INF8953CE(Fall 2020)
Machine Learning

Sarath Chandar Anbil Parthipan
Département Génie Informatique et Génie Logiciel
École Polytechnique de Montréal, Québec, Canada
`Programming Assignment #2`

Student's name, Identification: Moses Openja, 2033785

Date of the submission: 26/10/2020

1 Generating synthetic data using single class multivariate Gaussian distribution

- Generated data names are for question 1 are:

  DS1_train.csv, DS1_test.csv, DS1_valid.csv

  The corresponding code is label Question 1 until beginning of Question 2

- Testing the performance of DS1 dataset with K-NN classification of $K = 5$ got: Accuracy: 0.578750, Precision: 0.583113, Recall: 0.552500, F1 score: 0.567394

2 GDA model using the maximum likelihood approach

a). For DS1, report the best fit accuracy, precision, recall and F-measure achieved by the classifier.

- Accuracy: 0.966250, Precision: 0.966276, Recall: 0.966250, F1 score: 0.966250.

  The performance above is based on the test set, as required in the assignment.

  File Name: "Assignment2_2033785_2_ 1_a.txt"

b). Report the coefficients learnt

- w0: [1.35464974 1.31507699 1.3314178 1.2815658 1.32796116 1.3255876 1.2801238 1.30097595 1.29455227 1.19711815 1.33164364 1.30814516 1.24153705 1.30350525 1.36083324 1.34496766 1.37031654 1.3047932 1.34410696 1.29102756]

  w1: [2.11957295 2.09784241 2.0733269 2.08536989 2.07177151 2.09712021 2.08439585 2.12640347 2.12214562 2.16672049 2.05441559 2.12151196 2.13261964 2.13715674 2.07409241 2.12984784 2.0945371 2.12444338 2.08053467 2.06898178]

  File Name: "Assignment2_2033785_2_ 1_b.txt"

  The complete values including the sigma can be found in the .txt file.

  The corresponding code for Question 2 is labeled as Question2 until beginning of Question 3 in the notebook file.

3 For DS1, use k-NN to learn a classifier

Table 1: Performance results using KNN classification on the DS1 train and test sets

K	precision	recall	f1-score	accuracy
2	0.542500	0.542500	0.542500	0.542500
4	0.552500	0.552500	0.552500	0.552500
6	0.570000	0.570000	0.570000	0.570000
8	0.553750	0.553750	0.553750	0.553750
10	0.535000	0.535000	0.535000	0.535000
12	0.570000	0.570000	0.570000	0.570000
14	0.565000	0.565000	0.565000	0.565000
16	0.568750	0.568750	0.568750	0.568750
18	0.566250	0.566250	0.566250	0.566250
20	0.573750	0.573750	0.573750	0.573750
25	0.557500	0.557500	0.557500	0.557500
30	0.556250	0.556250	0.556250	0.556250
35	0.540000	0.540000	0.540000	0.540000
40	0.537500	0.537500	0.537500	0.537500
45	0.542500	0.542500	0.542500	0.542500
50	0.540000	0.540000	0.540000	0.540000

Table 2: Performance results using KNN classification on the DS1 train and validate sets

K	precision	recall	f1-score	accuracy
2	0.543590	0.530000	0.536709	0.542500
4	0.554974	0.530000	0.542199	0.552500
6	0.572539	0.552500	0.562341	0.570000
8	0.558266	0.515000	0.535761	0.553750
10	0.538043	0.495000	0.515625	0.535000
12	0.576923	0.525000	0.549738	0.570000
14	0.573446	0.507500	0.538462	0.565000
16	0.577031	0.515000	0.544254	0.568750
18	0.574648	0.510000	0.540397	0.566250
20	0.583569	0.515000	0.547145	0.573750
25	0.564246	0.505000	0.532982	0.557500
30	0.561983	0.510000	0.534731	0.556250
35	0.544444	0.490000	0.515789	0.540000
40	0.542614	0.477500	0.507979	0.537500
45	0.548023	0.485000	0.514589	0.542500
50	0.546784	0.467500	0.504043	0.540000

- File Name: "Assignment2_2033785_3_ 1_a.txt"

a). Does this classifier performs better than GDA or worse? Are there particular values of k which perform better? Why does this happen ? Use F1-Measure for model selection.

- As shown in Table (1 and 2), KNN classifier performed much worse than the GDA model above. This could be due to the fact that KNN is completely non-parametric approach which makes no assumptions about the shape of the decision boundary. The decision boundary is linear given the DS1 dataset is a binary classification problem.

b). Report the best fit accuracy, precision, recall and f-measure achieved by this classifier

- Precision: 0.573750, Recall: 0.573750, F1 score: 0.573750, Accuracy: 0.573750.

  With value of $K = 20$

  File Name: "Assignment2_2033785_3_ 1_b.txt"

  The corresponding source code for Question 3 is labeled as Question3 until beginning of Question 4 in the notebook file.

