Linda Zhang

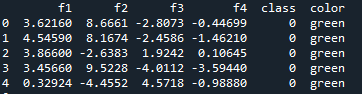
Class: CS 677

Date:4/2/2022

Homework #3

Question 1.1(row 20 to row 30)

I loaded the data into a dataframe and added 0 if it was green and 1 if it was red.



Question 1.2(row 32 to row 56)

I created a dataframe with calculations for mean and standard deviation for each feature.



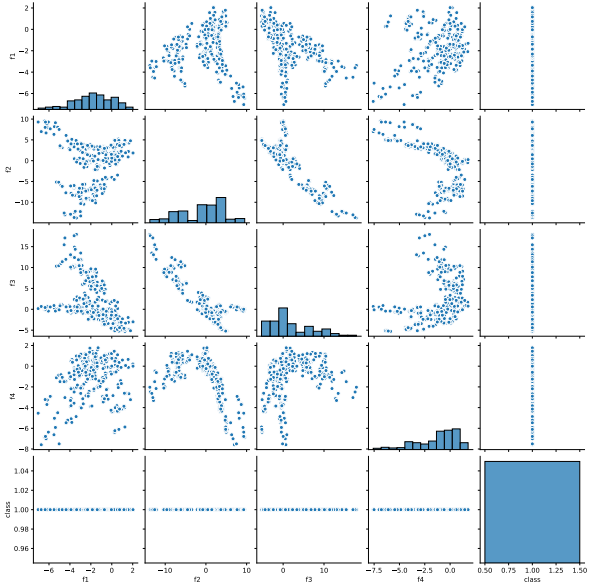
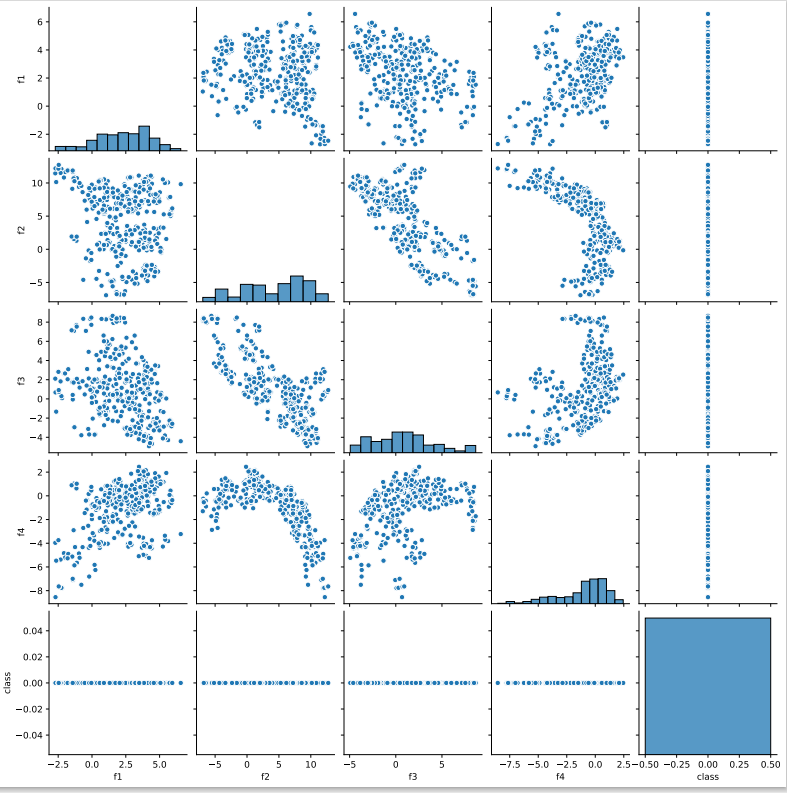
Question 1.3

From the table we can tell right away that the overall mean for feature 4 is negative for fake, real and overall bills and also really close in number. Feature 2 has the highest difference of means when comparing the fake bill and a real bill.

Question 2.1(row 58-69)

Splitting the data using the seaborn package. I created the graph and saved it into a pdf

Good bills: Bad Bills:



Question 2.2(row 72 to row 88)

The classifier I came up with had three rules:

1.f1>-4

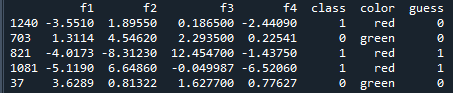
2.f2>-5

3.f3>-5

I created a function called guess that would take in the f1, f2 and f3 value and if it fits my classfier it would return it as a real bill. The results of the guess would be added into a new column called guess.

Question 2.3(row 90 to row 96)

I applied it to the x test dataset and printed the results



Question 2.4(row 98 to row 123)

Created a function that compared the class labels with the true labels. The function returns a table with values.

