2019W-T3 BDM 3014 - Introduction to Artificial Intelligence 01

Lab 01 – Logistic Regression

Lucicarla Silva de Santana	C0724974
Valeria Ferreira de Almada Nobrega	C0724858
Rafael Andrade Da Conceicao	C0725132
Tulio Fernandes	C0722179
Javier Navarro Gonzalez	C0725418

Logistic regression - Python output

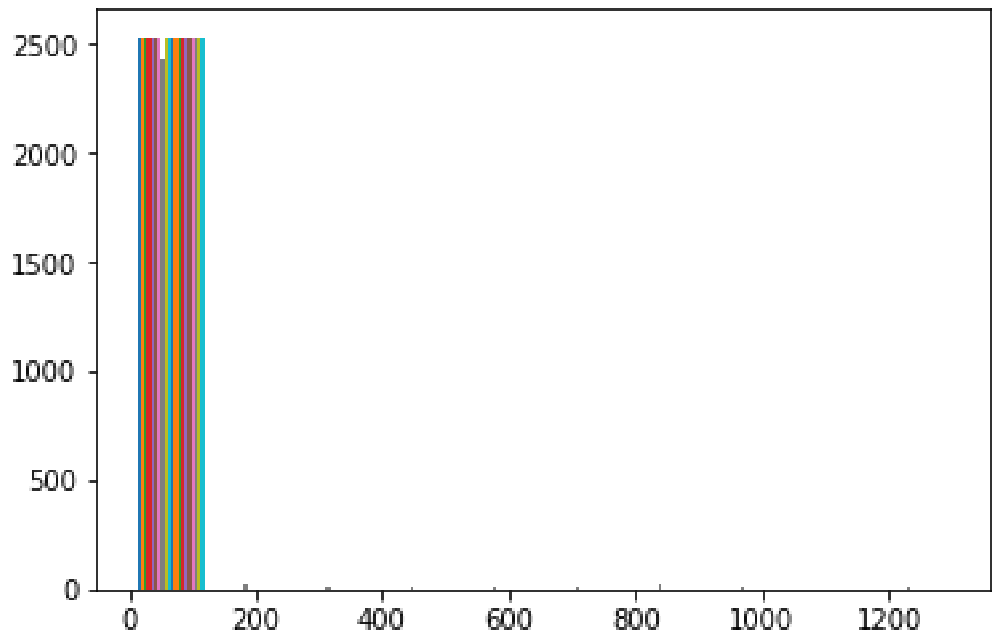
1. We will use acoustic features to distinguish a male voice from female. Load the dataset from “voice.csv”, identify the target variable and do a one-hot encoding for the same. Split the dataset in train-test with 20% of the data kept aside for testing.

```
In [8]: print (dataset_voice)
    meanfreq        sd    median  ...    dfrange   modindx  label
0   0.059781  0.064241  0.032027  ...   0.000000  0.000000   male
1   0.066009  0.067310  0.040229  ...   0.046875  0.052632   male
2   0.077316  0.083829  0.036718  ...   0.007812  0.046512   male
3   0.151228  0.072111  0.158011  ...   0.554688  0.247119   male
4   0.135120  0.079146  0.124656  ...   5.476562  0.208274   male
5   0.132786  0.079557  0.119090  ...   2.718750  0.125160   male
6   0.150762  0.074463  0.160106  ...   5.304688  0.123992   male
7   0.160514  0.076767  0.144337  ...   0.531250  0.283937   male
8   0.142239  0.078018  0.138587  ...   2.156250  0.148272   male
9   0.134329  0.080350  0.121451  ...   4.679688  0.089920   male
10  0.157021  0.071943  0.168160  ...   2.804688  0.200000   male
11  0.138551  0.077054  0.127527  ...   2.710938  0.132351   male
12  0.137343  0.080877  0.124263  ...   5.000000  0.088500   male
13  0.181225  0.060042  0.190953  ...   2.796875  0.416550   male
14  0.183115  0.066982  0.191233  ...   6.539062  0.139332   male
15  0.174272  0.069411  0.190874  ...   6.992188  0.209311   male
16  0.190846  0.065790  0.207951  ...   6.312500  0.254780   male
17  0.171247  0.074872  0.152807  ...   0.562500  0.138355   male
18  0.168346  0.074121  0.145618  ...   3.117188  0.059537   male
19  0.173631  0.073352  0.153569  ...   2.812500  0.068124   male
20  0.172754  0.076903  0.177736  ...   0.710938  0.235069   male
21  0.181015  0.074369  0.169299  ...   3.687500  0.059940   male
22  0.163536  0.072449  0.145543  ...   0.437500  0.091699   male
23  0.170213  0.075105  0.146053  ...   0.554688  0.161791   male
24  0.160422  0.076615  0.144824  ...   3.945312  0.073890   male
25  0.164700  0.075362  0.147018  ...   1.054688  0.125926   male
```