4 Synthetic data set with mixture of 3 Gaussians

- Generated data names are for question 4 are:

  DS2_train.csv, DS2_test.csv, DS2_validate.csv

  The corresponding source code for this part is labeled as Question4 until start of Question 5 in the notebook file.

5 Now perform the experiments in questions 2 and 3 again, but now using DS2.

1. Estimate the parameters of the GDA model using the maximum likelihood approach

a). For DS1, report the best fit accuracy, precision, recall and F-measure achieved by the classifier.

- Accuracy: 0.348750, Precision: 0.357836, Recall: 0.339353, F1 score: 0.319012

  File Name: "Assignment2_2033785_5_ 1_a.txt"

b). Report the coefficients learnt

- w0: [1.29724275 1.25102015 1.25004288 1.28276062 1.2989058 1.31526389 1.25354354 1.23550585 1.26595639 1.3037435 1.23173239 1.31843288 1.36190463 1.28937024 1.29656502 1.34739472 1.37227629 1.29930885 1.2857059 1.34835781]

  w1: [0.92480382 0.91899941 1.04501533 0.96098354 1.00304952 0.97531307 0.95780066 0.95276701 0.96189318 0.93158991 0.98127981 1.04586067 0.95713262 0.98932708 1.02272357 0.97261192 0.98520985 0.98567112 0.97661387 0.96837571]

  w2: [1.39698335 1.36873928 1.45029711 1.41682151 1.45625061 1.32099039 1.41018726 1.42241944 1.45919498 1.35199226 1.37948593 1.32646174 1.33457397 1.37718355 1.38782479 1.29662418 1.48695119 1.32193448 1.42819359 1.45026341]

  For complete results:

  File Name: "Assignment2_2033785_5_ 1_b.txt"

Table 3: Performance results using KNN classification on the DS2 train and test sets

K	precision	recall	f1-score	accuracy
2	0.765114	0.732873	0.746064	0.808750
4	0.822172	0.728152	0.754473	0.833750
6	0.846831	0.740105	0.769710	0.841250
8	0.859159	0.694358	0.721160	0.828750
10	0.860319	0.689397	0.712655	0.835000
12	0.876920	0.704402	0.733356	0.838750
14	0.875071	0.686617	0.707264	0.838750
16	0.871801	0.676856	0.692126	0.837500
18	0.869734	0.674576	0.690244	0.835000
20	0.886367	0.657015	0.661999	0.831250
25	0.881400	0.648299	0.651022	0.823750
30	0.878603	0.628865	0.616465	0.817500
35	0.875967	0.616688	0.593820	0.813750
40	0.872188	0.612212	0.589453	0.807500
45	0.869721	0.602404	0.571807	0.802500
50	0.869283	0.601075	0.570195	0.800000

Table 4: Performance results using KNN classification on the DS2 train and validate sets

K	precision	recall	f1-score	accuracy
2	0.784622	0.745737	0.761230	0.831250
4	0.816232	0.733402	0.758021	0.848750
6	0.873950	0.743451	0.775561	0.866250
8	0.869872	0.717134	0.746390	0.856250
10	0.870350	0.696943	0.721114	0.850000
12	0.877127	0.704381	0.728226	0.860000
14	0.884943	0.691222	0.711258	0.855000
16	0.871573	0.667044	0.677065	0.843750
18	0.874869	0.671870	0.684180	0.846250
20	0.873480	0.666361	0.674148	0.847500
25	0.886115	0.634610	0.629459	0.826250
30	0.887646	0.633569	0.624977	0.828750
35	0.885169	0.627510	0.615490	0.825000
40	0.877862	0.609737	0.586559	0.813750
45	0.877430	0.609871	0.586605	0.813750
50	0.875564	0.602782	0.575803	0.808750

5) 2. Does k-NN classifier perform better than GDA or worse? Are there particular values of k which perform better? Why does this happen ?

- As shown in Table (3 and 4), KNN classifier performed far better than the GDA model above on DS2 dataset. This is because DS2 contains decision boundary that is with highly non-linear unlike in DS1. KNN performed better is considered a completely non-parametric approach which makes no assumptions about the shape of the decision boundary.

- File Name: "Assignment2_2033785_5_ 2_.txt"

5) 3. Report the best fit accuracy, precision, recall and f-measure achieved by this classifier.

- Precision: 0.846831, recall: 0.740105, f1 score: 0.769710, accuracy: 0.841250.

  The best fit results for $K = 6$.

  File Name: "Assignment2_2033785_5_ 3_.txt"

  The corresponding source code for all questions 5 is in the notebook labelled from question 5

6 Comment on any similarities and differences between the performance of both classifiers on datasets DS1 and DS2?

- GDA model performed better on DS1 because its a binary problem with linear decision boundary unlike in DS2 which is highly non-linear decision boundary.