Yubin Ye U53651145

Question 1

1. load the data into dataframe and add column ”color”. For each class 0, this should contain ”green” and for each class 1 it should contain ”red”.
2. for each class and for each feature f1, f2, f3, f4, compute its mean μ() and standard deviation σ(). Round the results to 2 decimal places and summarize them in a table as shown below:

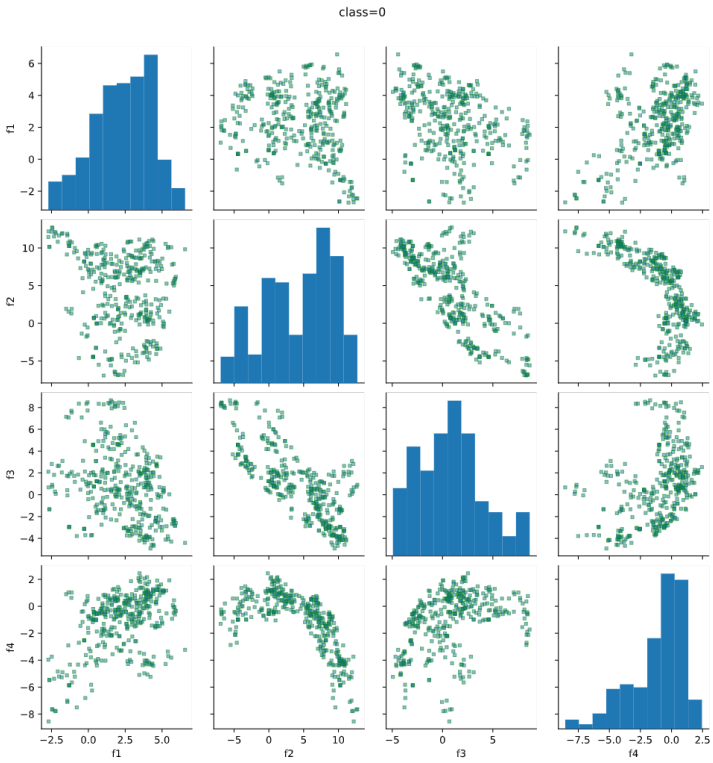
class	u(f1)	std(f1)	u(f2)	std(f2)	u(f3)	std(f3)	u(f4)	std(f4)
0	2.28	2.02	4.26	5.14	0.8	3.24	-1.15	2.13
1	-1.87	1.88	-0.99	5.4	2.15	5.26	-1.25	2.07
all	0.43	2.84	1.92	5.87	1.4	4.31	-1.19	2.1

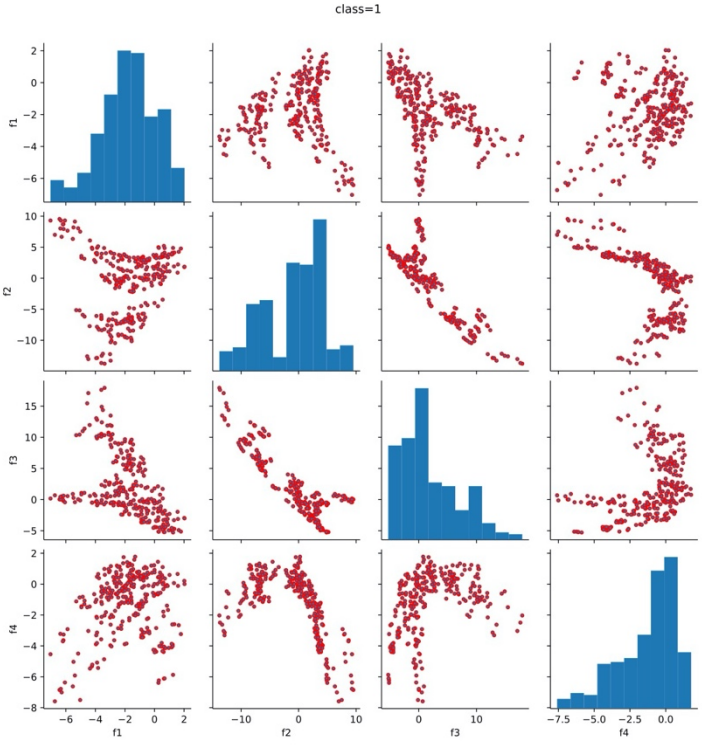
3. examine your table. Are there any obvious patterns in the distribution of banknotes in each class?