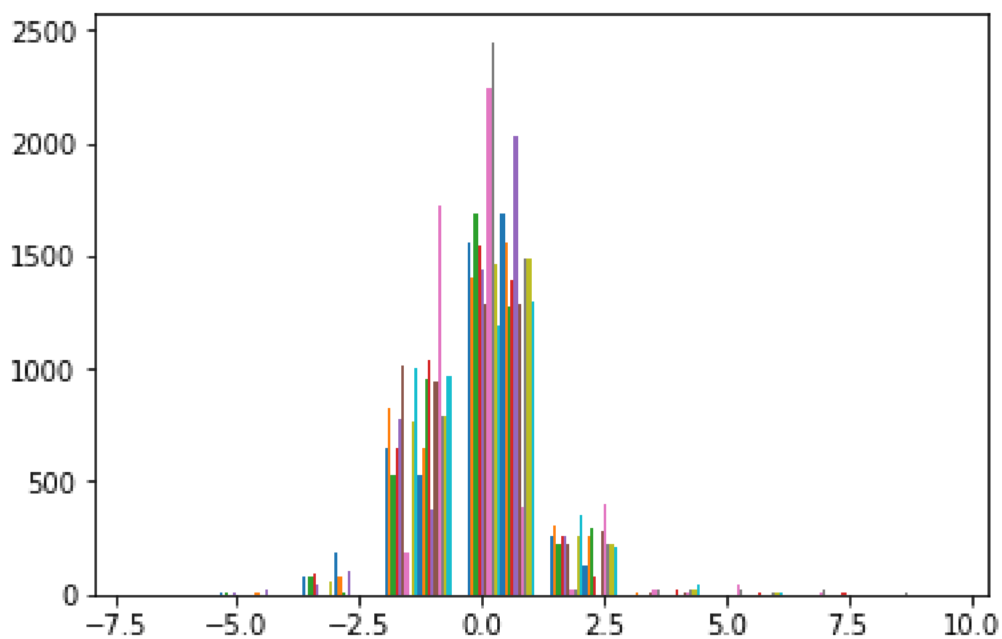
Voice data set loaded

```
In [31]: genre_count = dataset_voice['label'].value_counts()

In [32]: print(genre_count)
male      1584
female    1584
Name: label, dtype: int64

In [33]:
```

