CS185C:
Final Project
Malware Classification

Jordan Conragan, Brett Dispoto

May 7, 2020

Contents

1 Preprocessing

Most of the machine learning techniques used in this report use variations of the same preprocessing steps. Here are the preprocessing steps taken for the methods described in this report. The preprocessing was done in python3, and the relevant files can be found in the `preprocessing` directory of our submission.

1. Download the dataset.
2. Split the dataset into directories based upon their family label. (Already completed by the dataset provider.)
3. For each malware family, the following steps were then taken:
   (a) Read through all of the files, count the occurance of each uniqe opcode across all files.
   (b) Take the n (turning parameter) most common Opcodes, and convert them to an ASCII symbol for observation sybmols for our HMMs. The Opcodes which are not within the n most common will be converted to an "other" symbol. This will reduce noise in our model.
   (c) Once each opcode is assigned a symbol, we again read through the files and convert the opcodes to symbols.
      i. If bagging is being used, make copies of **each** converted malware file, which will later be split up accordingly during training.
      ii. Otherwise, if boosting or stacking is being used, we can simply dump the converted opcodes (symbols) for the entire family into one huge file. This file will be our observation sequence.