CIS 572 Assignment 3

Steven Walton

May 2, 2019

Problem 1

An analyst wants to classify a number of customers based on some given attributes: total number of accounts, credit utilization (amount of used credit divided by total credit amount), percentage of on-time payments, age of credit history, and inquiries (number of time that the customer requested a new credit account, whether accepted or not). The analyst acquired some labeled information as shown in the following table:

ID	Total Accounts	Utilization	Payment History	Age of History (days)	Inquiries	Label
1	8	15%	100%	1000	5	GOOD
2	15	19%	90%	2500	8	BAD
3	10	35%	100%	500	8	BAD
4	11	40%	95%	2000	6	BAD
5	12	10%	99%	3000	6	GOOD
6	18	15%	100%	2000	5	GOOD
7	3	21%	100%	1500	7	BAD
8	14	4%	100%	3500	5	GOOD
9	13	5%	100%	3000	3	GOOD
10	6	25%	94%	2800	9	BAD

Consider the following three accounts to be labeled:

Total Accounts	Utilization	Payment History	Age of History (days)	Inquiries	Label
20	50%	90%	4500	12	P1
8	10%	100%	550	4	P2
9	13%	99%	3000	6	P3

(a) Before using nearest neighbor methods to make predictions, how would you recommend processing or transforming the data? Why? Make any changes you think appropriate to the data before continuing on to the next two parts.

(b) What are the predicted labels P1, P2, and P3 using 1-NN with L 1 distance? Assume that percentages are represented as their corresponding decimal numbers, so 95% = 0.95. Show your work.

(c) Keep the information of customers 7, 8, 9, and 10 as validation data, and find the best K value for the K-NN algorithm. If the best value of K is not equal to 1, find the new predictions for P1, P2, and P3. Show your work.

Part 1)

Before doing nearest neighbor to make predictions I would suggest dividing the Age of History column by 1000. Smaller numbers make things easier. I would also convert the percentages to their decimal form and relabel "GOOD" = 1 and "BAD" = 0 This gives us the following table

ID	Total Accounts	Utilization	Payment History	Age of History (days)	Inquiries	Label
1	8	0.15	1.0	1	5	1
2	15	0.19	0.90	2.5	8	0
3	10	0.35	1.0	0.5	8	0
4	11	0.40	0.95	2	6	0
5	12	0.10	0.99	3	6	1
6	18	0.15	1.0	2	5	1
7	3	0.21	1.0	1.5	7	0
8	14	0.4	1.0	3.5	5	1
9	13	0.5	1.0	3	3	1
10	6	0.25	0.94	2.8	9	0

This puts everything in a way that a computer can understand and we are manipulating numbers in a consistent way that will bake in a bias into the data.
An example of baking in bias would be to normalize any column, because we would then add information, where we know what the largest number is. This would be a problem because Age of History for P1 would be ¿1. So let's not do that.

Part 2)

The easiest thing to do is put everything into arrays and calculate the distance in each metric. We can do this simply with python my doing the following:

```
for i in range(len(historical_data)):
    print(np.linalg.norm(historical_data[i][:-1] - p1,1))
 22.95
>>> 11.309999999999999
 16.25
 17.65
 15.99
 11.95
 25.39
 14.559999999999999
 18.05
 18.99
```

```
for i in range(len(historical_data)):
    print(np.linalg.norm(historical_data[i][:-1]-p2,1))
>>> 1.5
 13.139999999999999
 8.3
 6.8
 8.46
 12.5
 9.06
 10.01
 8.5
 9.46
```

```
for i in range(len(historical_data)):
    print(np.linalg.norm(historical_data[i][:-1]-p3,1))
 4.03
 8.649999999999999
 7.73
 3.31
>>> 3.03
 11.03
 8.59
 6.6
 7.09
 6.37
```

We use numpy's linear algebra package to call upon norm. Norm defaults to the Euclidean norm, but by passing 1 to the second parameter we can get the Manhattan distance.
We can easily see that the closest label (marked with >>>) is that of ID=2, which is labeled BAD, therefore we should label P1 as BAD.
Doing this again we see that P2 is closest to ID=1, so GOOD.

P3 is closest to ID=5, so GOOD
Giving us

Total Accounts	Utilization	Payment History	Age of History (days)	Inquiries	Label
20	0.50	0.9	4.5	12	BAD
8	0.10	1.0	0.55	4	GOOD
9	0.13	0.99	3	6	GOOD