Dataset

For the dataset, each training example has 2 files, the byte code file and the asm file which contacts the information extracted from the IDA disassembler. Currently the dataset consists of information extracted out from the asm file. 2 sets of features were extracted out, first is the frequency of the opcode. Second is the 2gram. There were a total of 597 different opcodes found in the examples. Therefore there is a possibility of a total of 5972 opcode combination which is a huge number. After processing and finding the frequency of all the 2 gram, trimming was done such that any feature that was 0 is removed. This lowered the number of features to around 30k. Another set of trimming was done, as there will be certain 2grams that occurs perhaps only once, those can also be removed as it is unlikely that it will provide any relevant data. The occurrence threshold for removal was set to 40. The idea is that of the 9 classes, there was a particular class that had a very small number of example with only 42 training data provided. Therefore for a particular 2gram to be relevant, it should have at least a total of 40 occurrences across the dataset. Further trimming was done till there was only 10k feature remaining.

|  |  |
| --- | --- |
| Number of Occurrences | Number of Features |
| Above 10 | 15537 |
| Above 20 | 12239 |
| Above 40 | 9478 |

Another issue was obfustication where IDA was unable to decipher the opcode and thus the features were unable to be extracted. Each training example was therefore examined, if they had the .text header in the file then it would used, else it would not be able to provide data for the features and so will be skipped.

Approach

For the training and testing set, it is split 75% and 25% with the ratio of classes kept the same. This is very important as the number of examples provided for a particular class is very small with only 42 examples.

Currently for feature selection, for random forest I have decided to use feature wrapper, however due to the size of the feature set, this could take a while to generate. The eta for the feature wrapper to complete is roughly 10 days. If it is unsuccessful, I will change tactic or not use feature selection at all. As for logistic regression and SVM, I have decided to use Embedded method and decided to use L2 loss for feature selection as that would be the most efficient.

Trial Run Data without Regularization/Feature Selection

|  |  |
| --- | --- |
| Decision Tree | 0.95 |
| SVM | 0.91 |
| Logistic Regression | 0.93 |

The issue I am currently facing is that for SVM and Logistic Regression, there is a convergence issue where they are unable to converge.