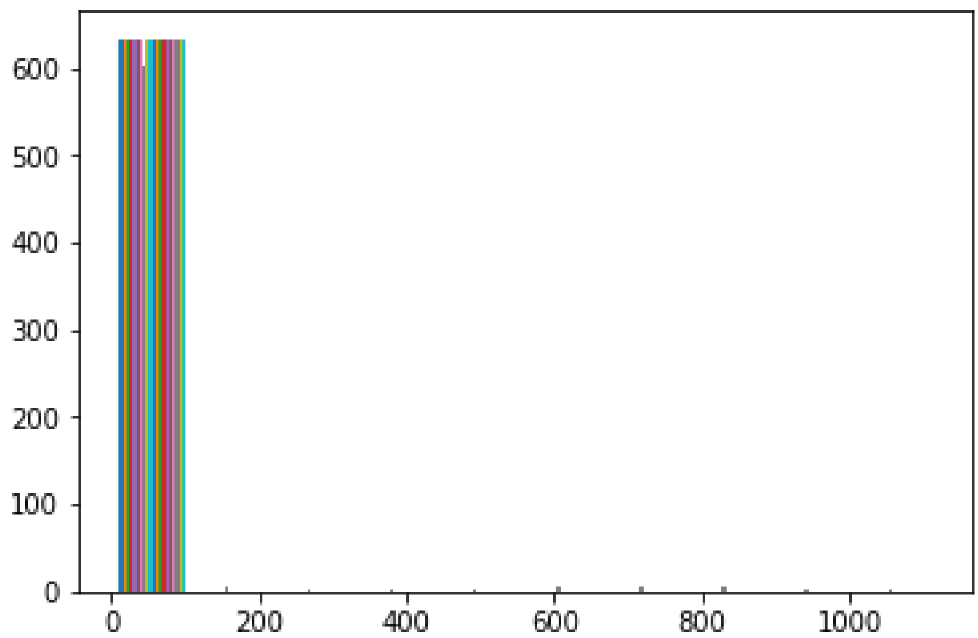
Count of male and female



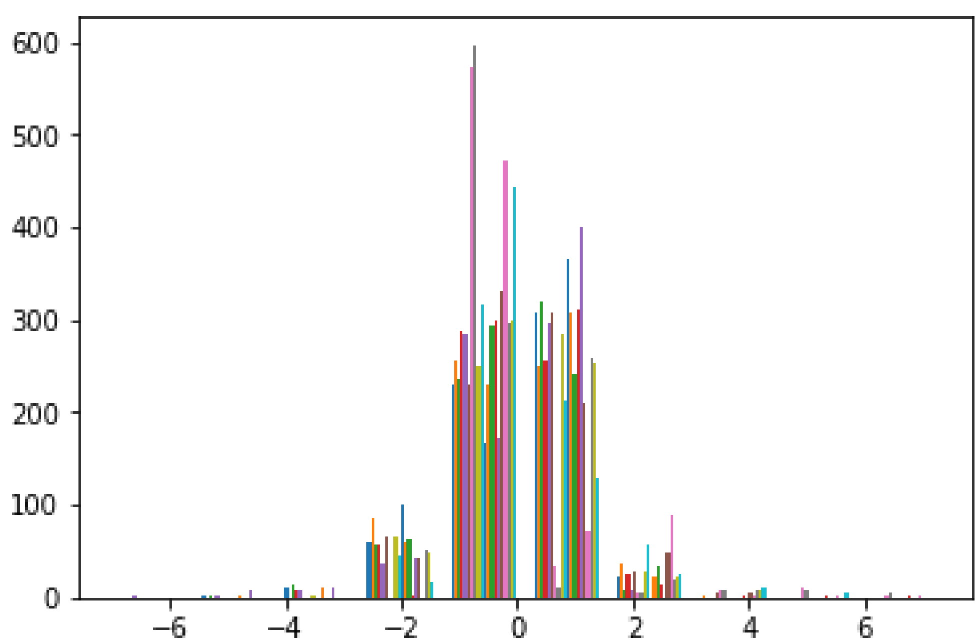
X_train before scaling



X_train after scaling



X_test before scaling



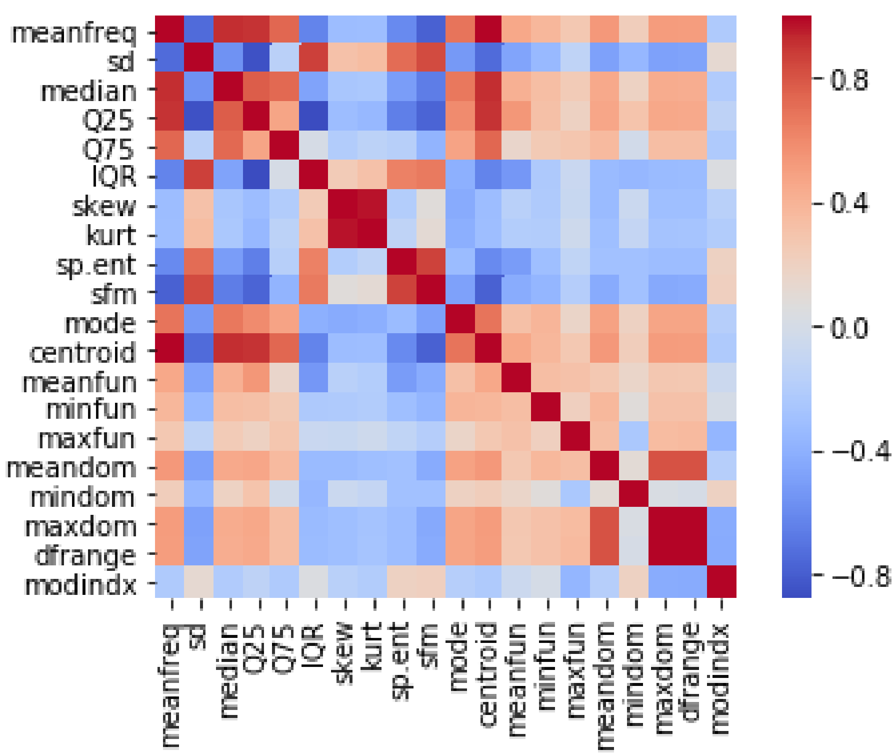
X_test after scaling

2. Fit a logistic regression model and measure the accuracy on the test set.

```
Predicted    0    1  All
True
0          293    8  301
1            5  328  333
All        298  336  634

In [68]: print(metrics.accuracy_s
0.9794952681388013
```

Results of prediction using all the columns and accuracy

3. Compute the correlation matrix that describes the dependence between all predictors and identify the predictors that are highly correlated. Plot the correlation matrix using seaborn heatmap.



Heatmap of correlation

4. Based on correlation remove those predictors that are correlated and fit a logistic regression model again and compare the accuracy with that of previous model.

```
Predicted    0    1  All
True
0          294    7  301
1            7  326  333
All        301  333  634

In [95]: print(metrics.accuracy_score(y_test, y_pred))
0.9779179810725552
```

Confusion matrix and accuracy using 5 main features determined by RFE

Logistic regression - R output

1. We will use acoustic features to distinguish a male voice from female. Load the dataset from "voice.csv", identify the target variable and do a one-hot encoding for the same. Split the dataset in train-test with 20% of the data kept aside for testing.

```
> head(dataset_voice)
    meanfreq         sd     median         Q25        Q75        IQR      skew        kurt    sp.ent
