Scott Griffy

Robert Handy

Spring 2019

CS545 - Machine Learning

Teacher: Anthony Rhodes

Assignment: Final Project Write-Up

Presented: June 10th, 2019

The EMBER dataset is a set of binaries. The white paper can be found at: `https://arxiv.org/abs/1804.04637` and the code used to manipulate the dataset can be found at `https://github.com/endgameinc/ember`. The dataset is designed to assist with training malware classifiers using machine learning to classify malicious binaries. The dataset consists of a training set of 300K malware binaries, 300K benign binaries (binaries that do not have an adverse affect on computers), 300K unlabeled binaries (these could be either malicious or benign, it is not known which). The dataset also includes a test set consisting of 100K malicious binaries and 100K benign binaries.

Finding data to use for machine learning models that classify malware is a difficult task. One of the problems in acquiring and publishing a dataset is licensing issues surrounding the binaries in the dataset. Many binaries are not permitted for distribution because of commercial licenses. The authors of the EMBER dataset avoided this by releasing features of binaries in place of the actual binaries.

The authors of EMBER removed the actual binaries from this dataset. This protects the researchers performing the machine learning algorithms. If malicious binaries were included in the dataset, the researchers could potentially infect themselves while training a model.

In our project, we used this dataset to train different machine learning models. We used a Bayesian learning algorithm on the labeled data of the EMBER dataset, and a K-means

algorithm on the unlabeled data.

1 Bayesian

Your stuff here

2 K-Means

The unlabeled data in the dataset is meant to be used for semi-unsupervised learning. This was out of scope for this project, but hopefully the unsupervised analysis we performed on the data is interesting enough to be a significant contribution to understanding how the EMBER dataset can be used for machine learning models.

3 Conclusion

Write this later.