COMP 551

Assignment 2 report

Yang Lu

260631276

1.

DS1 train saved to “DS1\_train.csv”

DS1 valid saved to “DS1\_valid.csv”

DS1 test saved to “DS1\_test.csv”

2.1.(a)

Best fit accuracy, recall, precision and F-measure achieved:

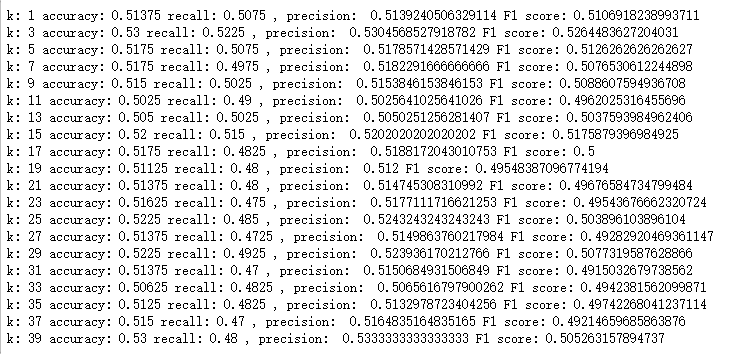
|  |  |
| --- | --- |
| accuracy | 0.96625 |
| recall | 0.965 |
| precision | 0.9674185463659147 |
| F-measure | 0.9662077596996245 |

2.1.(b)

Coefficients reported in “Assignment2\_260631276\_2\_1\_b.txt”.

3.(a)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| K | F-measure | K | F-measure | K | F-measure |
| 1 | 0.51069 | 15 | 0.51758 | 29 | 0.50773 |
| 3 | 0.52644 | 17 | 0.5 | 31 | 0.49150 |
| 5 | 0.51262 | 19 | 0.49548 | 33 | 0.49424 |
| 7 | 0.50765 | 21 | 0.49677 | 35 | 0.49742 |
| 9 | 0.50886 | 23 | 0.49544 | 37 | 0.49215 |
| 11 | 0.49620 | 25 | 0.50390 | 39 |  |
| 13 | 0.50375 | 27 | 0.49283 |  |  |



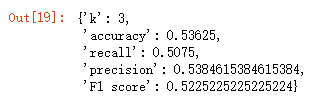
Based on the F-measure, k-NN classifier performs clearly worse than GDA method.

One possible explanation to this is that the data points in two sets are linearly separable due to their features while we are trying to apply k-NN classifier to get a non-linear boundary of classification.

The best k found by using validation set is 3 this time, although there is no huge difference in F-measure with corresponding k values.

3.(b)

k: 3 accuracy: 0.53625 recall: 0.5075 , precision: 0.5384615384615384 F1 score: 0.5225225225225224



4.

DS2 train set saved to “DS2\_train.csv”

DS2 valid set saved to “DS2\_valid.csv”

DS2 test set saved to “DS2\_test.csv”

5.1.a

accuracy: 0.45875

recall: 0.45

precision: 0.4580152671755725

F1 score: 0.45397225725094575



5.1.b

Coefficients reported in “Assignment2\_260631276\_5\_1\_b.txt”.

5.2

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| K | F-measure | K | F-measure | K | F-measure |
| 1 | 0.51661 | 15 | 0.52774 | 29 | 0.49564 |
| 3 | 0.48805 | 17 | 0.50428 | 31 | 0.49938 |
| 5 | 0.49685 | 19 | 0.50801 | 33 | 0.50932 |
| 7 | 0.51111 | 21 | 0.53580 | 35 | 0.51238 |
| 9 | 0.51244 | 23 | 0.52410 | 37 | 0.51644 |
| 11 | 0.52812 | 25 | 0.52891 | 39 | 0.52760 |
| 13 | 0.51733 | 27 | 0.50860 |  |  |

Using DS2 dataset, the k-NN classifier perform slightly better than GDA. The best k we can get using DS2 validation set is k=21, although there is no outstanding variation in F-measure as k varies.

The fact that GDA perform badly may suggest that the datasets are not linearly separable.

As for the reason why k-NN classifier also have a bad performance, one possible reason is that the size of training set is not large enough to predict objects with 20 features correctly. Also, the dataset is not normalized so that some features might be dominant than others(which is not intended to be) since they might possess a relatively larger values.

5.3

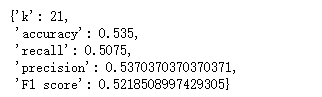
Using k=21, the best fit measurement using test set is:

k: 21 accuracy: 0.535

recall: 0.5075

precision: 0.5370370370370371

F1 score: 0.5218508997429305



6.

For GDA approach, the performance drop dramatically as we move from DS1 to DS2 since we generate DS2 data from mixtures of three gaussian distributions, which is not likely to be linearly separable.

For the K-NN classifier, since we do not increase the size of dataset used for training and we do not add weight or normalize the data, the performance of K-NN classifier remains around 0.5 F1 score for both DS1 and DS2.

Both of them do not work well for DS2, but for DS1, GDA performs much better than K-NN classifier since the data is linearly separable.