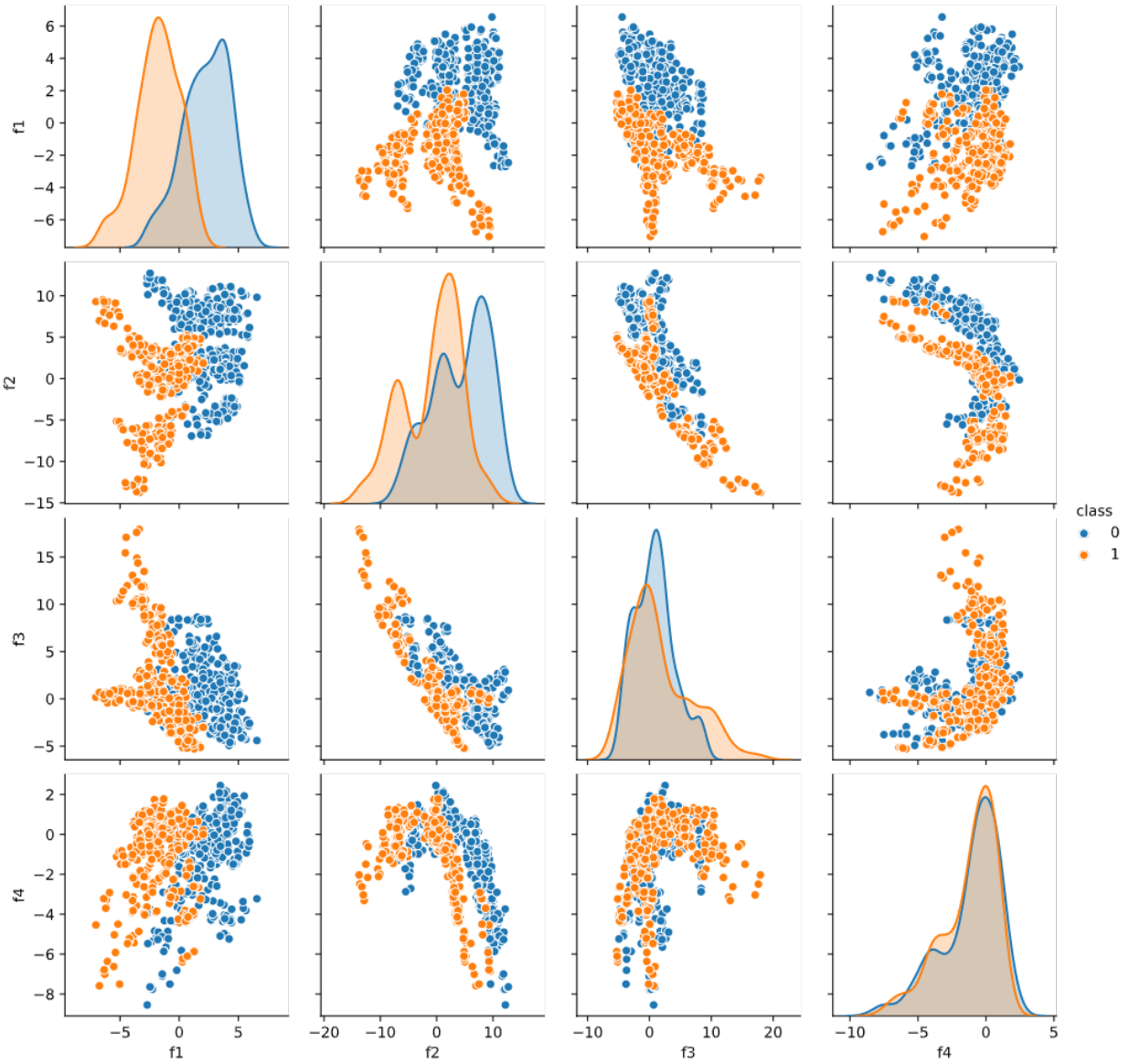
If class = 0, mean of f1, f2 are positive; If class = 1, mean of f1, f2 are negative.
Mean of f1, f2, f3, and f4 if class=1 are all bigger than if class=0

Question 2

1. split your dataset X into training Xtrain and Xtesting parts (50/50 split). Using ”pairplot” from seaborn package, plot pairwise relationships in Xtrain separately for class 0 and class 1.
Save your results into 2 pdf files ”good bills.pdf” and ”fake bills.pdf”







2. visually examine your results. Come up with three simple comparisons that you think may be sufficient to detect a fake bill.

```
if (f_1 > 0) or (f_2 > 10) or (f_3 < 5 and f_4 < -10):
        X = ”good”
else:
        X = ”fake”
```

3. apply your simple classifier to Xtest and compute predicted class labels
4. comparing your predicted class labels with true labels, com- pute the following:
(a) TP - true positives (your predicted label is + and true label is +)
(b) FP - false positives (your predicted label is + but true label is −

(c) TN - true negativess (your predicted label is − and true label is −
(d) FN - false negatives (your predicted label is − but true label is +
(e) TPR = TP/(TP + FN) - true positive rate. This is the fraction of positive labels that your predicted correctly. This is also called sensitivity, recall or hit rate.
(f) TNR = TN/(TN + FP) - true negative rate. This is the fraction of negative labels that your predicted correctly. This is also called specificity or selectivity.

