Report document for Assignment 2 CISC 471

Ngoc Nguyen – 20188794

Task 1.

1. Bar plot of data distribution for the 2 classes

Chart, bar chart

Description automatically generated

Task 2. 1. Research a high-level description of PCA method and how it is used for reducing the length of the input feature vector

PCA trades a little accuracy for simplicity, while reducing the number of variables of the dataset to make analyzing data much easier and faster. In high level term, this means reducing the dimension of the feature space. This also means dimensionality reduction. There are five main steps of the PCA process. Firstly, we will perform standardization ensure each feature contributes equally to the analysis. Secondly, we perform covariance matrix computation, this means analyzing the relationship between variables and their correlation to reduce redundant information. Thirdly, we compute eigenvectors and eigenvalues of the covariance matrix to identify the principal components. Fourthly, we assemble the feature vector, which contains columns the eigenvectors of the components we decided to keep. Finally, we recast the data along the principle components axes.

PCA 3D plot:

Chart, scatter chart

Description automatically generated

Task 3.

Task 3

Justification for learning rate, batch size, number of epochs, size of the hidden layer, number of hidden layers

The main focus for the ANN model is to keep the neural network simple and the learning rate low,

thus my ANN model only comprised of 3 hidden layers and one Dropout layer to avoid overfitting.

For the number of epochs, I tested with (10, 20, 50, 100) epochs, and consistently find that accuracy peaks around 10th-15th epochs and then lower afterwards.

With a smaller learning rate, I keep the batch size small as well. I tried common number batch size of (16, 32, 64) and finds that 16 minimizes overfitting the most, since the two other batch size at last epoch produces 1.0 accuracy.

7. How to ensure results are reproducible

Using random.seed() to ensure that all future replication of the process is generalized and no test run is biase

8. Which provides the best performance, why?

Decision Tree provides the best performance in terms of accuracy. With a dataset with such high dimension, DT is useful to identify the hidden pattern in the dataset. Furthermore, the dataset is not heavily processed and we have a small number of samples, thus DT outperforms the other model when working with a simplier input dataset.

9. Observations and conclusions

For Logistic Regression, Decision Tree, and Random Forest, the accuracy is lower with PCA data set compared with the full dataset. The reason why is because with a dimensionally reduced dataset, we loss a lot of information and complexity of the original dataset. Furthermore, our dataset is small and number of samples are limited, thus coupled with reduced dimensions really limited the features and predictive information that these models can use during training.

However, for Artificial Neural Network, the accuracy is higher with the dimensionally reduced data set compared to the full dataset.

Neural networks are effective at performing dimensionality reduction, thus coupling PCA before NN training can reduce a lot of the work of dimensionality reduction being done within NN and instead the model focuses on understanding the reduced dataset right away.