Inferences made and results - Vaidhyesh Padma Sundar

Two algorithms were run after the data was normalized to a range of [-1,1] and a Principal Component Analysis was performed. The data looked separable, suggesting that a classification model should be able to decently classify the data.

I tried to run 4 different algorithms, SVM, Naive Bayes, Neural network and Logistic regression. It was found that, SVM took forever to train, which makes sense because SVM works well on smaller datasets. With Naïve Bayes however, the runtime was quicker, but the accuracy was a meager 0.71.

Logistic regression on another hand, gave an overall 0.86 accuracy with 0.93 precision, 0.91 recall for class 0 and 0.47 precision, 0.52 recall for class 1. It shows that the prediction for class 0 is more accurate than predictions for class 1. This behavior is expected because the data was unbalanced. Just 4000 samples for Class 1 and 90000 samples for Class 0. To address this problem, I used a Random under sampler, which scaled the data of both the classes equally.

Then I ran the same preprocessed and cleaned data under a neural network with 8,8,8 hidden layers running with ‘relu’, the rectified linear unit activation function and ‘Adam’, a stochastic gradient-based optimizer. With this, I was able to achieve overall accuracy of 0.88 with 0.91 precision, 0.95 recall for class 0 and 0.54 precision, 0.37 recall for class 1.

The neural network has a 0.02 accuracy advantage over the Logisitic regression. Looking at the confusion matrices of Logisitic regression and neural networks, it is observed that the neural network has a higher “True Positive” value but has a lower “True Negative” value compared to Logisitic regression. Also, the “False Positive” of Neural Networks is lower than Logisitic regression suggesting that the neural network is going to work better at situations where it needs to detect the samples of Class 0 more accurately. However, Logisitic regression seems to work better for Class 1, which has fewer samples to begin with. I would therefore consider using Logisitic regression, when I need to classify samples with Class 1 more accurately.

I have attached the confusion matrices for logistic regression and neural networks below for the reader's reference.

Logisitic Regression

	Class 0	Class 1
Class 0	9112	888
Class 1	723	777

Neural Network

	Class 0	Class 1
Class 0	9523	477
Class 1	944	556