Marisa Paone
MET CS677
7/24/2023
Homework #3
NOTE: Since my training and testing sets are split in half each time the program is run, different results are received. I included a txt file of the run of the program I completed and saved it as output.txt

Question 1.

Please see main.py.

1. *Summarize your results into a table.*

class	μ(f_1)	σ(f_1)	μ(f_2)	σ(f_2)	μ(f_3)	σ(f_3)	μ(f_4)	σ(f_4)
0	2.28	2.02	4.26	5.14	0.80	3.24	-1.15	2.13
1	-1.87	1.88	-0.99	5.4	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

2. *Examine your table. Are there any obvious patterns in the distribution of banknotes in each class?*
   Knowing that class 0 are good banknotes and class 1 are counterfeit banknotes, the obvious patterns in the distribution of banknotes are for class 1. The mean for features 1, 2, and 4 are negative and close to a value of negative 1. Also, for class 1, the standard deviation for feature 3 is higher than class 0 by 2 (greater variability). For class 0, the mean values are positive (except for feature 4) and have a value of 2 or greater. The standard deviation of features 2 and 3 for class 0 banknotes is relatively smaller than the standard deviation for class 1 banknotes.

Question 2.

Please see main.py, good_bills.pdf and fake_bills.pdf

3. *Visually examine your results of the pairplots. Come up with three simple comparisons that you think may be sufficient to detect a fake bill.*
   My simple classifier is a banknote is good if f_1 > -2.5, and f_4 > -2, and f_3 < 8.
5. *Summarize your true label findings in the table below:*

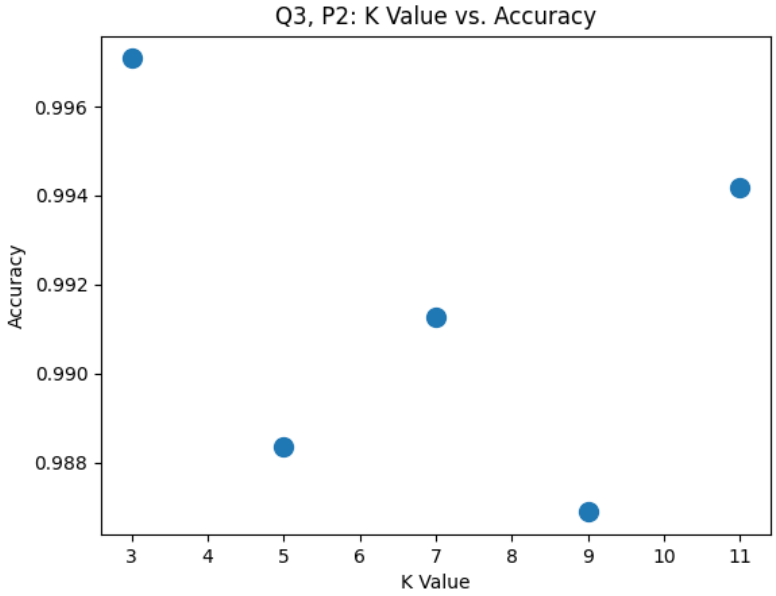
True Positives	False Positives	True Negatives	False Negatives	Accuracy	TPR	TNR
268	132	165	121	0.631195	0.67	0.576923

6. *Does your simple classifier give you higher accuracy on identifying fake bills or real bills? Is your accuracy better than 50%?*
   My simple classifier gives higher accuracy on identifying real bills since my TPR is higher than my TNR. My accuracy is better than 50%, it falls around 63% and the lowest I received from running the program a couple of times is 60%.

Question 3.

Please see main.py.

2. *Plot a graph showing the accuracy of k-NN classifiers. What is the optimal value k* of k?*



The optimal k value for this run of the program is k = 3, followed by k = 11.

3. *Use the optimal value k* to compute performance measures and summarize them in a table.*
   k = 3:

True Positives	False Positives	True Negatives	False Negatives	Accuracy	TPR	TNR
387	2	297	0	0.997085	0.994859	1

4. *Is your k-NN classifier better than your simple classifier for any of the measures from the previous table?*
   Yes, my k-NN classifier is very close to 100% accurate. It is 99.7% accurate in predicting bills. It is better than my simple classifier for each measure in this table.
5. *Consider a bill x that contains the last 4 digits of your BUID as feature values. What is the class label predicted for this bill by your simple classifier and k-NN using k*?*
   Simple Classifier Prediction: green
   k-NN prediction: green

Question 4.

Please see main.py.

2. *Did accuracy increase with k* in any of the 4 cases compared with accuracy when all 4 features are used?*
   Yes, accuracy increased when removing feature 4. It increased to 100% accurate!
3. *Which feature, when removed, contributed the most to loss of accuracy?*
   Removing feature 1 contributed to the most loss of accuracy. It decreased from 99.7% to 95.0% accurate.
4. *Which feature, when removed, contributed the least to loss of accuracy?*
   Removing feature 3 contributed to the lease loss of accuracy. It decreased from 99.7% to 98.3% accurate.

Question 5.

Please see main.py.

2. *Summarize your performance measures for logistic regression in a table.*

True Positives	False Positives	True Negatives	False Negatives	Accuracy	TPR	TNR
374	15	296	1	0.976676	0.96144	0.996633

3. *Is your logistic regression better than your simple classifier for any of the measures from the previous table?*
   Yes logistic regression is better than my simple classifier for all measures.
4. *Is your logistic regression better than your k-NN classifier(using the best k*) for any of the measures from the previous table?*
   No, logistic regression is a smidge worse (97.7%) than my k-NN classifier with k=3 (99.7%).
5. *What is the class label predicted for the bill x by logistic regression? Is It the same label predicted by k-NN?*
   Yes, it is the same label predicted by k-NN, it is green.

Question 6.

Please see main.py.

2. *Did accuracy increase in any of the 4 cases compared with accuracy when all 4 features are used?*
   Accuracy increased when dropping feature 4, it increased from 97.7% to 97.8%!
3. *Which feature, when removed, contributed the most to the loss of accuracy?*
   Removing feature 1 contributed the most to the loss of accuracy. It went from 97.7% to 81.0%.
4. *Which feature, when removed, contributed the least to the loss of accuracy?*
   Removing feature 2 contributed the least to the loss of accuracy. It went from 97.7% to 90.2%.
5. *Is the relative significance of features the same as you obtained using k-NN?*
   Yes, it is the same except for the last question (contributing to the least to loss of accuracy). In both cases though, removing feature 4 increased accuracy, and removing feature 1 decreased accuracy the most. For k-NN removing feature 3 contributed to the least loss of accuracy, and for logistic regression removing feature 2 contributed to the least loss of accuracy.