Machine Learning Engineer Nanodegree

Capstone Proposal

Anand Kumar Cheerla

June 24, 2018

Domain Background

The project deals with the computer science domain where vulnerability of online data is increased, The goal is to predict the websites either legitimate or phishy by considering the features that every website have like URL, IP address etc.

It is very important to solve this issue because now-a-days everything is getting online and usage of computers and other electronic devices is rapidly increasing . The usage of internet has increased a lot. At the same time the number of cyber crimes has increased immensely because the vulnerability of websites and lack of security for the online data leads to cyber crimes.

The online money transactions are much more vulnerable with phishy websites because most people don't even find the difference between legitimate websites and the phishy websites. They mimic the original websites very well. It is not easy to detect as there are huge number of websites, so we make use of machine learning algorithms to predict the good and bad websites.

Paper on Applying Machine Learning for Detecting phishy websites through classification:

https://www.ijcaonline.org/archives/volume147/number5/aldiabat-2016-ijca-911061.pdf

Problem Statement

By making use of the dataset features (which contains website information as individual attributes) to train the model that can generalize the data using machine learning algorithm, so that we can predict the vulnerability of any other website whether it is legitimate or phishy by using the features of the website (URL Length ,IP Address ,Sub-domain etc).We will use classification algorithm to do the task for us.

Datasets and Inputs

The data set is taken from the UCI Machine learning repository(https://archive.ics.uci.edu/ml/datasets/Website+Phishing) in which data is collected from the different online sources .Phishy websites are collected from the phishtank data archive(www.phishtank.com) and the legitimate websites are collected from Yahoo.There are total of 1353 instances of website information in this dataset .In that there are total 548 legitimate websites and 702 phishing websites and 103 suspicious URLs. There are total of 11 features which holds categorical values (Legitimate , suspicious , phishy) and due to convenience these categorical values were changed to numerical data(1, 0 , -1) as there are only three categories. Based on these features ,we can predict whether website is safe or not by training the data with those features and predict the output label.

Solution Statement

The solution to this problem is to make use of machine learning algorithms which yields best accuracy and best result. As it is a classification problem which classifies the three categories (Legitimate, suspicious, phishy), we can use supervised learning algorithms. Train the model using supervised learning algorithm based on the feature set by separating the training and testing data. Based on the Evaluation metrics we have, test the model and optimise the model that best generalizes the data. I want to use Logistic Regression as a benchmark algorithm followed by optimised algorithm which yields best results.

Benchmark Model

The benchmark model that I want to use here is Logistic Regression as it is classification problem. Based on the performance of the model (like time taken to train and the accuracy score)derived from the Evaluation metrics,I will compare with the optimised model.The one which yields best results and better accuracy is chosen as the best model.

Evaluation Metrics

The Evaluation metric I want to use here is MCC (Matthews Correlation Coefficient) based on the predicted output label and the actual output label data. Here Accuracy score won’t workout because the classes are imbalanced. MCC is a balanced measure even if the classes are of different sizes. F-beta score is also a good metric when accuracy score fails to evaluate the model. F-beta score considers both precision and recall .F-beta score is a way of combining these two.

https://en.wikipedia.org/wiki/Matthews_correlation_coefficient

Project Design

Exploring the data : Load the dataset which is used and then explore the data based on the given feature set and make necessary approximations yourself like number of records and number of features and the important features which are responsible for classification.

DataPreprocessing:

For this data there is no need to scale or normalize the data because it contains numerical feature values(-1,0,1) which represents categorical values. This step is done only when the data contains the missing values, they must be removed. If there are any non-numerical data they must be encoded using one hot encoding because most of the machine learning algorithms expect the data to be numeric.

Shuffle and Split data: We need to split the data into training and testing sets. The split is made like 80% of the data is used for training and 20% of the data is used for testing. The rule is to never use testing data to train the model inorder to detect overfitting and to know whether the model generalizes the data well. we can do this splitting by using the predefined functions also.

Evaluating Model Performance: By using the benchmark model ,We need to use Evaluation metrics to determine the performance of the model. Evaluation metrics like MCC is used. This is useful because we can make optimizations based on the performance.

Choose Best Model: Choosing the best model is the important step in the machine learning algorithm, because different algorithms fits different types of data at different situations. Choosing the appropriate model will give the best results. The models I want to try here are support vector machines (time taking but yields good result) and AdaBoost classifier(which boosts the performance).Based on the score ,I can choose the best model.

Model Tuning: Model tuning is necessary to improve the performance of the model. Tuning the parameters is a tedious task, so we can use grid search technique to tune the parameters so that the model performs very well.