Yazan Abughazaleh
CS 4375.004

ML Algorithms from Scratch

The goal of this project was to build a logistic regression and Naïve Bayes model for the titanic data set using C++. The models are designed to predict if a person survived on the titanic based off of the predictors of age, sex, and passenger class. The passenger class, sex, and survived attributes are all qualitative while age is a quantitative value.

Model Output

The first output is that of the logistic regression model. The model trains itself on 800 observations and performs predictions on the rest.

```
Opening titanic.csv File
Header: "","pclass","survived","sex","age"
Training Model!
The model took 177177 milliseconds to train.
Coefficients for Logistic Regression:
w0 (Intercept): 0.999877 , w1 (Slope): -2.41086

Metrics:
Accuracy: 0.784553
Sensitivity: 0.695652
Specificity: 0.862595
Program ended with exit code: 0
```

In the output for logistic regression we can see the coefficients generated for the boundary line and create an equation of y = -2.41086x + .999877 where y is the log odds of a passenger surviving and x is the sex of the passenger. The model also outputs three metrics based on the confusion matrix, which are accuracy, sensitivity, and specificity. The accuracy, which is .78 indicates that the algorithm correctly identified 78% of the cases in the test data set. The sensitivity, which is the true positive rate of the model is around 70% and the specificity, which measures the true negative rate is at 86%. This is indicative that the model is a decent classifier.

The Naive Bayes model also uses the same data set as logistic regression for training. The output shows the prior probabilities as well as the first few probabilities from the predictor.

```
Opening titanic.csv File
Header: "","pclass","survived","sex","age"
Prob No: 0.61
Prob Yes: 0.39
Likelyhood for p(pclass|survived)
Pclass     1  2  3
survived
   no
   yes
0.172131 0.22541 0.602459
0.416667 0.262821 0.320513
Likelyhood for p(sex|survived)
Sex        0  1
survived
   no
   yes
0.159836 0.840164
0.679487 0.320513
Age mean:
Not survived: 30.4182
Survived: 28.8261
Age variance:
Not survived: 14.2938
Survived: 14.4159
Time taken to train the model with the training data: 4 milliseconds.
Evaluating Model
            [0]                 [1]
[    0]   0.42069          0.57931
[    1]  0.793882         0.206118
[    2]  0.871101         0.128899
[    3]  0.225887         0.774113
[    4]  0.145994         0.854006
[    5]  0.167164         0.832836
[    6]  0.890018         0.109982
[    7]  0.867962         0.132038
[    8]  0.883091         0.116909
[    9]  0.788101         0.211899
[   10]  0.675258         0.324742
```

The first metric shown is the Apriori probabilities of survival based on the training data. The probability for not surviving is at 69% while survival is at 31% based on the data. The model than also calculates the conditional probabilities for the predictors of passenger class, age, and sex, and combines them with the Apriori to make predictions. In the first entry, it is shown that the probability of survival is at 58% while the probability that the passenger does not survive is at 42%.

Generative vs. Discriminative Classifiers

Classification algorithms like logistic regression and Naïve Bayes fall into two categories called discriminative and generative. Both of the algorithms share the similarity in how they calculate conditional probability to map data to classification labels. The difference is in how the algorithms calculate the conditional probabilities. A discriminative algorithm calculates a boundary line of conditional probabilities to separate data into two classes, while generative algorithms typically rely on Bayes' theorem to find joint probabilities (https://medium.com/@mlengineer/generative-and-discriminative-models-af5637a66a3).

Both categories of classification are forms of supervised learning. An example of a generative classifier is Naïve Bayes, while a discriminative classifier is logistic regression (https://towardsdatascience.com/generative-vs-discriminative-classifiers-in-machine-learning-9ee265be859e). Generative models typically are more sensitive to outliers because the probability distribution can be greatly altered. Generative models typically require less data to train than discriminative classifiers Discriminative models may be easier to train but may not work on data where the distribution is complex (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.). Both the algorithms have advantages and disadvantages and the best one must be picked depending on the use case.

Reproducible research is an important aspect of machine learning research because it is used to help researchers confirm the results of a study and avoid creating research that has already been performed. The National Science Foundation (NSF) defines research as "the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator". The importance of reproducibility is that it not only helps ensure the correctness of a study, but also builds confidence in what is done to produce the results. Currently in the field of machine learning, there are many barriers to the reproducibility of research, which include source code being kept as confidential. The implications of not having reproducibility in ML research is that unvalidated research has the ability to spread biases in other research which can spread inaccuracies in results and lead to bad results (https://blog.ml.cmu.edu/2020/08/31/5-reproducibility/).

To ensure research is reproducible in machine learning, it should be considered from the very beginning of the project. Research should produce clear documentation from day one and should explain the steps taken and elaborate on why steps were performed to create a successful output. This pattern of documentation should continue throughout the entire duration of the project (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). One focus of producing reproducible research is to ensure that the codebase is open and follows a pipelining checklist. This is to ensure that data information is not missing or omitted from the documentation, which is one of the leading

causes of research being non-reproducible (https://towardsdatascience.com/reproducible-machine-learning-cf1841606805).