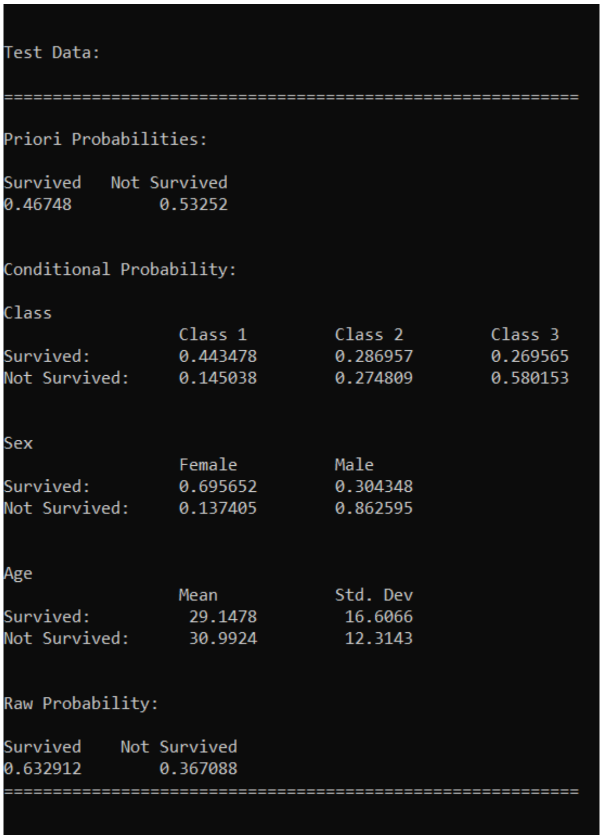
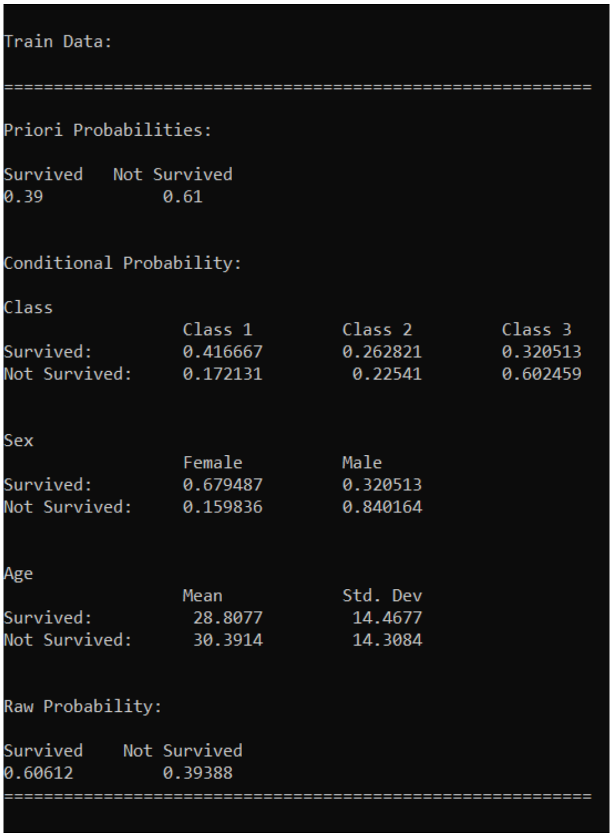
Umar Ali-Salaam
Cory Pekkala

ML Algorithms from Scratch

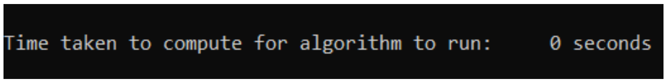
a) <u>Naive Bayes:</u>



```
Test Data:

===========================================================

Priori Probabilities:

Survived   Not Survived
0.46748         0.53252


Conditional Probability:

Class
                  Class 1         Class 2         Class 3
Survived:         0.443478        0.286957        0.269565
Not Survived:     0.145038        0.274809        0.580153


Sex
                  Female          Male
Survived:         0.695652        0.304348
Not Survived:     0.137405        0.862595


Age
                  Mean            Std. Dev
Survived:          29.1478         16.6066
Not Survived:      30.9924         12.3143


Raw Probability:

Survived    Not Survived
0.632912        0.367088
===========================================================
```



```
Train Data:

============================================================

Priori Probabilities:

Survived   Not Survived
0.39            0.61


Conditional Probability:

Class
                  Class 1         Class 2         Class 3
Survived:         0.416667        0.262821        0.320513
Not Survived:     0.172131         0.22541        0.602459


Sex
                  Female          Male
Survived:         0.679487        0.320513
Not Survived:     0.159836        0.840164


Age
                  Mean            Std. Dev
Survived:          28.8077         14.4677
Not Survived:      30.3914         14.3084


Raw Probability:

Survived    Not Survived
0.60612         0.39388
============================================================
```



```
Time taken to compute for algorithm to run:     0 seconds
```

