Final Project Report

Github Repo: <https://github.com/eez9/C470-FInal-Project.git>

In the amazing world of AI and machine learning, the evaluation of the classification algorithm, is important to check for the efficiency in that they would have in the real world. I went ahead and dived into the evaluation of three different classification algorithms which were Naïve Bayes, Logistic Regression and KNN I attempted to train and test each algorithm on the dataset gathered from the spambase.csv file. That file contained features that were gotten from emails that were viewed either as spam or not spam. This includes some attributes like frequencies, character frequencies and some other metrics.

**Naïve Bayes:** This is an algorithm that is a probabilistic classifier based on the Bayes theorem and this one assumes independence amongst features. The implementation of Naïve Bayes requires the learning of prior probabilities and conditional probabilities of each of the classes from the training data. A prediction will then be made when the posterior probability is calculated for each class and selects the class that has the highest possible probability. For this test I attempted to calculate the Accuracy, true positives, false positives, and the area under the curve. The picture below shows all of the findings.

A screenshot of a computer

Description automatically generated

**Logistic Regression:** This algorithm is a linear classifier that shows the probability of a binary outcome. The implementation for logistic regression requires that the algorithm is trained using the gradient descent to assume the weight and bias parameters. A prediction would then be made by learning weights and bias from the input features that would pass them using the sigmoid equation. For this test I attempted to calculate the Accuracy, true positives, false positives, and the area under the curve, but unfortunately, I did not have that great of luck. I am not able to show any of the findings as there was a few errors in my code I couldn’t figure out. But on the github page you can see my attempt at the logistic regression algorithm.

**KNN:** This algorithm in a non parametric, algorithm and KNN classifies new instances based on the similarity to other existing instances inside of the training data. Again this algorithm requires that the number of nearest neighbors is specified In order to consider. Once that happens then a prediction can be made by having a majority vote amongst the number of nearest neighbors in the new instance. For this test I attempted to calculate the Accuracy, true positives, false positives, and the area under the curve, but unfortunately, I do not believe the outcomes were fully correct as both my true and false positives were at 0 and I had some really low accuracy and are under curve. In the picture below you can see my results for the KNN algorithm.

A screenshot of a computer

Description automatically generated

Overall learning about this algorithms and trying to implementing was extremely challenging but I did manage to have fun with it and I at least learned how the worked. I do look forward to figuring out how I can make these algorithms fully work with the email spam base data after the due date as I did find this project very interesting.