Dillon Carter (dwc180002)

03/04

ML Algorithms from Scratch

```
C:\Users\Pickle Mustard\Documents\Machine Learning Projects\4375-Intro-to-Machine-Learning\Project_3_ML_Algorithms_From_Scratch>Naive.exe
Opening file titanic_project.csv
Reading line 1
Heading: "","pclass","survived","sex","age"
Size of training set: 800
Size of test set: 246
New length: 1046
NUm Sur: 317
[,]     [,1]    [,2]
[1,]    0.60375 0.39625
[,]     [,1]    [,2]    [,3]
[1,]    0       0       0.249211
[2,]    0.429022        0.375394        0.274448
[,]     [,1]    [,2]
[1,]    0.233438        0.709779
[2,]    1.29022 0.290221
Means
Means
Means
Means
---Confusion Matrix---
[,]     [1]     [0]
[1]     0       110
[0]     0       136
---Accuracy---
0.552846
---Sensitivity---
0
---Specificity---
1
Closing file Titanic.csv

C:\Users\Pickle Mustard\Documents\Machine Learning Projects\4375-Intro-to-Machine-Learning\Project_3_ML_Algorithms_From_Scratch>
```

```
C:\Users\Pickle Mustard\Documents\Machine Learning Projects\4375-Intro-to-Machine-Learning\Project_3_ML_Algorithms_From_Scratch>a.exe
Opening file titanic_project.csv
Reading line 1
Heading: "","pclass","survived","sex","age"
Size of training set: 800
Size of test set: 246
New length: 1046
Weights: 1.11204 | -2.60396
---Confusion Matrix---
[,]     [1]     [0]
[1]     67      43
[0]     22      114
---Accuracy---
0.735772
---Sensitivity---
0.609091
---Specificity---
0.838235
Closing file Titanic.csv

C:\Users\Pickle Mustard\Documents\Machine Learning Projects\4375-Intro-to-Machine-Learning\Project_3_ML_Algorithms_From_Scratch>
```

The runtime of the logistic regression algorithm was much higher than the Naïve Bayes. The gradient descent took multiple loops to complete, running up the time. However, the logistic regression algorithm was a lot more accurate than the Naïve Bayes, having a much higher accuracy.

The generative classification algorithms take a long time to run. The gradient descent algorithm is an example of trading time for accuracy. The accuracy can be much higher than discriminative classifiers. Finding the local minima of a logistic curve is very accurate when making decision on large sets of data.

Discriminative algorithms are generally great for smaller sets of data. As gradient descent can have long runtimes with no matter the size of the data set, it can be faster to take a prediction for the end results given the predicting factors than to find a minima. It will be less reliable on larger data sets.

Looking at reproducibility, I came across these two sources: https://www.science.org/doi/10.1126/scitranslmed.abb1655 and https://arxiv.org/abs/2108.12383. One of the most major challenges when it comes to these types of algorithms is creating a reproducible environment so that others can take what you have researched, repeat it, and evaluate your conclusions. To that end, the results that they get must be as close as possible to the initial results produced during research. Otherwise, the original conclusions can't be evaluated. One of the ways in which it is implemented is keeping the training and test data sets consistent across different runs. In my code, I used a set seed when dividing the training sets, but another method is just to keep the datasets entirely separated.