Machine Learning from Scratch

Spencer Gray and Michael Stinnett

Part A:

Logistic Regression Code Output:

```
Opening file titanic_project.csv
heading: "","pclass","survived","sex","age"
Closing file titanic_project.csv.
B0: -1.26106
B1: -0.424828
TP: 80
FP: 17
TN: 113
FN: 35
accuracy: 0.787755
sensitivity: 0.695652
sensitivity: 0.869231
Time to Run: 23 milliseconds
```

Naive Bayes Code Output:

```
Opening file titanic_project.csv.
Closing file titanic_project.csv.

Time to train: 109 microseconds

PRIOR PROBABILITIES
Survived: 0.39 Died: 0.61

LIKELIHOODS
pclass
pclass1 -> Died: 0.172131 Survived: 0.416667
pclass2 -> Died: 0.22541 Survived: 0.262821
pclass3 -> Died: 0.602459 Survived: 0.320513
Sex
female -> Died: 0.159836 Survived: 0.679487
male -> Died: 0.840164 Survived: 0.320513
Age
Mean -> Died: 30.4899 Survived: 28.8261
Standard Deviation -> Died: 14.3233 Survived: 14.4622

METRICS
Confusion Matrix
113   35
18   80
Accuracy: 0.784553
Sensitivity: 0.862595
Specificity: 0.695652
PS C:\Users\mpsti>
```

Part B:

Both algorithms had pretty good accuracy. Each predictor seemed to be important. The results tended to say that a young, female, in passenger class 1 would be the most likely person to survive. Whereas, an older, male, in passenger class 3 would be the most likely to die.

Logistic regression had a slightly higher accuracy probably due to the larger nature of the data set. Also, in this case it seems clear cut to divide the y values into discriminative camps of died or did not because of how telling each of the predictors are. Thus, making the discriminative algorithm slightly better. However, naive bayes ran much faster.

Part C:

There are two main types of classifiers. Those types are Discriminative and Generative. This assignment allowed us to write both types from scratch. Logistic Regression is a discriminative classifier while Naive Bayes is a Generative classifier.

Discriminative classifiers are made explicitly to discriminate between Y values. That makes discriminative algorithms only application to be classifiers. Whereas, Generative models try to find where a data point might be placed in Y space. This is looking more at pure probabilities, thus the raw probabilities can be extracted easier for further applications elsewhere. Ultimately, generative tends to have more bias and disriminatives tend to have more variance.

Sources:
https://www.analyticsvidhya.com/blog/2021/07/deep-understanding-of-discriminative-and-generative-models-in-machine-learning/#:~:text=Discriminative%20models%20draw%20boundaries%20in,the%20labels%20of%20the%20data.

Part D:
Reproducible research in machine learning is being able to run a machine learning algorithm on similar data and receive the same results. This is important because it legitimizes the machine learning algorithm. If your model can only work in certain perfect conditions, then it might be a coincidence that your model is predicting/learning correctly. However, it is a very difficult problem to master. With data that changes rapidly. Hardware that gets upgraded constantly. Maybe even a power outage and system failure causes your work to be unreproducible.

Reproducibility can be better implemented with a clear set of training data for the model. Another important part is to document everything while developing the algorithms. Always consider the pipeline model, that the algorithm is never going to stop development and whenever you stop you are pipelining all of your work down to the next person who works on it.

Sources:
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.

https://arxiv.org/abs/2108.12383