CSc 180 Project 2

Alec Resha 301201706 (Solo Project)

Due 10/14

Problem Statement

The problem being addressed in this project was classifying connections as an attack or not. Based on a dataset of information on connections along with whether they were an attack or not, two different ML models predicted whether an input was an attack or not. This can be expanded from a binary classification to a categorical one with relative ease. For a standard neural network, outputting an array of the types of attack with the likelyhood and taking the highest probability works well, and cnn could be adapted to give x outputs rather than just 2 for binary.

The data was structured as a standard TSV file. The largest issues with the raw data was to drop missing and duplicate rows, but requires very little other setup.

Methodology

To begin processing the data, two approaches are possible. Each row can be encoded to be a binary 1 for normal connection or 0 for an attack, the other option is to expand on the output and encode each type of attack separately as well. The first option gives a very accurate yes/no on whether a connection is an attack. The second option gives the ability to also output the predicted kind of attack with lower, but still very high, accuracy.
The next step is to normalize the numerical columns and encode the remaining text columns to make the data more compatible with ML models.
Once normalized and encoded, the data is split and fed to each kind of model.

Experimental Results and Analysis

For each model, I made 5 variations for the sake of runtime. In each one I varied a different aspect of it, including the number of layers, activation, optimizer, and for the CNN models the kernal size and count. Two of the top 5 models by RMSE value were Fully Connected Neural Networks, the third was a Convolutional Neural Network. The accuracy and averaged F1 reflected this as well.

Task Division and Reflection

This project was done solo. The largest challenge that I faced was getting the data into the CNN model. This was solved by checking the notes again on how to reshape matrices for CNN models. Besides this, setting up an initial CNN model was the most difficult since it was new, but rereading the notes and documentation helped immensely.

I learned a lot about the CNN model. I had a hard time following the process in the lectures, but after working out the data and model myself the process clicked.