McGill University

COMP 551 - Applied Machine Learning

Assignment 2 Report

Written by:
Asher Wright - 260559393

Due to:
Prof. Chandar

February 12, 2018

Linear Classification and Nearest Neighbour Classification

1. Dataset 1

The first dataset, DS1, was created as indicated in the handout. Please note that I decided to split up the data into different text files. I first split it into testing and training (as was given in assignment 1), and then I further split it up into positive and negative classes. Another, more scalable, option would be to add a column to the data called "label", but with two classes I found this method easier.

2. LDA with Dataset 1

Below are the computed performance measures for dataset 1 with LDA classification.

Best fit accuracy	95.08 %
Precision	95.77 %
Recall	94.33 %
F-measure	95.05 %

The coefficients learnt were the following:
w = [-14.70625916, 8.72333061, 6.06823166, 3.39412552, 10.23338177, 4.19114737, -17.87277904, 24.77351351, 30.34024193, -9.18815402, 13.27738285, 12.89340784, -16.32857775, -13.64881951, 5.89264418, -13.54701867, -30.67490656, 6.98491646, 0.99426985, 5.20896759]
w_0 = -28.0514877938

3. k-NN with Dataset 1

Figure 1 shows how the F_1 measure changed as k was increased from 1 to 150. I only chose odd numbers of k to prevent having to deal with ties. As it would be messy to list the performance for every 75 k value, I instead list the performances every 10, as well as the best. The performances for each value of k are given below in Table 1.

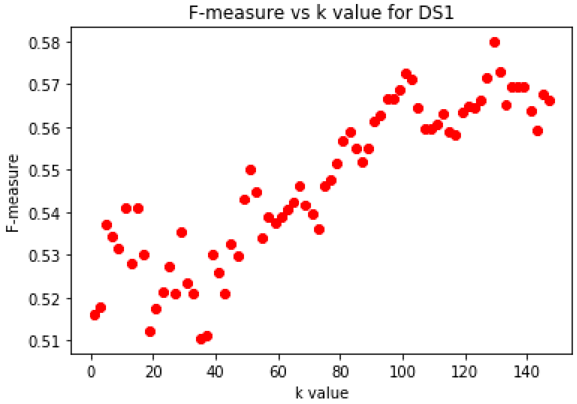
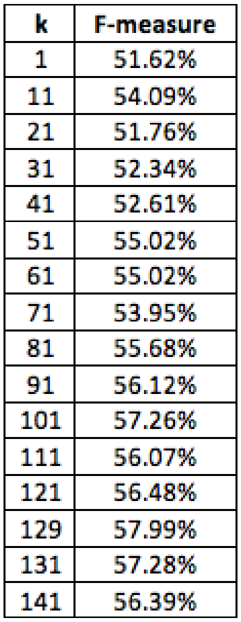


Figure 1: F-measure vs k value for NN, dataset 1

This classifier does much worse than LDA. The linear approach yields around a 95% F_1 measure, whereas this reaches about 58% at the maximum, at a k value of 129. Generally, the k-NN classifier improves as k increases, up to about 100, where it then fluctuates. Depending on the

Table 1: F-measure values



k	F-measure
1	51.62%
11	54.09%
21	51.76%
31	52.34%
41	52.61%
51	55.02%
61	55.02%
71	53.95%
81	55.68%
91	56.12%
101	57.26%
111	56.07%
121	56.48%
129	57.99%
131	57.28%
141	56.39%

(randomly) generated data, the optimal value of k could change greatly. It would be a good idea to use a separate set of data to properly select a value of k.

Below are the computed performance measures for dataset 1 with k-NN (k = 129) classification.

Best fit accuracy	56.42 %
Precision	55.97 %
Recall	60.17 %
F-measure	57.99 %

4. Dataset 2

Again, as for dataset 1, I decided to save multiple text files, split into the testing and training sets, as well as the positive and negatively classified sets. I did this instead of using a label column.

5. LDA & k-NN with Dataset 2

Below are the computed performance measures for dataset 2 with LDA classification.

Best fit accuracy	55.83 %
Precision	55.52 %
Recall	58.67 %
F-measure	57.05 %

Next, k-NN was performed with k varying again. The values for k this time can be seen in Figure 2, below. It is worth noting that there is no clear trend at all with how k changes and the F_1 measure for k-NN with dataset 2.

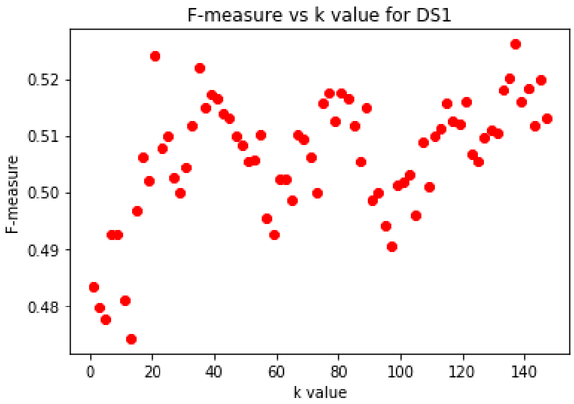


Figure 2: F-measure vs k value for NN, dataset 2

The k corresponding to the highest f value was k = 137, which yielded an F-measure of 52.61%. Using this value, the following performance measures were found for dataset 2 with k-NN classification.

Best fit accuracy	56.17 %
Precision	57.25 %
Recall	48.67 %
F-measure	52.61 %

Now both of the classification methods do not perform well. With many different blobs of data, neither method is able to classify well. LDA still performs better than k-NN, which is against what I would have guessed. Since there are different clumps of data, with certain covariances around certain means, I would have guessed that nearest neighbours would have been able to be quite accurate within those clumps, but this appears to not be the case.

6. Final comments

With the first dataset, which was created using a single gaussian, LDA classification was very effective. K-NN was not effective in this case. With the second dataset, both classifiers had about the same accuracy, which was very poor (56%). As well, the precision of each classifier was about the same. However, one thing that is interesting is that the recall of the k-NN classifier was much lower than that of LDA, and thus LDA had a higher F-measure.