COMP 551: Applied Machine Learning
Assignment #2
Report

By:
Weishi Wang (260540022)

Submitted to:
Prof. Sarath Chandar

Oct 21st, 2018

Question 1.

Question 1 is straightforward. A new folder called "DS1" is created to store all DS1 data including the training set, validation set and test set. They are tagged with "0" and "1", which represent negative and positive class, and are stored in "DS1_train.txt", "DS1_valid.txt" and "DS1_test.txt". Code is available in jupyter notebook.

Question 2.

1. (a)

The best fit accuracy is: 0.9575
The best fit precision is: 0.9598
The best fit recall is: 0.9550
The best fit F-measure: 0.9574

These data are also saved in "Assignment_260540022_2_1_a.txt"

(b)

The coefficient learnt including $\emptyset, \mu_0, \mu_1, S$, are saved in the file called "Assignment2_260540022_2_1_b.txt".

Question 3.

(a)

A plot of all 4 metrics over different k values is generated. The values of k are all chosen to be odd numbers to avoid the situation when two classes are tied for even k's. The plot evaluates the accuracy, precision, recall and F-measure on every odd k from 1 to 200. The plot is shown in figure 3.1. According to the plot, we can determine that the performance of the k-NN is much worse than GDA, since all metrics of k-NN classifier for odd k from 1 to 200 are below 0.6, whereas all metrics for GDA are above 0.95. From the plot, we see that when k is 113, the F-measure is the largest. This happens because when k is at certain value, the noise can be reduced and makes the classifier overall smoother. Similar to linear regression, this is at a point where there is no overfitting or underfitting, which makes it the best k value.

(b)

According to the plot on this dataset, the best k value is 113. When k is 113:

The accuracy: 0.56625
The precision: 0.56955

The recall: 0.54250
The f-measure: 0.55570

These values are also stored in “Assignment2_260540022_3_b.txt”

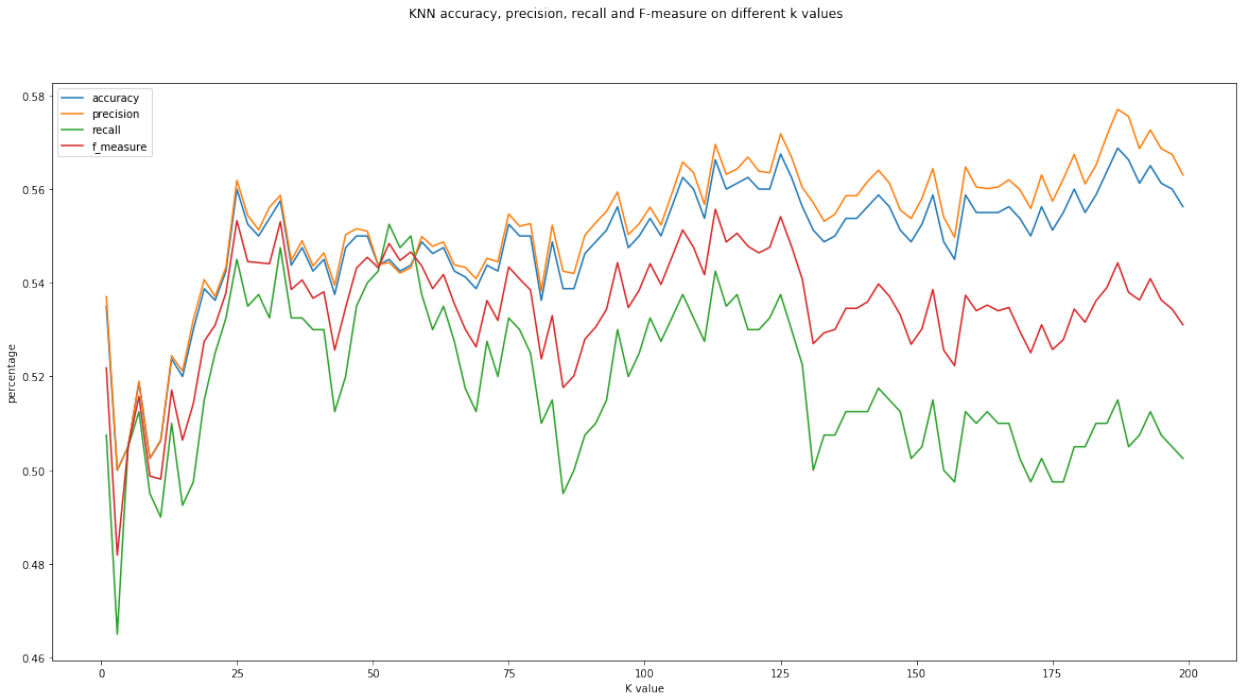

Figure 3.1: KNN accuracy, precision, recall and F-measure on different k values

Question 4.

