Project: Machine Learning

Overview

The goal of this project is to apply machine learning on one or more of a variety of different cybersecurity problems. Solving these problems may involve large amounts of unstructured and/or structured data. The team working on this project will need to construct training, validation, and test sets and perform standard cross validation using one or more machine learning algorithms on that data.

The user should be able to apply transformations on the data to create new features and remove other features before applying machine learning. Other more advanced features to be implemented would involve discretizing data or constructing graphs from the data. A tool should be developed that will be modular and present simple controls to the user to explore various options.

Project Ideas

1) We are currently not doing machine learning using ASCII strings that can be extracted from malware. We can provide a dataset of strings that can be used as raw data that can be processed for problems that can be solved using machine learning. For example, we one perform a classification problem which involves providing strings from different malware families to a machine learning classification algorithms, and one can build a classification model that predict the family of malware of "unseen" malware. Certain transformations can be made to the strings before they are fed as input to the classification algorithms.

2) Currently we are leveraging deep learning algorithms that are exactly know for their human readability. That is, it is hard to gain intuition into what a deep neural network (DNN) has learned. Due to their complexity, i.e., large number of layers and weights, it is also difficult to understand why DNNs make specific predictions. We suggest looking into different methods of interpreting the DNN models induced and the predictions made. One method may involve extracting a decision tree from a trained network by finding the best decision tree that achieves the closest accuracy to the trained network. A second method may involve extracting information from the parameters of the trained DNNs. A third method might involve computing the importance of each input feature in the classification task.

The project will involve working with various data sets (discrete numbers, continuous numbers, text, and graph-based representations). There are quite a few challenges, specifically in writing data transformation scripts and training on large amounts of data. We suggest using scikit-learn (scikit-learn.org) for this project and Python for the scripts.

Objectives

1. Construct a large data set that can be fed into a machine learning algorithm.
2. Develop transformations that can be applied to the data before training on the data
3. Perform standard cross-validation and report accuracies.
4. Train on graph-based data representations that depict relationships in the data

Data

We will provide several sample cybersecurity datasets and will outline several prediction problems that can be solved by applying machine learning to the data. We will provide you with support with understanding any transformations that may be necessary. We will provide a large dataset for final performance evaluation of your machine learning algorithms.

Considerations

We do not expect it to be perfect, but we do expect the application of machine learning to some dataset. Achieving realistic accuracies will require using large datasets. A simply user interface should be created to allow various different transformations and to allow easy application of machine learning to the data.

Technology

- Machine learning and transformations should be performed using Python. Use scikit-learn for the learning algorithms. TensorFlow or Theano can be used for the DNNs.

Related Links

1. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.16.6788&rep=rep1&type=pdf

2. Chapter 5 in http://www2.docm.mmu.ac.uk/STAFF/D.Dancey/researchfiles/dancey-phd.pdf

3. http://www.kdnuggets.com/2015/04/model-interpretability-neural-networks-deep-learning.html