CS4375.004
Parker Tate
Kyle Ayiku

ML Algorithms from Scratch

1. I could not manage to read the data into an eigen library matrix, which is what I did my logistic regression implementation in.

2.

```
A-priori probabilities:
       0        1
0.610000 0.390000

Likelihood values for p(survived | pclass)
          1        2        3
0: 0.172131 0.225410 0.602459
1: 0.416667 0.262821 0.320513

Likelihood values for p(survived | sex)
     Female     Male
0: 0.159836 0.840164
1: 0.679487 0.320513

Age matrix
              0         1
Mean: 30.418203 28.826122
SD:   14.308467 14.439002

Accuracy: 0.784553
Sensitivity: 0.695652
Specificity: 0.862595
Training Time: 4509 microseconds

Program terminated.
C:\Users\tateo\source\repos\hw3_ml\x64\Debug\hw3_ml.exe (process 17284) exited with code 0.
Press any key to close this window . . .
```

Generative Classifiers & Discriminative Classifiers:

Generative classifiers are the parameters estimated for the probabilities P(Y) and P(X, Y) which the statistical method naïve Bayes does whereas discriminative classifiers are the parameters estimated for P(Y|X) with logistic regression as a primary method. The former type of classifiers attempt to have the model for a class, which is essentially how data would be retrieved from the inputs. On the other hand, discriminative classifiers identify which of the input features can be the most distinguishable from each class. Computationally, generative classifiers use its models to obtain knowledge for the joint probability distribution by implementing Bayes theorem to determine the conditional probability, which is known as P(X|Y), then selecting the label that is most appropriate. Conversely, in the realm of a discriminative model, the posterior probability is

calculated, known as P(Y|X) or a map is “drawn” from an input, being X to the label, being Y in an attempt to find out what the decision boundary is.

In essence, both generative and discriminative classifiers are methods which attempt to predict the conditional probability in their respective models. As a result, generative models are superior for smaller data sets since they are not only designed to handle both supervised and unsupervised learning tasks, but they converge at a quicker rate. Contrarily, discriminative classifiers are quite efficient at handling classification tasks since their primary objective is to find solutions for problems that are not on the generic side. In conclusion, both types of modeling algorithms are quite utilizable in the field of machine learning as users can discover more ways of organizing and using data to their advantage to use in different applications across technology and beyond.

Reproducible Research in Machine Learning:

Reproducible research is the process of recording and releasing results determined for an impact evaluation. This method permits users to provide and retrieve information for and from data similar or identical to the original studies which further proves the conclusions resulted from such studies. For reproducible research, there are four vital components: data documentation, data publication, code publication, and output publication.

Data documentation is the part which data is recorded, sorted, and sampled to ensure the reasons of why data is presented the way it is. Next, data publication is the data being released to the public once the data cleaning, collection, and analysis is finished. Then, code publication is the component that requires users to publish the data and reproducible code through thorough and meticulous data management and substantiation for the document. A prime example of the whole reproducible research process is how I have personally utilized R and GitHub to first implement the coding algorithms to complete a few homework assignments, which the code and the document supplementing the code are uploaded into the latter application, available for anyone and everyone to view, determining the credibility and validity of the reproducibility. In conclusion, reproducibility is essential to machine learning as the review of code, data and findings and other discoveries alike can path the way into how this field would not only be further developed, but also, possibly even placing limits on paper documentation.

References:

https://www.analyticsvidhya.com/blog/2021/07/deep-understanding-of-discriminative-and-generative-models-in-machine-learning/

https://blog.ml.cmu.edu/2020/08/31/5-reproducibility/

https://dimewiki.worldbank.org/Reproducible_Research