5. summarize your findings in the table as shown below:

TP	FP	TN	FN	accuracy	TPR	TNR
352	47	254	33	0.88	0.91	0.84

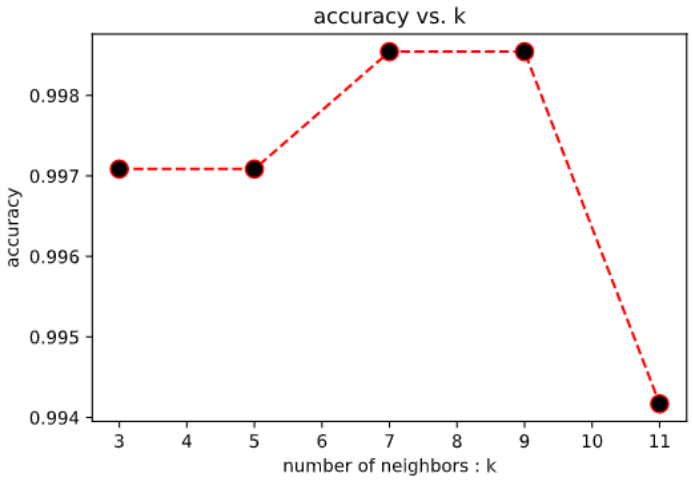
6. does you simple classifier gives you higher accuracy on identifying ”fake” bills or ”real” bills” Is your accuracy better than 50% (”coin” flipping)?
Yes, it is better than 50%

Question 3

(use k-NN classifier using sklearn library)

1. take k = 3,5,7,9,11. Use the same Xtrain and Xtest as before. For each k, train your k-NN classifier on Xtrain and compute its accuracy for Xtest
when k=3, accuracy is 0.9971
when k=5, accuracy is 0.9971
when k=7, accuracy is 0.9985
when k=9, accuracy is 0.9985
when k=11, accuracy is 0.9942

2. plot a graph showing the accuracy. On x axis you plot k and on y-axis you plot accuracy. What is the optimal value k∗ of k?



3. use the optimal value k∗ to compute performance measures and summarize them in the table

Best k = 7 or 9, they have same result:

TP	FP	TN	FN	accuracy	TPR	TNR
384	0	301	1	0.9985	0.9974	1

4. is your k-NN classifier better than your simple classifier for any of the measures from the previous table?
Yes, it is better

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? What is the label for this bill predicted by k-NN using the best k∗?
my predicted label is 0
k-NN predicted label is 0

Question 4

1. take your best value k . For each of the four features f1,...,f4, drop that feature from both Xtrain and Xtest. Train your classifier on the ”truncated” Xtrain and pre- dict labels on Xtest using just 3 remaining features.
if missing f1, accuracy is 0.9519
if missing f2, accuracy is 0.9708
if missing f3, accuracy is 0.9679
if missing f4, accuracy is 0.9927

2. did accuracy increase in any of the 4 cases compared with accuracy when all 4 features are used?
No, they all decreased
3. which feature, when removed, contributed the most to loss of accuracy?
F1
4. which feature, when removed, contributed the least to loss of accuracy?
F4

Question 5

use logistic (regression classifier using sklearn library
1. Use the same Xtrain and Xtest as before. Train your logistic regression classifier on Xtrain and compute its accuracy for Xtest
Accuracy is 0.9781
2. summarize your performance measures in the table TP FP TN FN accuracy TPR TNR

TP	FP	TN	FN	accuracy	TPR	TNR
373	3	298	12	0.9781	0.9688	0.99

3. is your logistic regression better than your simple classifier for any of the measures from the previous table?
Yes, it is better than my rules prediction.

4. is your logistic regression better than your k-NN classifier (using the best k∗) for any of the measures from the previous table?
Yes, it is better than when we miss f1 or f2 or f3 and use k-NN.

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 x by logistic regression? Is it the same label as predicted by k-NN?
logistic regression label is 0.
Yes, it is the same label as predicted by k-NN

Question 6
1. For each of the four features f1, . . . , f4, drop that feature from both Xtrain and Xtest. Train your logistic regression classifier on the ”truncated” Xtrain and predict labels on Xtest using just 3 remaining features.
if missing f1, accuracy is 0.8047
if missing f2, accuracy is 0.9184
if missing f3, accuracy is 0.8746
if missing f4, accuracy is 0.9781

2. did accuracy increase in any of the 4 cases compared with accuracy when all 4 features are used?
No, all accuracies are not increase
3. which feature, when removed, contributed the most to loss of accuracy?
F1
4. which feature, when removed, contributed the least to loss of accuracy?
F4
5. is relative significance of features the same as you obtained using k-NN?
Yes,
F1 is the most significant features in both k-NN and logistic results.
f4 is the least significant features in both k-NN and logistic results.