Fernando Colman and Linus Fackler
CS 4375.003
October 2nd, 2022
ML Algorithms from Scratch

## Code Results

```
Opening titanic.csv
Headings: pclass,survived,sex,age
Closing titanic.csv
Number of observations: 1046
Coefficients :
0: 159.97
1: -23.7194
2: -149.28
3: -8.71368
Accuracy: 1
Sensitivity: 1
Specificity: 1
Duration of Algorithm: 14.136
```

Logistic Regression: To carry out this logistic regression algorithm I used a very high amount of epochs to try and get the weights of the coefficients to as accurate an amount as I could get them. Even though this algorithm was run on a strong computer, it still took 15 seconds to run the program giving us the above coefficient and metric values.

```
Prior probability, survived=no, survived=yes.
0.610000 0.390000

Likelihood values for p(pclass|survived):
0.172131 0.225410 0.602459
0.416667 0.262821 0.320513

Likelihood values for p(sex|survived):
0.159836 0.840164
0.679487 0.320513

Applied to first 5 test observations:
0.701500 0.298500
0.540072 0.459928
1.000000 0.000000
1.000000 0.000000
1.000000 0.000000
Elapsed time in milliseconds: 2 ms
```

Naïve Bayes: As seen in the 2ms runtime, the Naive Bayes algorithm is a very fast algorithm, therefore usable to make predictions in real time, other than Logistic Regression

## Generative Classifiers vs Discriminative Classifiers

Logistic Regression is a discriminative classifier because it estimates the parameters of P(Y|X), while Naïve Bayes is a generative classifier because it estimates parameters for P(Y) and P(X|Y). In terms of which one of these is better, it truly depends on the dataset that they are being used on. Naïve Bayes handles small datasets better, however, logistic regression is only going to get better and better as the dataset grows. Naïve Bayes also has a higher variance than logistic regression but a lower variance.

For our dataset specifically, I think that due to both the low number of predictors and the large number of observations that we had, our logistic regression algorithm ended up being a good choice to predict whether a passenger had survived or drowned. I believe that had the dataset had more factors and predictors then it would have a very different story and Naïve Bayes would have probably been a better option.

## Reproducible Research in Machine Learning

This is the process of repeatedly running your algorithm on certain datasets and obtaining the same/similar result on a particular project. Encompassing design, reporting, data analysis, and interpretation, this process adds value to any continuous integration/delivery cycle. It ultimately smooths the process of making changes to such cycle, without having to implement a new system every time [1].

This effectively states the problem that researchers are trying to solve. 50% of researchers were unable to reproduce their own experiments, making research in the field of Machine Learning harder than it should be.

A reproducible Machine Learning application is therefore built to scale with your business growth, essentially adaptable to the growing demand for speed and volume of model execution. The benefit of such scalability creates trust and credibility with the overall product and is becoming more and more essential [2].

[1] https://arxiv.org/abs/2108.12383
[2] https://www.decisivedge.com/blog/the-importance-of-reproducibility-in-machine-learning-applications/#:~:text=Reproducibility%20with%20respect%20to%20machine,reporting%2C%20data%20analysis%20and%20interpretation