Name: Bushra Rahman
Class: CS 4375.004
Assignment: ML Algorithms from Scratch

Logistic Regression & Naive Bayes From Scratch

The purpose of this program is to perform logistic regression and Naive Bayes on a predetermined set of training data, and output the resultant metrics. Here is an example showing the code output:

```
Opening file titanic_project.csv
Closing file titanic_project.csv

First coefficient w0 is 0.999877
Second coefficient w1 is -2.41086
Execution time for training data was 12943070 microseconds.
_____________________________________________________________
Apriori probabilities are 0.61 and 0.39

Conditional probabilities for pclass are:
0.172131 0.22541 0.602459
0.416667 0.262821 0.320513

Conditional probabilities for sex are:
0.159836 0.840164
0.679487 0.320513

Means of age for not survived and survived are:
18.5551 11.2422

Variances of age for not survived and survived are:
210.734 201.895
Execution time was 9629 microseconds.
```

Analysis of the Results

The algorithms from scratch performed fairly well on the Titanic test data. Both the logistic regression algorithm and the Naive Bayes algorithm in C++ worked nearly as well as the R implementation, producing nearly the same coefficients and probabilities for both. However, there was still some discrepancy. Additionally, the C++ implementation was notably time-consuming, in both the raw implementation as well as the execution time. The C++ implementation took many lines of code and many microseconds to do what an R implementation could do much faster.

Generative vs. Discriminative Classifiers

Generative and discriminative classifiers describe two different types of classification algorithms. In generative classifiers, the probability is modeled from the likelihood P(X|Y) for the predictor X and target Y. Then, using the Bayes Rule, the posterior P(Y|X) is calculated [1]. This is reflected in Naive Bayes,

which is a generative classifier that estimates parameters for likelihood P(X|Y) and prior P(Y). Meanwhile, in discriminative classifiers, the posterior is modeled directly, as seen in logistic regression where the parameters for P(Y|X) are directly estimated [1, 2].

These methods reflect the different usages of the two algorithms: generative classifiers focus on how the predictor and target variables occur together, while discriminative classifiers focus on finding the boundaries that separate classes in the data [2]. Additionally, Naive Bayes tends to perform better on small amounts of data, while logistic regression is not as sensitive to the size of the data [2]. However, one similarity between Naive Bayes and logistic regression is that they are both high-bias low-variance algorithms. In terms of their internal logic, logistic regression goes through an iterative optimization process, while Naive Bayes needs to do only one pass over the data to estimate its parameters.

Reproducible Research in ML

Reproducible research refers to research whose results and means of producing those results can be replicated by other researchers using different data [3, 4]. Reproducibility is important because it strengthens the accuracy of the results and shares the methods involved with the greater scientific community, thus ensuring transparency and peer reviewability [3]. Reproducibility is generally considered to be the norm in scientific research, but machine learning in particular is currently experiencing a reproducibility crisis [3, 4]. This crisis appears to be affecting not only machine learning, but the general computer science field [4]. Reproducibility begins with transparency, in sharing one's code and test data with the public. This can be achieved through means as simple and accessible as GitHub, a widely popular code repository system [4]. Other general tips for implementing reproducibility include documenting one's environment (including code, data, dependencies, hardware, etc.), clearly describing the ML algorithm being used and its particular time/space complexity, and incorporating opposing research from other sources so as to open up a space for discussion [3, 4].

Sources

[1] Ng, Andrew Y.; Jordan, Michael I. "On Discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes." [Online]. *Advances in Neural Information Processing Systems 14 (NIPS 2001)*, 2001, https://ai.stanford.edu/~ang/papers/nips01-discriminativegenerative.pdf. [Accessed: 3-Mar-2023].

[2] Yıldırım, Soner. "Generative vs Discriminative Classifiers in Machine Learning." [Online]. *Towards Data Science*, 14 Nov, 2020, https://towardsdatascience.com/generative-vs-discriminative-classifiers-in-machine-learning-9ee265be859e. [Accessed: 3-Mar-2023].

[3] Ding, Zihao; Reddy, Aniketh; Joshi, Aparna. "5 – Reproducibility" [Online]. *ML CMU*, 31 August, 2020, https://blog.ml.cmu.edu/2020/08/31/5-reproducibility/. [Accessed: 3-Mar-2023].

[4] Ramón Fonseca Cacho, Jorge; Taghva, Kazem. "The State of Reproducible Research in Computer Science" [Online]. *Advances in Intelligent Systems and Computing*, 12 May, 2020, https://link.springer.com/chapter/10.1007/978-3-030-43020-7_68. [Accessed: 3-Mar-2023].