Similar to question 1, data are properly generated (see jupyter notebook for code). They are stored in the “DS2” folder, and are called “DS2_train.txt”, “DS2_valid.txt” and “DS2_test.txt”.

Question 5.

1. (a)

The best fit accuracy is: 0.4875
The best fit precision is: 0.4876
The best fit recall is: 0.4925
The best fit F-measure: 0.4900

These data are also saved in "Assignment_260540022_5_1_a.txt"

(b)

The coefficient learnt including $\emptyset, \mu_0, \mu_1, S$, are saved in the file called "Assignment2_260540022_5_1_b.txt".

2.

In certain region, k-NN performs better than GDA, i.e. odd value of k from 1 to 150. Then the performance of the k-NN drops below the performance of GDA. From the plot, we obtain the best k value, which is 95. The reason is similar to question 3(a), as when k equals to 95, the classifier does not compromise its predicting power trying to classify every data, nor generalize too much and become meaningless. For these particular dataset, the best k value is then 95.

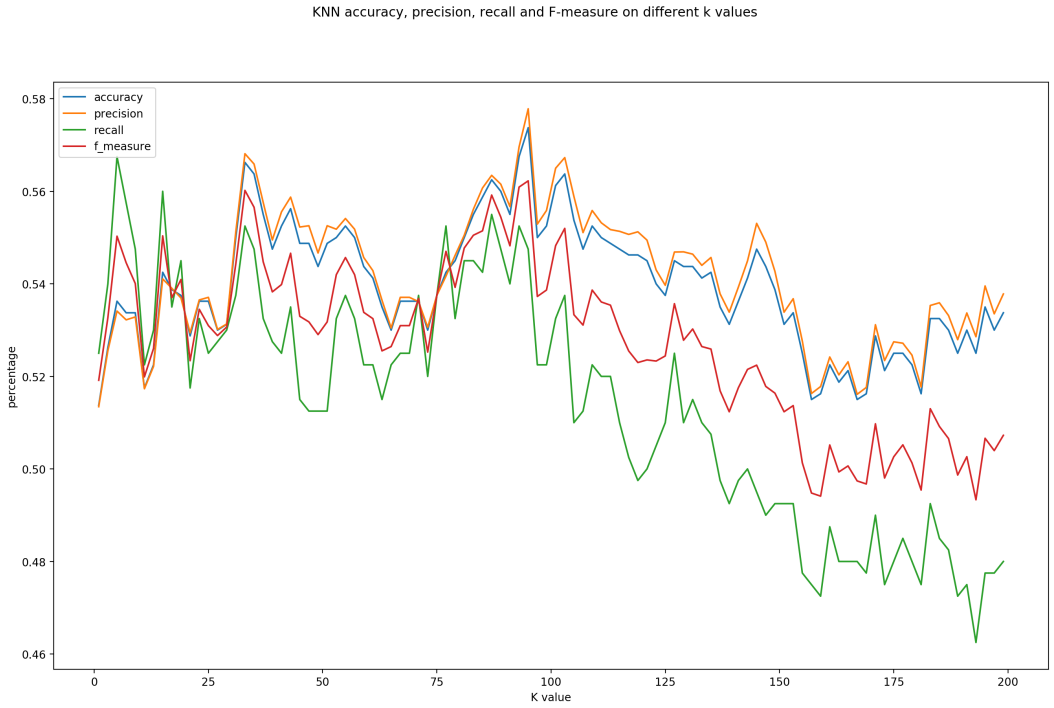

Figure 5.1: KNN accuracy, precision, recall and F-measure on different k values for dataset2

3.

According to the plot, the best k value is 95. When k is 95:

The accuracy: 0.57375
The precision: 0.57784
The recall: 0.54750
The f-measure: 0.56226

These values are also stored in "Assignment2_260540022_5_3.txt"

Question 6.

Differences:

- In dataset DS2, k-NN classifier can actually perform better than GDA model for certain k values. However, in dataset DS1, GDA always perform better with all metrics above 0.95.

- In DS1, only one Gaussian model is used for one class, therefore the generated data are in 2 clusters, which makes linear classifier such as GDA very powerful in this case. In DS2, each class contains 3 Gaussian with means close to the means of the other class, and cause the data unevenly distributed in space. In this case, linear classifier loses its power and model like k-NN actually perform better.

Similarities:

- The performance of k-NN for both datasets is not great, about 0.55 for all metrics. It is favorable in DS2 only because its GDA performance is too poor.

- The performance of k-NN is unstable for both datasets. We can see up and down in both plots.

- For both datasets, the best k value is in the middle range.