Kexi's reflection

Kexi Huang

December 2019

1 Introduction to the inference goal and baseline model

1.1 Inference Goal

The inference goal of this project is chosen to do a multivariate classifier. We have classified the "normal" feature into 4 labels, and our goal is to analyze which type the given request is. Since different attack behave differently, for example, the "src_bytes" of "neptune" always appears as 0, while the "src_bytes" of "smurf" is in general 520, we can learn different features for different attacks.

1.2 Baseline Model

Decision tree is decided to be our baseline model for its simplicity and resemblance with most of our models. It is a traditional method for classification and it behaves good and steady in most conditions. Therefore, we want to see whether our more complicated models will have better behaviors than this simple and widely used baseline model, and whether it is worthy of spending more time on training and predicting process. Also, since decision tree can be the most traditional model of tree models, we want to see whether the numerical method like Multi-Layer Perceptron performs better in such classification tasks.

2 Introduction to the Multi-Layer Perceptron and comparison

Multi-Layer Perceptron (MLP) is a class of feedforward artificial neural network (ANN). It can be seen as a logistic regression classifier where the input is first transformed using a learnt non-linear transformation. There are generally more than three layers in the neural network, which are called input layer, hidden layer(s), and output layer.

The design inspiration of MLP is derived from the structure of the human brain. The neurons in each layer in the model are analogous to human neurons. When the human brain gets raw information, the neurons of the human brain are activated to extract features and finally "calculates" the truth of the information. Similarly, when the MLP model obtains raw data as its input, its neurons simulate the processing mechanism of the human brain through a series of linear and non-linear functions, and extract the characteristics of the data and classify them.

MLP can be seen as the prototype of the Artificial Neural Network (ANN), which connects the human brain to data analytics. Compared with traditional classification algorithms, such as decision tree and random forest, MLP focuses more on the analysis of the value, rather than simply searching for the similarities and differences between the data. MLP is a function fitting algorithm, and it treats each data as a known point in the space, and uses numerical analysis to fit a complicated function. This function calculates the corresponding category based on the input features. We use MLP to classify because of the belief that if there is a relationship between features, the relationship is generally presented as a relatively smooth function.

3 Differences

The main difference is the changing of the inference goal of this project. Compared to the binary classifier in the last project, we proposed to do a multivariate classifier this time, aiming not only to

detect abnormal requests, but also to analyze what type the attack is. This change brings unforeseen challenges to the project, since some algorithms are not designed to do such complicated task and we have to find out a way to predict on 4 columns at the same time.

Also, part of the difference between this project and that in Assessment 1 can be seen in the data processing part. We have focused more on the data processing part rather than using all of the features to train a model. We have looked into the behavior of each column and have found that there are columns with little information for our analysis. Also, the non-numeric columns are encoded using one-hot encoding, instead of simply removing them from the data set, and as we can see in the importance analysis in random forest model, the "service" feature contributes a lot to the last result.

What is more, we have used more complicated algorithms compared to those in the last assessment. For me, I have chosen Multi-Layer Perceptron this time, and have learned the idea of connecting human brains to data analytics. The design of the algorithm is attractive, for its creative idea and analytical background.