Divya Thomas

Class: CS 677

Date: 4/1/2023

Non-code answers to assignment.

**Question 1-1.** Pandas frame is printed in question1.py file.

**Question 1-2.** Please see calculated output below.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| class | µ(f1) | σ(f1) | µ(f2) | σ(f2) | µ(f3) | σ(f3) | µ(f4) | σ(f4) |
| 0 | 2.28 | 2.02 | 4.26 | 5.14 | 0.80 | 3.24 | -1.15 | 2.12 |
| 1 | -1.87 | 1.88 | -0.99 | 5.40 | 2.15 | 5.26 | -1.25 | 2.07 |
| all | 0.43 | 2.84 | 1.92 | 5.87 | 1.40 | 4.31 | -1.19 | 2.10 |

**Question 1-3.**

For class 0, all means were positive with the exception of f4. For class 1, all means were negative with the exception of f3. Overall, all means were positive except for f4, which contains negative means for both classes. As for standard deviations all class categories stayed within 1.0 of each other with the exception of f3 that had a larger discrepancy.

**Question 2-1.** Please see question2.py for the pairplot data and the two pdf files attached (good\_bills.pdf and fake\_bills.pdf) in the datasets directory.

**Question 2-2.**

From the pairplot generated from the training data:

* It seems that f1 >= 0 is mostly green (good) while red (bad) is f1<0
* It seems that majority of green is f2 > 5 and red is f2<5
* It seems that majority of green is f3 <= 6 while red is dispersed throughout

**Question 2-3.** Test dataframe is altered to add predictions in question2.py

**Question 2-4.** Please see output in question2.py.

**Question 2-5.** Since the test data is compiled randomly, the counts may differentiate for each run. After a few runs, I can confidently say these counts and accuracy/tpr/tnr calculations are very consistent throughout with below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP** | **FP** | **TN** | **FN** | **Accuracy (%)** | **TPR** | **TNR** |
| 164 | 1 | 311 | 210 | 69.24 | 0.44 | 1 |

**Question 2-6.**

My simple classifier is actually much more accurate than I expected for the simplicity of it. It has almost 100% accuracy for predicting “fake” bills, when “real” bills is only 44%. Overall accuracy is better than 50%, ranging from 64-70% consistently.

**Question 3-1.**

3: 99.85422740524781

5: 99.85422740524781

7: 100.0

9: 99.41690962099126

11: 99.41690962099126

**Question 3-2.** 7 is the optimal k value with 100% accuracy.

Chart, line chart

Description automatically generated

**Question 3-3.** Performance measures for k = 7.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **TP** | **FP** | **TN** | **FN** | **Accuracy (%)** | | **TPR** | **TNR** |
| 385 | 0 | 301 | 0 | 100 | 1 | | 1 |

**Question 3-4.**

The k-NN classifier is extremely better than my simple classifier in the previous question. All performance measures have improved, except for TNR, which stayed at the max value of 1.

**Question 3-5.**

Simple Classifier:

f1 f2 f3 f4 prediction

0 5 7 3 2 **green**

7-NN Classifier:

f1 f2 f3 f4 prediction

0 5 7 3 2 **green**

**Question 4-1.**

f1 missing: 96.20991253644316

f2 missing: 97.23032069970844

f3 missing: 96.35568513119533

f4 missing: 98.39650145772595

**Question 4-2.**

The accuracy did not increase in any of the 4 cases. The accuracy went down for al four scenarios tested compared to the case when all four features were used.

**Question 4-3.**

It seems that accuracy is most lost when the f1 feature is missing. The f3 feature was a close second to this, being only 0.1% higher in accuracy than f1.

**Question 4-4.**

The feature that contributed least to the loss of accuracy is f4, with the highest accuracy of 98.4%.

**Question 5-1.** Logistic regression accuracy: 98.83%

**Question 5-2.** Performance measures for Logistic Regression

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TP** | **FP** | **TN** | **FN** | **Accuracy (%)** | **TPR** | **TNR** |
| 377 | 0 | 301 | 8 | 98.83 | 0.98 | 1 |

**Question 5-3.**

Yes, the logistic regression predictions are much better, across all measures, than my simple classifier. It has the same measure for TNR, at max value.

**Question** **5-4.**

No, the logistic regression and k=7 k-NN classifiers had the same perfect measures in false positives, true negatives, and TNR. This shows that they are equally as successful in detecting fake bills. However, the logistic regression scored less successfully for detecting good bills, especially since the k-NN had a perfect score for that as well. Therefore, overall accuracy for logistic regression, at 98.83% was less than k-NN’s 100%.

**Question 5-5.**

Logistic Regression Classifier:

f1 f2 f3 f4 prediction

0 5 7 3 2 **green**

The logistic regression predicted the same label as both the simple and k-NN classifiers.

**Question 6-1.**

Accuracy with f1 missing: 81.77842565597668

Accuracy with f2 missing: 89.067055393586

Accuracy with f3 missing: 87.17201166180757

Accuracy with f4 missing: 98.68804664723032

**Question 6-2.**

The accuracy of all four cases decreased compared to when all features are included when using logistic regression. The closest to the original, though, was when f4 is missing. This is only about 0.1% less in accuracy than the case with all four features included.

**Question 6-3.**

All cases, except for f4 missing, decreased in accuracy substantially compared to the original. However, the feature removal that resulted in the greatest loss to accuracy was f1, which has a 17% decrease in accuracy.

**Question 6-4.**

As mentioned in 6-2, f4 being missing contributed almost nothing to the loss of accuracy, with only a 0.1% drop.

**Question 6-5.**

Yes, both the logistic regression and k-NN data show that missing f1 resulted in greatest loss to accuracy and missing f4 resulted in the least amount of loss to accuracy.