Mason Cushing

CS 4375.004

Professor Karen Mazidi

March 4th, 2023. Repository at https://github.com/mta825/Cushing_Portfolio_CS4375

Reproducible Research and Classification Models

Logistic Regression

```
HEADING: "","pclass","survived","sex","age"

FINAL WEIGHTS: 1.00635 -2.40586

CONFUSION MATRIX
113 35
18 80

ACCURACY: 0.784553
SENSITIVITY: 0.816327
SPECIFICITY: 0.763514
```

Naïve Bayes

```
HEADING: "","pclass","survived","sex","age"

INITIAL LIKELIHOODS FOR PREDICTORS:
SEX VERSUS SURVIVAL
0.159836 0.840164
0.679487 0.320513

PASSENGER CLASS VERSUS SURVIVAL
0.172131 0.22541 0.602459
0.416667 0.262821 0.320513

AGE MEAN AND STANDARD DEVIATION:
30.4182 28.8261
14.3231 14.4622

OVERALL SURVIVAL PRIORS:
0.61 0.39

PRIORS FOR THE PREDICTORS IN THE TRAINING DATA:
0.3625 0.6375
0.2675 0.24 0.4925

CONFUSION MATRIX
113 35
18 80

ACCURACY: 0.784553
SENSITIVITY: 0.816327
SPECIFICITY: 0.763514
```

Above are pasted runs of the code linked near this project. My first observation for this data is the identical confusion matrices. Despite using different models and a different subset of predictors, our predictions for the data points being used for testing were exactly identical in their end result, leaving our accuracy at 78.4553%, our sensitivity at 81.6327%, and our specificity at 76.3514%.

The predictors for the logistic regression were approximately an intercept of 1 and a weight of negative 2.4 on sex, which was the only predictor used. This algorithm would simply predict that the person survived if sex equaled zero and that they died if sex equaled one. This actually resulted in a decent accuracy for the problem.

When the second model was used however, the same results were obtained, regardless of the fact that the other predictors were being used. This means that these predictors were being effectively ignored. This is very unusual for these models, and I would have expected that some differing results would have been obtained.

The difference between these two models lies in their nature. Logistic regression is an example of a discriminative classifier, and Naïve Bayes is an example of a generative classifier. According to Yıldırım, “Generative classifiers learn the joint probability distribution … to be able to explain how the data is generated” while “Discriminative classifiers try to find boundaries that separate classes.” In other words, discriminative classifiers focus more on finding the dividing between the data points and the classes it is told to distinguish, while generative classifiers reason that since most of the time the data points with these features (or values in these ranges) were classified this way, that this data point should also be classified this way.

The two classifiers are similar in that they both are used to classify two (and only exactly two) classes of data without being given the classification, and they require data to train in which the label is given to them. They are very different in the actual math behind how they work, but as seen in my above data, they can arrive at very similar results due to their design.

Reproducible research in machine learning is, as LeVeque, Mitchell, and Stodden put it, making research results "available in such a way that published computational results can be conveniently reproduced." This is important because without making results reproducible, they cannot be confirmed by others in the field. Peer review, as this confirmation is called, is a vital component of research and science, as it makes any research that you do checked by others in your field to avoid errors. Reproducibility can be implemented fairly easily by directly showing every step in the calculations of the research in a report or, if applicable, show the code used to create the output being used in the research.

However, as Hemant notes, "Cost and budget constraints are another area impacted by reproducibility. Without the details, (…) adopting new algorithms can run into huge costs and considerable research effort, only to lead into inconclusive results." Without knowing how the output for the research was created, considerable resources might be wasted. Furthermore, some machine learning applications take a very long time to train, and the cost of reproducing the output is a vital point to know before someone may continue your research.

Works Cited

Hemant, P. (2020, April 7). *Reproducible machine learning*. Medium. Retrieved March 4, 2023, from https://towardsdatascience.com/reproducible-machine-learning-cf1841606805

LeVeque, Mitchell, and Stodden. *Reproducible research for scientific computing: Tools and strategies ...* (n.d.). Retrieved March 4, 2023, from https://staff.washington.edu/rjl/pubs/cise12/CiSE12.pdf

Yıldırım, S. (2020, November 14). *Generative vs Discriminative classifiers in machine learning*. Medium. Retrieved March 4, 2023, from https://towardsdatascience.com/generative-vs-discriminative-classifiers-in-machine-learning-9ee265be859e