Machine Learning Classification From Scratch in C++

Zachary Canoot and Gray Simpson

Runs of Code

Logistic Regression

```
Run the 'Sex' Predictor Scenario at 600 iterations
Coefficients:
Intercept: 0.999877
Coefficient 1: -2.41086
Accuracy: 0.784553
Sensitivity: 0.695652
Specificity: 0.862595
Training Time: 290ms
Confusion Matrix:
P. -> 0 1
0 113 35
1 18 80
```

Naive Bayes

```
Naive Bayes Machine Learning from Scratch in C++
Attempting to open file 'titanic_project.csv',
File opened successfully.
Reading heading: "","pclass","survived","sex","age"
Length is now 1046
Now closing file 'titanic_project.csv'
Head of data:
Survived: 0 Class: 3 Age: 19 Sex: 1
Survived: 1 Class: 3 Age: 22 Sex: 0
Survived: 1 Class: 3 Age: 20 Sex: 1
Survived: 0 Class: 3 Age: 1 Sex: 0
Survived: 0 Class: 2 Age: 63 Sex: 1
Survived: 0 Class: 1 Age: 38 Sex: 1
```

```
Survived: 0 Class: 3 Age: 19 Sex: 1
Survived: 0 Class: 2 Age: 39 Sex: 1
Survived: 1 Class: 3 Age: 17 Sex: 0
Survived: 0 Class: 3 Age: 3 Sex: 0

The Naive Bayes calculation took 1 milliseconds.

Here are the data values and metrics.

Survived: 312 prob 0.39  Dead: 488 0.61
Age Died Mean: 30.3914       Age Survived Mean: 28.8077
Age Died Variance: 205.15   Age Survived Variance 209.989

Pclass Likelihoods
     0     1     2
0  0.172131  0.22541  0.602459
1  0.416667  0.262821  0.320513

Sex Likelihoods
     0     1
0  0.159836  0.840164
1  0.679487  0.320513

Confusion Matrix
     0     1
0  113  35
1  18  80

Accuracy: 0.784553
Sensitivity: 0.695652
Specificity: 0.862595


Process returned 0 (0x0)   execution time : 0.075 s
Press any key to continue.
```

Result Comparison

One thing we can notice very quickly from both algorithms' runs is that they have the same confusion matrices, which leads to the same accuracy, sensitivity, and specificity. Though we took very different methods to get to our results, one involving a slow-train and slowly gaining more, neither ended up "better" than the other, in this instance, though Naive Bayes was certainly faster.

Considering we trained off of small dataset of 800 observations, as data increases, this result could change. However, in our instance, given the same data, both Logistic Regression and Naive Bayes had the same results, despite the fact that Logistic Regression based this data solely off of sex and Naive Bayes off of a combination of sex, age, and passenger class.

To look at Naive Bayes more closely, we can see that the passenger class and sex had strong effects on the results, sex the strongest, while data collected for age showed that they were more similar between those who died and lived. On the other hand, Logistic Regression considers the sex of each observation/person, and produced the a model dividing the data based on the probability that someone survived given their gender. It was the only predictor so it was the *only* impact on the model. The fact the results then ended up the same may hint towards an error, or a very manicured dataset. Whatever the case they are both accurate!

To note on our high specificity, we would argue that when deciding how likely you are to die on the Titanic, we'd prefer to know how likely we are to die. In that case, a high specificity is a good thing. That is because it is the amount of true negatives out of total negatives. Who knows, there is a Titanic 2 sailing soon. Guess they should fill the whole boat with men based on this model!

Generative Classifiers vs Discriminative Classifiers

Generative classifiers and discriminative classifiers are different ways of meeting the same goal, completing a classification model. The goal of a classification model is to estimate the probability of a certain outcome of Y given X, P(Y=1|X). Generative models find the probability of Y and the conditional probability of X given Y. This *generates* P(Y|X) given P(Y) and P(X|Y). You can also instead use discriminative classifiers that simply assume some form of P(Y|X) and fit their assumption to the data using training data, in that way it *discriminates* by separating the data.

It is easier to understand in the context of this assignment. Logistic Regression is Discriminative while Naive Bayes is Generative. Naive Bayes makes a set of probabilities that, given any data, can generate a probabilistic classification. Compare that to logistic regression, that fits a function that divides the data into a classification. It follows that Naive Bayes would be better at producing a model to make new data that fits the model, while Logistic Regression is good at labeling data.

Sources

(1) Mazidi, K. (2020). *Machine learning Handbook Using R and Python* Creative Commons.

(2) https://www.analyticsvidhya.com/blog/2021/07/deep-understanding-of-discriminative-and-generative-models-in-machine-learning/ (Educated most of my description)

Reproducible Research In Machine Learning

Reproducible research is research that can be done multiple times with the same or similar results, despite whether it is done by the same person who initially created it or someone else entirely. It is important most largely for the proper understanding of data (1). If it can be understood, both the data and the efforts taken to get that data can be best known, and then compared with new methods of investigation.

It is also important due to being able to validate research results between different people and over the years (2). Reproducible research gives us both transparency and confidence, so that when we explain and are specific with what we do, the same processes can be exactly replicated (1). In theory, this should be simple with computer science, since source code is more often available and computers are overwhelmingly reliable, but somehow this isn't always the case (2).

Between random sampling, hidden neural network layers, unseeded data separation, and non-deterministic calculations, there are many causes of differences in computer science program reproduction (1). This can be helped by careful documentation, but without universal methods and standards, there is no way to ensure perfect reproducibility. Clear descriptions of algorithms with complexity and sample sizes, readily available source code, clear descriptions of data collection, ranges of hyperparameters, clear definitions of statistics used, and even the computer's infrastructure are all good sorts of documentation to bring us closer. In fact, it is becoming more common to require the code and data for it to be given when submitting experiments done on a computer(2).

Sources

(1) https://blog.ml.cmu.edu/2020/08/31/5-reproducibility/
(2) https://scholar.google.com/scholar_url?url=https://www.jmlr.org/papers/volume22/20-303/20-303.pdf&hl=en&sa=X&ei=GgA6Y7K5IIjaygS29LHABw&scisig=AAGBfm2VFR-Pd9ZP5Pzjz10KkPAhMQk4rw&oi=scholarr
(3)