

ML Algorithms from Scratch
Shiyu Yan, (First Name Last Name)

a. Copy/paste runs of your code showing the output (coefficients and metrics), and run times

```
Microsoft Visual Studio Debug Console

Opening file titanic_project.csv.
Reading line 1:
Heading: ,pclass,survived,sex,age
coefficients: -0.88
coefficients: -1.55
coefficients: 0.0971258
coefficients: -2.51266
coefficients: 2.25571
coefficients: -2.8935
coefficients: 3.14607
coefficients: -4.03765
coefficients: 2.89304
coefficients: -5.6648
coefficients: 2.71203
coefficients: -6.59219
coefficients: 2.82841
coefficients: -6.62275
coefficients: 3.21443
coefficients: -5.76553
coefficients: 3.56346
coefficients: -4.2773
coefficients: 2.53579
coefficients: -3.4657
accuracy: 0.784553
sensitivity: 0.862595
specificity: 0.695652
RunTime: 0.0045215
```

b. Analyze the results of your algorithms on the Titanic data

In the logistic regression, accuracy 0.784553 means that the model is correctly predicting the outcome 78.46% of the time. Sensitivity 0.862595 means that it identifies 86.26% of the positive cases in this dataset. Specificity 0.695652 shows that it identifies 69.57% of the negative cases in this dataset. Total run time of the logistic regression in 10 echoes is 0.0045215.

```

A-priori probabilities:
    0    1
0.621818 0.378182
Conditional probabilities:
    pclass
          1          2          3
perished 0.172131 0.22541 0.602459
survived 0.416667 0.262821 0.320513
sex
    female    male
0 0.159836 0.840164
1 0.679487 0.320513
age
mean 30.4182 28.8261
std 205.153 209.155
TEST DATA
Predictions
perished    survived
0.8019370.198063
0.8766440.123356
0.2349190.765081
0.1522390.847761
0.1724780.827522
Sensitivity: 60%
Specificity: 91.6031%
Accuracy: 69.5652%
Time measured: 0.0028963 seconds

```

In my Naive Bayes model I got an accuracy of 69.5652 percent. This is not the best possible accuracy but overall, 70 percent accuracy is good. This means that 70 percent of the predictions are correct. The sensitivity is 60%. That shows that it can identify 60 percent of the survived cases accurately. The Specificity is 91.6 percent. This is a very good value and means that it can accurately identify 91.6% of the perished values in the data set. The total run time of the model was 0.0015 seconds which is very fast.

c. Write two paragraphs comparing and contrasting generative classifiers versus discriminative classifiers. Cite any sources you use.

Generative classifiers are based on modeling the probability distribution of the data for each class. If a new observation is given to the generative classifiers, it will predict the class that is most likely to have generated a new observation. Naive Bayes is a good example of generative classifiers. It directly estimates parameters for $P(Y)$ and $P(X|Y)$. The strengths of generative classifiers are hard to overfit because they model the probability distribution of the data. The weaknesses of generative classifiers require more data to train the probability distribution for each class.

Discriminative classifiers model the input features to predict the possible classes. It means that if a dataset has specific features, discriminative classifiers can identify other data with those features to predict the possible classes. Logistic regression is one of discriminative classifiers. It directly estimates the parameters of $P(Y|X)$. Discriminative classifiers are capable of handling

complex decision boundaries, making them flexible in predicting the possible classes. However, the bad points for discriminative classifiers are easier to overfit if the dataset has limited data.

Citation:

Yildirim, Soner. "Generative vs Discriminative Classifiers in Machine Learning." Medium, Towards Data Science, 14 Nov. 2020, <https://towardsdatascience.com/generative-vs-discriminative-classifiers-in-machine-learning-9ee265be859e#:~:text=Logistic%20regression%2C%20SVM%2C%20and%20tree,are%20examples%20of%20discriminative%20classifiers.>

d. Google this phrase: reproducible research in machine learning. Using 2-3 sources, at least one of which should be academic, write a couple of paragraphs of what this means, why it is important, and how reproducibility can be implemented. Cite your sources using any format.

Reproducible research in machine learning ensures that the algorithm can be executed multiple times on the same dataset, producing identical or comparable results. It is essential for ensuring the validity and reliability of research in machine learning, as it permits other researchers to test, identify and correct potential errors and biases, and develop more sophisticated research based on the findings. According to "*Reproducible Research for Scientific Computing: Tools and Strategies for Changing the Culture*" in this article, it described the concept of reproducibility in scientific research and highlighted the importance of reproducible research in machine learning. It also provides many examples of reproducible research in different fields, including genomics and astrophysics. The author believed that we need to strengthen the standards and guidelines for reproducibility and improve their implementation. For the implementation, we can use github or R notebooks to display code documentation and testing tracking.

Citation:

LeVeque, Randall J., Ian M. Mitchell, and Victoria Stodden. "Reproducible research for scientific computing: Tools and strategies for changing the culture." *Computing in Science & Engineering* 14.04 (2012): 13-17.

Heil, B.J., Hoffman, M.M., Markowetz, F. et al. Reproducibility standards for machine learning in the life sciences. *Nat Methods* 18, 1132–1135 (2021). <https://doi.org/10.1038/s41592-021-01256-7>.