b) The a priori probabilities are surprisingly different for the test data, 0.46748 for survived and 0.53252 for not survived, compared to 0.39 and 0.61 for train data. I think that's down to the sample sizes of both sets, if it were say 10k or more data points I think the values would be more similar.

The conditional probability of class shows a higher probability of surviving if you were in passenger class 1. Classes 2 and 3 don't have much of a difference in survivability, but are lower than class 1 in both the test and train data.

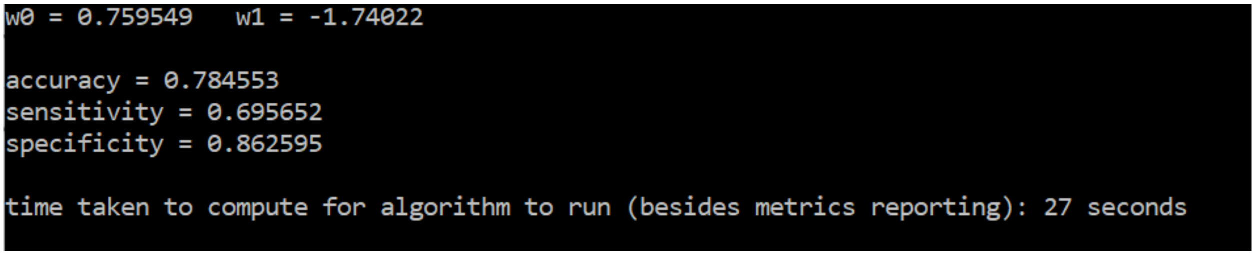
The conditional probability of sex shows a higher probability of women surviving over men (0.696 to 0.304) in the test data. The conditional probability of sex shows a higher probability of women surviving over men (0.679 to 0.320) in the train data. That's a pretty significant difference between the two probabilities, that likely shows the priority of the people on the ship, women were a higher priority.

The conditional probability of age shows that the mean age of survivors was slightly lower than those who didn’t survive (29.1 to 31) for test data, and (28.8 to 30.4) for train data. The difference isn’t very much, but the mean is consistent between both the train and test data. The standard deviations surprisingly varied between the two sets of data (16.6 and 12.3) for test data and (14.5 and 14.3) for train data. Again, I think that points to the small size of the data, as to why there’s differences between the two data sets.

The raw probability was calculated by using the Gaussian or Normal Distribution equation using both variance and mean. The entire test and train data sets were run through the equation. The raw probabilities were close in value of surviving and not surviving at (0.633 and 0.367) for test, (0.606 and 0.394) for train. This shows that the data for surviving is more frequent around the mean, than not surviving, given the size of each value.

Logistic Regression

a)



```
w0 = 0.759549   w1 = -1.74022

accuracy = 0.784553
sensitivity = 0.695652
specificity = 0.862595

time taken to compute for algorithm to run (besides metrics reporting): 27 seconds
```

b)

The intercept was 0.759549, the weight on the independent variable (sex) was -1.74022. These were the weights to minimize the error between the sex and survived training data in gradient descent.

Accuracy was 0.784553, this is the result of the ratio between the sum of our model's correct predictions and all of the predictions it had made [(tp+tn)/(tp+tn+fp+fn)]. So, roughly 78.46% of the time, when shown the test observations of the sex feature, our model was correctly predicting the contents of the test observations of the survived feature.

Sensitivity was 0.695652, this represents how correctly our model was able to predict based on the test proportion of the sex feature, when the test proportion of survived feature contained a success (1). So for roughly 69.57% of successes within the survived test proportion, our model was correctly identifying them as a success (1) and for about 30.43% of that portion, there were incorrect predictions.

