Kyle Chan kxc180021, Ryan Banafshay rcb170002

ML Algorithms from Scratch

**Logistic Regression**

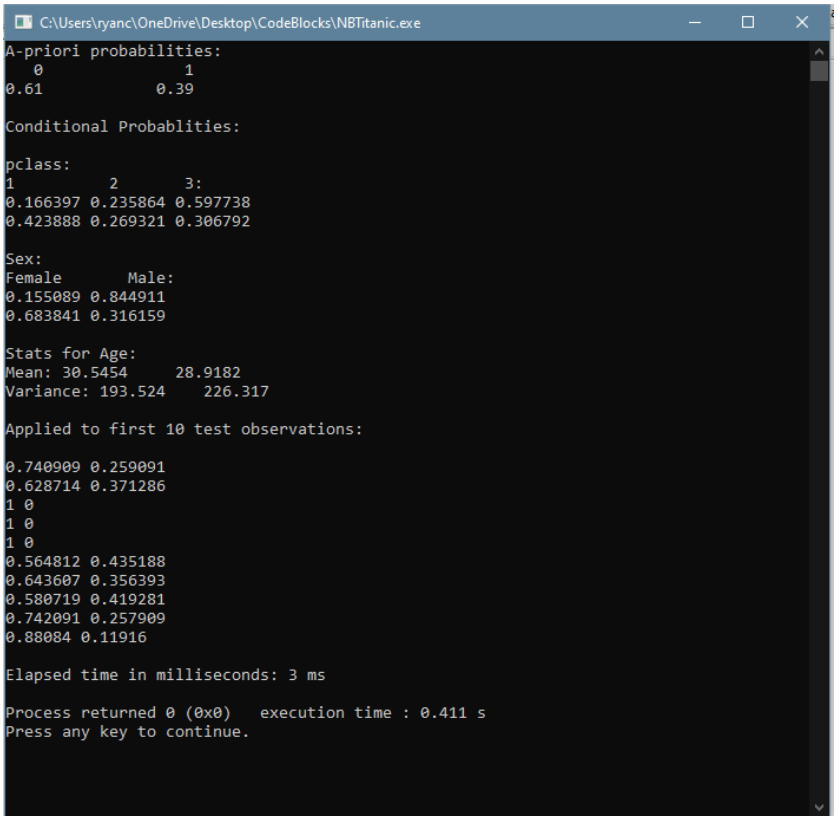
```
fireb@DESKTOP-TVV6QUH MINGW64 ~/Deskto
$ ./outfile
Covariance: -0.127856
Correlation: -0.538
Time to train: 85 ms
Weight: -1.41099

Test Data Stats:
Men survived: 80
Men dead: 19
Women survived: 35
Women dead: 113

Correctly Predicted Survived: 80
Correctly Predicted Dead: 113
Incorrectly Predicted Survived: 19
Incorrectly Predicted Dead: 35

Sensitivity: 0.695652
Specificity: 0.856061
Accuracy: 0.781377
```

We used sex as the only predictor for whether a passenger survives in our linear regression model. The weights vector is affected by the gradient descent function run through 500 epochs. The gradient vector identifies the line with the best logistic fit, from which we have determined a coefficient of -1.41. In other words, the likelihood of survival decreases by log odds -1.41 when the value of sex increases by a factor of 1. This explains why our model anticipates the survival of all men. According to the test data, only 80 out of 109 men in total survived. Since sex is the only predictor in this model and men in the test data are more likely predicted to survive, all women were predicted to have died. In reality, 35 out of 148 women survived in the test data.

Our linear regression model accurately predicted over 78% of survival observations in the test data. The accuracy of our model was measured by the proportion of times it accurately predicted whether a passenger would live or died. Calculating the true positive and negative rates is done via sensitivity and specificity, respectively. Overall, our model had a higher true negative rate.

Kyle Chan kxc180021, Ryan Banafshay rcb170002

**Naïve Bayes**




We built a Naïve Bayes analysis of the data from scratch. We obtained the final values by using a-priori probabilities based on survival data, conditional probabilities for passenger class and sex, as well as the mean and variance of ages. The conditional probabilities for the qualitative data suggests that Females had a higher survival probability than Males, and that passengers in Class 3 had the highest probability of mortality, though not by much. Since age is quantitative, we had to use mean and variance values to calculate probabilities. We saw the average death was at age 30 and average survived age was 28, not to significant of a difference between the two. We did see a high variance though in the ages which suggests that there were a large number of standard deviations.

Finally, we tested the values on the train data to obtain the final probability matrix for a Naïve Bayes analysis. Based on our results on the test data, we can see that our predictors suggest a high probability of belonging to class 0, which is the dead class in the survival matrix. This was consistent throughout multiple iterations of observations.

**Generative Classifiers vs. Discriminative Classifiers**

A generative classifier will provide a probability for a date set while a discriminative classifier will make predictions on unseen data sets. How each uses the data provided determines how the two differ from each other. Discriminative classifiers identify the “prior probability and likelihood probability to calculate a probability” after receiving the data and basing their prediction on the training data (Goyal). Depending on the kind of data set, generative and discriminative also differ from one another.

Performance-wise, generative models are better as they can make strong assumptions from much less information compared to discriminative models. In contrast, discriminative classifiers can produce predictions even in the presence of incomplete data while generative classifiers must omit or fill in missing data.

**“Reproducible Research in Machine Learning”**

In broader terms, reproducibility is the ability for researchers to run multiple tests with similar results to ensure correctness in findings. More specifically in Machine learning, reproducibility is the ability to recreate a system workflow or algorithm to reproduce the same results as the original work. This is significant as when you design a machine learning algorithm, if it isn’t built with reproducibility in mind, it can make your work much more difficult or unlikely to be implemented in the future or by other computer scientists. This wastes both money and time as new machine learning algorithms will have to be constantly developed to accomplish goals that have already been addressed by previously made algorithms.

Reproducibility can be achieved through several different means. The most impactful is to keep proper documentation of your model that thoroughly details every important aspect. If proper records are kept at every step of implementation, it makes recreate the model relatively simple. Automating the process of creating your algorithm can also contribute to reproducibility. Many researchers also suggest taking the initiative to take a community approach to this, such as submitting to and reviewing journals that support reproducible research and employers hiring candidates with reproducible work. Researchers stressing these tactics shows how important they seem to believe this issue is.

**Sources**

**https://towardsdatascience.com/reproducible-machine-learning-cf1841606805**

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

**https://staff.washington.edu/rjl/pubs/cise12/CiSE12.pdf**