Team Marianne

Malware Classification

- Ankit Vaghela
- Vinay Kumar Bingi
- Jin Wang
- Niraj Kadam

Approach

- Extract Byte files and ASM files
- Convert to RDD and optionally to Dataframe
- Tokenize
- Convert to bigrams
- Extract Features (eg. TFIDF or count)
- Train and Predict

RDD Based Implementation

Changed loadLibSVMFile and _parse_libsvm_line

- Convert RDD to svmlib format

  {label index:feature ….}

- Converts to LabeledPoint RDD
- Loads RDD rather than extracts it from path
- Label was changed to (label-1)

Dataframe based Implementation

Dataframes

- Convert RDD to Dataframe of schema (id, feature, label)
- Transform dataframes
- Tokenize
- Convert to bigrams
- Extract Features (HashingTF+TFIDF)
- Train and Predict

Struggles and Learnings

- Struggled majorly with Google Cluster -- Get more knowledge
- Got 80% accuracy on small dataset but RDD was shuffled when ran on large dataset -- save and map IDs

Questions?