Joshua Durana

ML From Scratch

CS 4375.003

Outputs

```
--Logistic Regression Coefficients--
Intercept: -2.41085
Sex: 0.999869
--Metrics--
Accuracy: 0.787755
Sensitivity: 0.695652
Specificity: 0.869231
Algorithm Run Time: 0.453784 seconds



--Summary--
A-Priori Probabilities: 0.61 0.39

--Conditional Probabilities--
Class
0: 0.172131 0.22541 0.22541
1: 0.416667 0.262821 0.262821

Sex
0: 0.159836 0.840164
1: 0.679487 0.320513

Age
Mean: 30.4182 28.8261
Variance: 205.153 209.155

--Naive Bayes Metrics--
Accuracy: 0.784553
Sensitivity: 0.695652
Specificity: 0.862595
Algorithm Run Time: 0.0004653 seconds
```

B.

Both algorithms have very similar results and predictions from the Titanic data. Both the algorithm's accuracy, sensitivity, and specificity are very similar. The biggest difference between them is their run time. The Logistic Regression has a longer run time since Gradient Descent needs hundreds of iterations to get the coefficients. For Naive Bayes, it doesn't need multiple iterations to calculate the probabilities.

C.

Discriminative classifiers estimates P(Y|X) from the training data. They're given the data and use it to calculate the coefficients needed for the model. The classifier will output a model that will calculate the probability from a given set of predictors.

Generative classifiers estimates parameters for P(Y) and P(X|Y). P(X|Y) is estimated from a model, then applies Bayes Rules for the predictions.

Sources

- https://www.cs.cmu.edu/~epxing/Class/10701-11f/Lecture/lecture4-annotated.pdf\
- https://www.cs.toronto.edu/~urtasun/courses/CSC411_Fall16/08_generative.pdf

D.

Reproducibility in science is important because it allows us to figure out to confirm other researcher's findings. It allows researchers to prove whether a finding is valid by redoing the experiment. It's also important because it increases the reliability of an experiment. In machine learning, it's through using the same datasets and methods and trying to get the same results.

There are many reasons why reproducibility is a huge problem in machine learning. One big problem is that not a lot of scientists are sharing their code. Another aspect is that there are many random occurrences that can happen such as datasets being shuffled at initialization or changing floating-point values due to changes to hardware or software changes. Changes in libraries and software can make reproducing a model much harder.

To alleviate these problems, there are many methods to improve reproducibility. You need to describe how the algorithm works and its complexity analysis. You need to provide the source code and a sample of the data set used. Describing the hardware and software that were being used when the model was created. These are a small number of things to improve reproducibility.

Sources

- https://blog.ml.cmu.edu/2020/08/31/5-reproducibility/
- https://neptune.ai/blog/how-to-solve-reproducibility-in-ml