Question 2.5(row 126 to row 135)

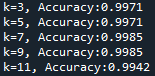


Question 2.6

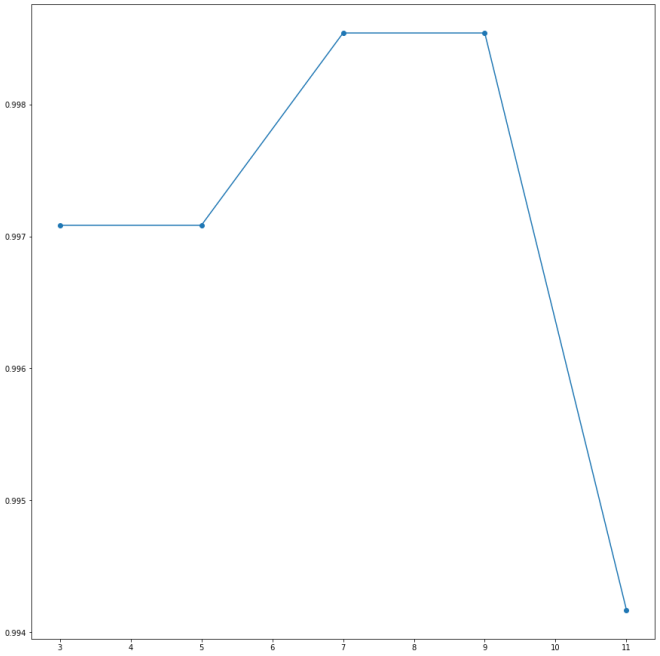
My simple classifier has an accuracy of 71.87%. Although this is not a very high percentage of accuracy it is better than 50%(or coin-flipping). The reason as to why I did not achieve a high percentage is because some of the features have similar ranges so it was hard to find the exact numbers based on the graphs.

Question 3.1(row 138 to row 169)

Set up the data to be used in a k-NN classifier. Created a for loop that will run each k-value into the k-NN classifier. I then calculated the accuracy of each k-value. As you can see the accuracy for the k-NN classifier is very high.



Question 3.2(row 172 to row 178)



Based on the graph we can tell right away that the highest accuracy for k is when k=7, k=8 and k=9.

Question 3.3(row 181 to row 222)

I used k=7 to get the optimal value of k and compared it to x test.



Question 3.4

My k-NN classifier is a lot better than the simple classifier that I have made. The k-NN classifier has an accuracy of 99.85% while my simple classifier only has an accuracy of 71.97%. The k-NN classifier is almost 100% accurate and that is ideal when it comes to machine learning.

Question 3.5(row 225 to row 240)

Using my BUID(1222) as a bill number and the feature would be as follows:

f1=1

f2=2

f3=2

f4=2

Both the simple classifier and the k-NN predicted that my bill was real.

Question 4.1(row 243 to row 267)

I created a function that sets up the feature selection. In the function that I created I made a new x test value that would select the columns that were given and use that to make the new prediction. I also used k=7 since it is k\* and the most optimal choice.

Without f1:0.971

Without f2:0.975

Without f3: 0.975

Without f4:0.994

Question 4.2

The accuracy of the test decreased when any one of the 4 test was removed. If we want the highest accuracy we should keep all four features in the test.

Question 4.3

The feature that contributed the most loss is feature 1. The accuracy dropped from 99.85% to 97.1%.

Question 4.4

The feature that contributed the least loss when removed is feature 4. The accuracy dropped from 99.85% to 99.4%.

Question 5.1(row 270-276)

Created a logistic regression to test x test

Question 5.2(row 279 to 314)

Created a dataframe to summarize the findings



Question 5.3 My logistic regression is better than my simple classifier because it has a higher percentage of accuracy.

Question 5.4

My K-NN classifier looking at the optimal k is higher than my logistic regression.

Question 5.5

My logistic regression table took my buid and guessed that given those 4 values as a feature it would return as a real bill

Question 6.1

I created a function that sets up the feature selection. In the function that I created I made a new x test value that would select the columns that were given and use that to make the new prediction.

Without f1:0.991

Without f2:0.885

Without f3: 0.911

Without f4:0.799

Question 6.2

The accuracy only increased if we removed feature 1 compared to when we consider all four features.

Question 6.3

When we remove feature 4 it contributed the most loss of accuracy because it went from 97.8% accurate to 79.9%

Question 6.4

When we remove feature 1 it contributed the least loss of accuracy because it went from 97.8% accurate to 99.1%

Question 6.5

The results are not the same but actually opposites. When I set up feature selection for logistic regression removing feature 4 contributed the most loss and feature 1 contributed the least loss. For the k-NN using the most optimal K the one that contributed the most loss is feature 1 and the one that cost the least loss is feature 4.