746924 0.708333 0.727117
with k = 17, 0.749167 0.745484 0.756667 0.751034
with k = 18, 0.741667 0.750865 0.723333 0.736842
with k = 19, 0.744167 0.742149 0.748333 0.745228
with k = 20, 0.740000 0.749135 0.721667 0.735144
with k = 21, 0.751667 0.745928 0.763333 0.754530
with k = 22, 0.748333 0.754266 0.736667 0.745363
with k = 23, 0.750833 0.744715 0.763333 0.753909
with k = 24, 0.751667 0.753356 0.748333 0.750836
with k = 25, 0.750000 0.744300 0.761667 0.752883
with k = 26, 0.747500 0.752122 0.738333 0.745164
with k = 27, 0.750833 0.746318 0.760000 0.753097
```

Figure 2: First row indicates measures in linear regression, all other rows are from k-NN classifier and indicate other measures based on value of k

4 Data Imputation

The following observations were made

- The data was completed using the mean of the set and stored in Q4data.csv, where last row is dependent variable and first 4 rows were discarded as they were not used for prediction.
- The data was completed using the mean, median, and also a random value between mean-std dev and mean+ std dev. Average residual error on regression on median and on random values was worse off in sme cases and better in some.

5 Linear Regression and Regularized Linear Regression

The following tasks were done:

- The data was stored according to the guidelines mentioned.
- Coefficients are stored in a file called Q5coeff.csv for coefficients for linear regression for the 5 cases. And for regularized linear regression, stored in a file called Q6coeff.csv.
- We observe that the error reduces significantly on doing regularized regression. The features whose coefficients are close to zero should not be picked and hence feature selection can be done.
- Value of λ observed is close to 0, close to 1, close to 5 or large value depending on the data split. Mostly close to 1 or close to extremes(i.e 0 or large value) are observed.

```
The residual error on best fit is 0.018291
The residual error on best fit is 0.017169
The residual error on best fit is 0.019278
The residual error on best fit is 0.020410
The residual error on best fit is 0.019989
The average residual error on best fit over 5 splits is 0.019027
split number 0:
with alpha = 5.100000, error is 0.017471
split number 1:
with alpha = 5.500000, error is 0.016663
split number 2:
with alpha = 0.300000, error is 0.018987
split number 3:
with alpha = 4.400000, error is 0.019475
split number 4:
with alpha = 9.900000, error is 0.018956
The average residual error on doing l1 regularized linear regression over 5 splits is 0.018310
```

Figure 3: Values of error for linear regression and l1 regularized linear regression

6 Logistic Regression

The following observations were made

- For the data set - *Accuracy* : 0.825000, *Precision* : 0.760000, *recall* : 0.950000, $f - measure$: 0.844444
- Performance results by professor boyd's code are *Accuracy* : 0.975000, *Precision* : 0.952381, *recall* : 1.000000, $f - measure$: 0.975610.
- L1 regularized logistic regression performs better.

```
Reading data...

Problem summary:
    [feature matrix]        dense matrix of (40 examples x 96 features)
    [mode]                  classify and test

Running classifier...

Classification result:
    [right predic. count]   39
    [wrong predic. count]   1
    [test error]            1 / 40 = 0.025

Timing:
    [read data]            0.00206 (sec)
    [solve problem]        7.9e-05 (sec)
    [write solution]      0.000377 (sec)
    [total time]           0.00252 (sec)
```

Figure 4: Performance results by professor boyd's code.

7 LCA Analysis and LDA Analysis

The following observations were made

- Performance results by LCA are $Accuracy$: 0.5, $Precision$: 0.5, $recall$: 1.000000, $f - measure$: 0.6666667.
- Performance results by LDA are $Accuracy$: 1.00000, $Precision$: 1.00000, $recall$: 1.00000, $f - measure$: 1.00000.
- We infer that LDA performs much better than LCA

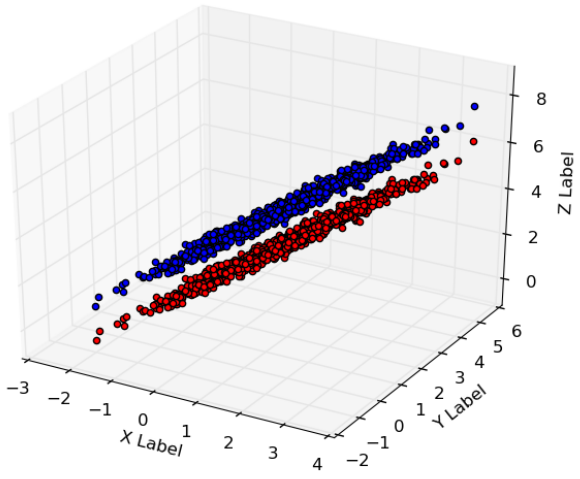

Figure 5: 3-D plot of dataset