1 0.05978098 0.06424127 0.03202691 0.015071489 0.09019344 0.07512195 12.863462  274.402906 0.8933694
2 0.06600874 0.06731003 0.04022873 0.019413867 0.09266619 0.07325232 22.423285  634.613855 0.8921932
3 0.07731550 0.08382942 0.03671846 0.008701057 0.13190802 0.12320696 30.757155 1024.927705 0.8463891
4 0.15122809 0.07211059 0.15801119 0.096581728 0.20795525 0.11137352  1.232831    4.177296 0.9633225
5 0.13512039 0.07914610 0.12465623 0.078720218 0.20604493 0.12732471  1.101174    4.333713 0.9719551
6 0.13278641 0.07955687 0.11908985 0.067957993 0.20959160 0.14163361  1.932562    8.308895 0.9631813
        sfm       mode   centroid    meanfun     minfun    maxfun     meandom    mindom    maxdom
1 0.4919178 0.00000000 0.05978098 0.08427911 0.01570167 0.2758621 0.007812500 0.0078125 0.0078125
2 0.5137238 0.00000000 0.06600874 0.10793655 0.01582591 0.2500000 0.009014423 0.0078125 0.0546875
3 0.4789050 0.00000000 0.07731550 0.09870626 0.01565558 0.2711864 0.007990057 0.0078125 0.0156250
4 0.7272318 0.08387819 0.15122809 0.08896485 0.01779755 0.2500000 0.201497396 0.0078125 0.5625000
5 0.7835681 0.10426140 0.13512039 0.10639784 0.01693122 0.2666667 0.712812500 0.0078125 5.4843750
6 0.7383070 0.11255543 0.13278641 0.11013192 0.01711230 0.2539683 0.298221983 0.0078125 2.7265625
    dfrange    modindx label
1 0.0000000 0.00000000  male
2 0.0468750 0.05263158  male
3 0.0078125 0.04651163  male
4 0.5546875 0.24711908  male
5 5.4765625 0.20827389  male
6 2.7187500 0.12515964  male
```

Voice data set loaded

```
> nrow(s_train)
[1] 2414
> nrow(s_test)
[1] 754
> 
```