Specificity was 0.862595, this represents how correctly our model was able to predict based on the test proportion of the sex feature, when the test proportion of survived feature contained a failure (0). So for roughly 86.26% of failures within the survived test proportion, our model was correctly identifying them as a success (0) and for about 13.74% of that portion, there were incorrect predictions.

All in all, our model was better at predicting failures (0) than successes (1).

c) Generative Classifiers, try to generate data based off of the individual classes, and determine which class is more likely to correlate or produce the given observation. An example of this would be an AI image generator that tries to make an image based off of user input. For instance, if the user wants an image of a spaceship it'll try to generate what it believes is an image of a spaceship. Discriminative Classifiers are essentially the inverse of Generative Classifiers, they try to determine what the actual classes are in a data set based off of the input data. An example of this would be when a Tesla car attempts to recognize its surroundings, it takes in loads of images and tries to determine what each object in the image is. For instance it can recognize if a stop light is green, yellow, or red based on how it looks, and relays that information back to the driver and the car.

   Both models have to go through a training phase, in which they have to generate or determine a class based off of their input data. Both models use conditional probability for their classification. Both models are making inferences and assumptions based on the input data, which means they won't always be 100% accurate, but they can always evolve and improve by increasing the amount of input data. As a final note it is found that Discriminative Classification usually performs better than Generative Classifiers, according to Andrew Ng.

Resources

automaticaddison, Author. "The Difference between Generative and Discriminative Classifiers." Automatic Addison, 19 Aug. 2019, https://automaticaddison.com/the-difference-between-generative-and-discriminative-classifiers/.

Ng, Andrew, and Michael Jordan. "On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes." Advances in Neural Information Processing Systems, 1 Jan. 1970, https://papers.nips.cc/paper/2001/hash/7b7a53e239400a13bd6be6c91c4f6c4e-Abstract.html.

Malhotra, Akanksha. "Generative Classifiers v/s Discriminative Classifiers." Medium, Medium, 16 Oct. 2019, https://medium.com/@akankshamalhotra24/generative-classifiers-v-s-discriminative-classifiers-1045f499d8cc.

d)

As artificial intelligence grows into the mainstream of tech and other industries, the trust in machine learning algorithms has become a center stage concern. Any algorithm that lacks the ability to be verified is risky to use as “unintended consequences” may arise (Hemant). There is a term for this concern in machine learning called *reproducibility.*

In general, and in machine learning, “reproducibility [refers to] being able to recreate a … workflow to reach the same conclusions as the original work” (Hemant). This is a critical component to many fields that deal with experimentation and analysis of data because if the results cannot be recreated, then the truth cannot be verified. For example, mathematics and the sciences operate in an environment built upon the trust of previous researchers who have discovered some truth in their respective universes of discussion. Trust is placed in those researchers that they had correctly diagnosed, interpreted, and analyzed their findings, so that new researchers can iterate upon their work (Heil, et. Al). Machine learning has become one of many tools employed by researchers to conduct their work, because unlike traditional, simpler, models, “machine-learning models are .. well suited to cope with the scale and complexity of [various scientific] data” (Heil, et. Al).

Although the problem of reproducibility is multifaceted, “missing information” is cited as one of the foremost “root cause[s] of all reproducibility issues” (Hemant). In many cases, while configuration information, such as the values of hyperparameters in an experiment might be disclosed, the “data, models, and code [are not always made] publicly available and usable by other” researchers (Heil, et. Al). This is a solvable problem, but a collective effort, in that it's the “scientific culture surrounding computational work…[that is making] it difficult to verify findings [and] efficiently build upon past research” (Stodden, et. Al).

Nonetheless, concerted efforts such as “record[ing] every step taken in building a model through documentation and version control” can help to reduce the effects of this problem. If enough researchers do this, then information-loss can be reduced, and could lead to more reproducible models and results.

Resources

Hemant, Preeti. “Reproducible Machine Learning.” Medium, Towards Data Science, 7 Apr. 2020, https://towardsdatascience.com/reproducible-machine-learning-cf1841606805.

Heil, Benjamin J., et al. “Reproducibility Standards for Machine Learning in the Life Sciences.” Nature News, Nature Publishing Group, 30 Aug. 2021, https://www.nature.com/articles/s41592-021-01256-7.

Stodden, V, et al. Setting the Default to Reproducible: Reproducibility in Computational and Experimental Mathematics. http://stodden.net/icerm_report.pdf.