Assignment: Improvements to malware detection and classification

Machine learning is not all about autonomous vehicles and terminator robots. Techniques such as principle component analysis (PCA) can be combined with other data exploration techniques to help us gain a deeper understanding of the world around us. Many machine learning (ML) techniques aspire to reduce the complexity of data to simplify comparison and classification.

Computational techniques for analysing characteristics of ‘things’ can help to identify patterns and attributes which can be used to identify thing such as which species of plant a cell belongs to, what are the key drivers for business profitability, and what traits are common in certain diseases.

Background

TOBORRM

TOBORRM is a new computer security start-up. They have traditionally worked on hacking and penetration testing but are branching out into machine learning and active network defence systems. TOBORRM has received a grant to research and develop detection technologies for malware.

TOBORRM Data Collection

Early work by TOBORRM saw their development team automate the collection of data from download sites. The team developed a toolkit that could scour the internet for files and download them. TOBORRM used their automated tools to collect the ***MLDATASET-200000-1612938401*** data. This dataset provides 200,000 samples of clean and malicious files which have been classified as ‘Clean’ or ‘Malware’, respectively.

The dataset gathered some basic statistics about file types, download locations and sizes. The programming team also created “CodeCheck” as an internal tool to try to identify some basic file properties (such as if the file is executable, or whether the file contains ‘recognisable text strings’). It is not known whether “CodeCheck” is reliable.

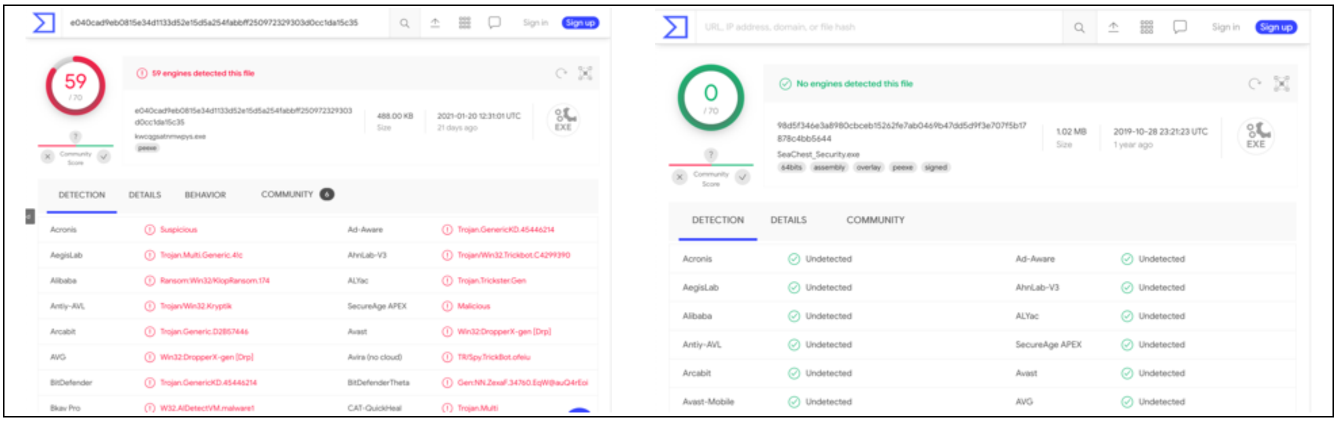
Unfortunately, the TOBORRM team does not understand the intricacies of machine learning models, and have developed the dataset without any consideration for scaling, categorisation of variables, encoding of data etc. The *MLDATASET-200000-1612938401* will require significant cleaning and preparation for it to be useful for data visualisation and machine learning.

TOBORRM Dataset Malware Classification

In order to classify malware, TOBORRM used only ‘old’ files that were likely to have been identified by other malware and virus scanners.