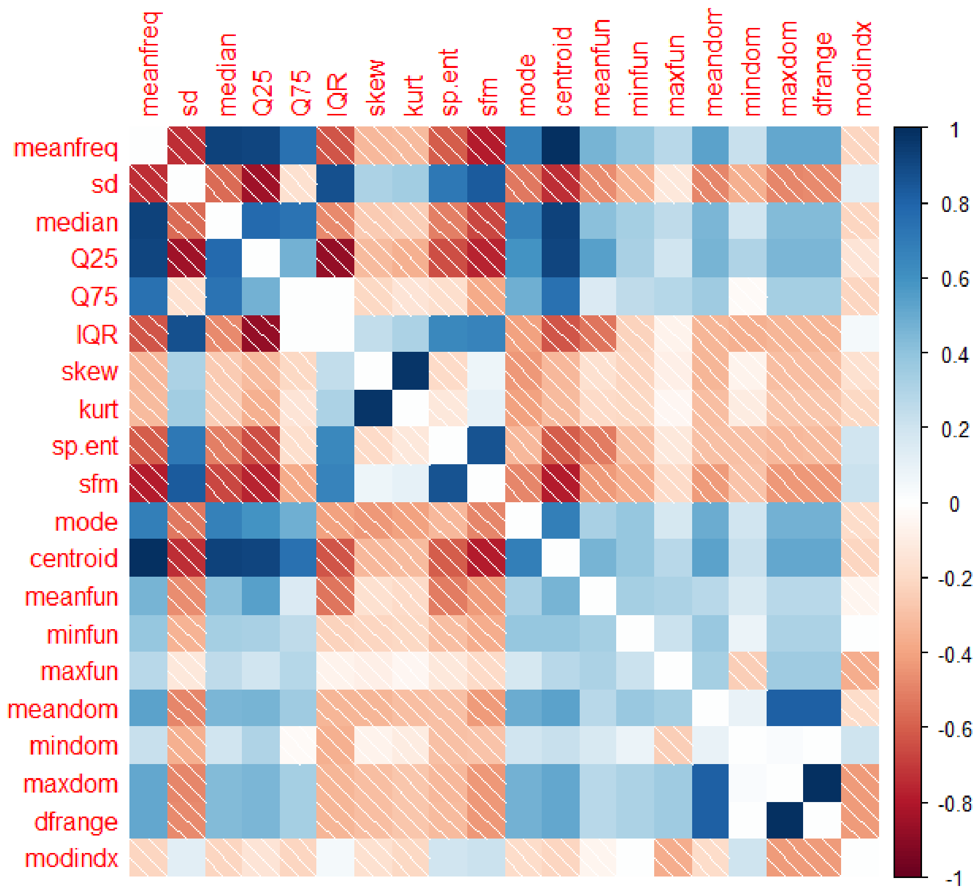
Split the dataset in train-test

2. Fit a logistic regression model and measure the accuracy on the test set.

```
> con_matrix_full
            Predicted_value
Actual_Value FALSE TRUE
       FALSE   369    9
       TRUE      8  369
> # 4.3 Accuracy
> ## Acc. 96% -> All columns
> (con_matrix_full[[1,1]] + con_matrix_full[[2,2]]) / sum(con_matrix_full)
[1] 0.9774834
```

Results of prediction using all the columns and accuracy

3. Compute the correlation matrix that describes the dependence between all predictors and identify the predictors that are highly correlated. Plot the correlation matrix using seaborn heatmap.



Correlation (corrplot)

4. Based on correlation remove those predictors that are correlated and fit a logistic regression model again and compare the accuracy with that of previous model.

```
> con_matrix_select <- table(Actual_Value=s_test$label, Predicted_value = logisticRes_select > 0.5)
> con_matrix_select
            Predicted_value
Actual_Value FALSE TRUE
       FALSE   366   12
       TRUE      8  369
> ## For model using columns from the dataset manually selected by their correlation score
> con_matrix_manual <- table(Actual_Value=s_test$label, Predicted_value = logisticRes_manual > 0.5)
> con_matrix_manual
            Predicted_value
Actual_Value FALSE TRUE
       FALSE   365   13
       TRUE      6  371
> ## Acc. 96% -> Columns automatically selected
> (con_matrix_select[[1,1]] + con_matrix_select[[2,2]]) / sum(con_matrix_select)
[1] 0.9735099
> ## Acc. 97% -> Columns manually selected
> (con_matrix_manual[[1,1]] + con_matrix_manual[[2,2]]) / sum(con_matrix_manual)
[1] 0.9748344
```

Confusion matrix and accuracy