Sovanna Ramirez & Sanjana Jadhav
CS 4375.004

ML Algorithms from Scratch

Runs of code showing the output (coefficients and metrics), and run times

- Logistic Regression:

```
Opening titanic_project.csv file
Reading Line 1
heading: "","pclass","survived","sex","age"
Number of Observations in Train Vectors: 800
Number of Observations in Test Vectors: 246
Closing titanic_project.csv file

Performing Logistic Regression:
Intercept: 0.999981
Sex Coefficiant: -2.41111
Time for Algorithm to Compute: 21 seconds

Using Test Data to Predict:
Accuracy: 0.784553
Sensitivity: 0.695652
Specificity: 0.862595
```

As we can see, it took the program 21 seconds to train and produce an intercept of 0.999 and a coefficient of -2.4111. There were several complex computations involved such as matrix multiplication, transposing a matrix, etc. When the test data was used to predict, the model resulted in an accuracy of 0.784553. Moreover, out of all positive cases, 0.695652 were correctly predicted. Out of all negative cases, 0.862595 were correctly predicted.

- Naive Bayes:

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

A-priori Probabilities:
1 (Alive): 0.39             0 (Deceased): 0.61

Conditional Probabilities:
p(pclass|survived)...
Pclass 1 & Alive: 0.416667  Pclass 1 & Deceased: 0.172131
Pclass 2 & Alive: 0.262821  Pclass 2 & Deceased: 0.22541
Pclass 3 & Alive: 0.320513  Pclass 3 & Deceased: 0.602459

p(sex|survived)...
Female & Alive: 0.679487    Female & Deceased: 0.159836
Male & Alive: 0.320513      Male & Deceased: 0.840164

p(age|survived)...
Mean Alive: 28.8261             Mean Deceased: 30.4182
Variance Alive: 14.4622     Variance Deceased: 14.3231

Training Time of Naive Bayes Algorithm: 0 s

Predictions on Test Data:
  [0]              [1]
[0]0.42088          0.57912
[1]0.793881         0.206119
[2]0.871139         0.128861
[3]0.226058         0.773942
[4]0.145902         0.854098
[5]0.165464         0.834536
[6]0.890094         0.109906
[7]0.867948         0.132052
[8]0.883218         0.116782
[9]0.788214         0.211786
```

With this model we start by finding the A-priori probabilities by taking the total count of survivors or deceased over the total count of observations within our train set. We see that the A-priori probability of survival was 39% and not survived (deceased) was 61%. Moving forward, we calculated our conditional probabilities of survival based on the different predictors we had pclass (1,2,3), sex (f-0, m-1), and age (variable). After coming up with our calculations, we could finally put it all together to utilize the Bayes Theorem to calculate the raw probabilities given pclass, sex, and age.

Sovanna Ramirez & Sanjana Jadhav
CS 4375.004

Generative classifiers versus Discriminative classifiers:

Logistic Regression is a Discriminative Classifier. In simpler words, the model tries to find a relationship between a given x and y. These classifiers are mostly used to distinguish boundaries such as survived/dead, old/young, etc. They are simpler to train and perform well with big datasets. They can be used for multi-class classification or binary classification. However, outliers in the data may affect the data as discriminative classifier algorithms tend to overfit the data.

Naïve Bayes is a Generative Classifier. This model tries to find the probability of each class given the data. They work better with smaller datasets. Depending on the dataset, this model could produce inaccurate results because some predictors may be dependent on others. Naïve Bayes assumes that all predictors are independent of each other. Generative Classifiers also work best when there is missing data, as they can "generate" new data based on given data.

Sources:

- ML Textbook (Class Notes)
- Discriminative Classifier
- Joint versus Conditional Densities
- Generative Model

Reproducible Research in Machine Learning

Reproducible research is the effort to duplicate the results of previous research using the same datasets, algorithms, tools, and procedures prior. The practice of reproducibility, does not always ensure the duplication of results which is why machine learning is experiencing a reproducibility crisis. It is understood that reproducibility is actually quite difficult to achieve in machine learning.

Reproducible research is important to machine learning because it helps to confirm findings and improve research rather than waste time repeating previous research. Moreover, data, and other factors as well, are always changing and it is essential that reproducibility can be achieved to reduce ambiguity in research findings. Researchers want to be able to confirm any new findings without getting confused with the data itself. Thus, reproducible research is important because it ensures that data stays consistent, and results are correct regardless of the changes of other factors.

Now, the implementation of reproducible research should be top of mind when starting any machine learning project. Ensuring reprodubility, can come from the quality of documentation – what steps were taken to successfully execute the project, plus any other specific and noteworthy details. Good documentation will be sure to include any environments, software, and approaches used when getting results. Moreover, reprodubility can be implemented through different types - methods reproducibility, results reproducibility, and inferential reproducibility.

Sovanna Ramirez & Sanjana Jadhav
CS 4375.004

MLA Citations

Ding, Zihao. "5 - Reproducibility." *Machine Learning Blog | ML@CMU | Carnegie Mellon University*, 24 Aug. 2020, https://blog.ml.cmu.edu/2020/08/31/5-reproducibility/.

"The Importance of Reproducibility in Machine Learning Applications." *DecisivEdge*, 7 Dec. 2022, https://www.decisivedge.com/blog/the-importance-of-reproducibility-in-machine-learning-applications/#:~:text=Reproducibility%20with%20respect%20to%20machine,reporting%2C%20data%20analysis%20and%20interpretation.

"Statistical Analyses and Reproducible Research." *Taylor & Francis*, https://www.tandfonline.com/doi/abs/10.1198/106186007X178663.