1. TOBORRM’s data collector would send the file to virustotal.com

2. files were tagged as “Malicious” if a majority of virustotal.com virus scanners recognised the file as containing malware (see Figure 1)
3. Files were tagged as “Clean” if ALL virustotal.com scanners identified the file as “Clean”. (see Figure 1)



59 engines detected this file

e040cad9eb0815e34d1133d52e15d5a254fabbff250972329303d0cc1da15c35
kwcqgsatnmwpys.exe
peexe

488.00 KB
Size

2021-01-20 12:31:01 UTC
21 days ago

EXE

Acronis	Suspicious	Ad-Aware	Trojan.GenericKD.45446214
AegisLab	Trojan.Multi.Generic.4!c	AhnLab-V3	Trojan/Win32.Trickbot.C4299390
Alibaba	Ransom:Win32/KlopRansom.174	ALYac	Trojan.Trickster.Gen
Antiy-AVL	Trojan/Win32.Kryptik	SecureAge APEX	Malicious
Arcabit	Trojan.Generic.D2B57446	Avast	Win32:DropperX-gen [Drp]
AVG	Win32:DropperX-gen [Drp]	Avira (no cloud)	TR/Spy.TrickBot.ofeiu
BitDefender	Trojan.GenericKD.45446214	BitDefenderTheta	Gen:NN.ZexaF.34760.EqW@auQ4rEoi
Bkav Pro	W32.AIDetectVM.malware1	CAT-QuickHeal	Trojan.Multi

No engines detected this file

98d5f346e3a8980cbceb15262fe7ab0469b47dd5d9f3e707f5b17878c4bb5644
SeaChest_Security.exe
64bits assembly overlay peexe signed

1.02 MB
Size

2019-10-28 23:21:23 UTC
1 year ago

EXE

Acronis	Undetected	Ad-Aware	Undetected
AegisLab	Undetected	AhnLab-V3	Undetected
Alibaba	Undetected	ALYac	Undetected
Antiy-AVL	Undetected	SecureAge APEX	Undetected
Arcabit	Undetected	Avast	Undetected
Avast-Mobile	Undetected	AVG	Undetected

Figure 1 - VirusTotal.com comparison of confirmed infected vs confirmed clean

As such, the ***“Actually Malicious”*** field can be considered to be a generally accurate classification for each downloaded sample.

Initially the security and software development teams believed they would be able to gain insight from various statistical analyses of the dataset. Their initial attempts to classify data lacked sensitivity and had many false positives, the results of TOBORRM’s analysis have been included in the **“Initial Statistical Analysis”** column of the data set and is provided for your information and comparison only.

SCENARIO

You have been brought on as part of a data analysis team to improve on their malware detection capabilities.

The basic analysis was conducted by TOBORRM staff based on their ‘gut feel’ and some basic statistical understanding. You will be trying to improve their initial statistical analysis by using various machine learning models for analysis and classification.

The raw data for your machine learning analysis is contained in the **MLDATASET-200000-1612938401.csv** file.

The variables in the dataset are as summarised in the table below.

Feature	Description	Data Type
Sample ID	ID number of the collected sample	Numeric
Download Source	A description of where the sample came from	Categorical
TLD	Top Level Domain of the site where the sample came from	Categorical
Download Speed	Speed recorded when obtaining the sample	Categorical
Ping Time To Server	Ping time to the server recorded when accessing the sample	Numeric
File Size (Bytes)	The size of the sample file	Numeric
How Many Times File Seen	How many other times this sample has been seen at other sites (and not downloaded)	Numeric
Executable Code Maybe Present in Headers	'CodeCheck' Program has flagged the file as possibly containing executable code in file headers	Binary
No Executable Code Found In Headers	'CodeCheck' Program has flagged the file as not containing executable code in the file headers	Binary
Calls to Low-Level System Libraries	When the file was opened or run, how many times were low-level Windows System libraries accessed	Numeric
Evidence of Code Obfuscation	'CodeCheck' Program indicates that the contents of the file may be Obfuscated	Binary
Threads Started	How many threads were started when this file was accessed or launched	Numeric
Mean Word Length of Extracted Strings	Mean length of text strings extracted from file using unix 'strings' program	Numeric
Similarity Score	An unknown scoring system used by 'CodeCheck' seems to be the score of how similar the file is to other files recognised by 'CodeCheck'	Numeric
Characters in URL	How long the URL is (after the .com / .net part). E.g. /index.html = 10 characters	Numeric
Actually Malicious	The correct classification for the file	Binary
Initial Statistical Analysis	Previous system performance of "FileSentry3000™ v1.0"	Binary

Your initial goals will be to

- Clean and prepare the data for data exploration and basic data analysis, and later (for Assignment 2) for ML modelling.
- Perform Principal Component Analysis (PCA) on the data.
- Identify features that may be useful for ML algorithms
- Create a brief report to the rest of the research team that will describe whether a subset of features could be used to effectively identify malicious files.

TASK

First, copy the code below to a R script. Enter your student ID into the command **set.seed(.)** and run the whole code. The code will create a sub-sample that is unique to you.

```r
#You may need to change/include the path of your working directory

#Import the dataset into R Studio.
dat <- read.csv("MLDATASET-200000-1612938401.csv",
                na.strings="", stringsAsFactors=TRUE)

set.seed(Enter your student ID here)

#Randomly select 500 rows
selected.rows <- sample(1:nrow(dat),size=500,replace=FALSE)

#Your sub-sample of 500 observations and excluding the 1st and last column
mydata <- dat[selected.rows,2:16]

dim(mydata)  #check the dimension of your sub-sample
```

You are to clean and perform basic data analysis on the relevant features in **mydata**, and as well as principle component analysis (PCA). This is to be done using "R". You will report on your findings.

Part 1 – Exploratory Data Analysis and Data Cleaning

(i) For each of your **categorical** or **binary** variables, determine the number (%) of instances for each of their categories and summarise them in a table as follows.

Categorical Feature	Category	N (%)
Feature 1	Category 1	10 (10%)
	Category 2	30 (30%)
	Category 3	60 (60%)
Feature 2 (Binary)	YES	75 (75%)
	NO	25 (25%)
…	…	…
Feature *k*	Category 1	25 (25%)
	Category 2	25 (25%)
	Category 3	15 (15%)
	Category 4	35 (35%)

(ii) Summarise each of your continuous/numeric variables in a table as follows.

Continuous Feature	N (%) missing	Min	Max	Mean	Median	Skewness
Feature 1						
Feature2						
….	….	….	….	….	….	….
Feature k						

(iii) Examine the results in sub-parts (i) and (ii). Are there any invalid categories/values for the categorical variables? If so, how will you deal with them and why? Is there any evidence of outliers for any of the continuous/numeric variables? If so, how many and what percentage are there and how will you deal with them? Justify your decision in the treatment of outliers (if any).

Part 2 – Perform PCA and Visualise Data

(i) Clean your data as you have suggested in Part 1 (iii) to make it usable in "R".

(ii) Export your "cleaned" data as follows. This file will need to be submitted along with you report.

```
#Write to a csv file.
write.csv(mydata,"mydata.csv")
```

** Do not read the data back in and use them **

(iii) Extract the data for the numeric features in **mydata**, along with **Actually.Malicious**, and store them as a data frame/tibble. Then, perform PCA using *prcomp(.)* in R, but only on the numeric features.

- Outline why you believe the data should or should not be scaled, i.e. standardised, when performing PCA.
- Outline the individual and cumulative proportions of variance explained by each of the first 4 components.
- Outline the coefficients (or loadings) for PC1 to PC4, and describe the loadings for the PC1 and PC2 only.
- Outline how many principal components are adequate to explain at least 50% of the variability in your data.

(iv) Create a scree plot and interpret.

(v) Create a biplot with PC1 and PC2 to help visualise the results of your PCA in the first two dimensions. Colour code the points with the variable **Actually.Malicious**. Write a paragraph to explain what your biplot is showing. That is, comment on the PCA plot, the loading plot individually, and then both plots combined (see Slides 28-29 of Module 3 notes) and outline and justify which (if any) of the features can help to distinguish Malicious and Non- Malicious files.

(vi) Based on the results from parts (iii) to (v), describe

- which dimension (choose one) can assist with the classification of malwares (Hint: project all the points in the PCA plot to PC1, i.e. horizontal axis and see whether there is good separation between the points for malicious and non- malicious files. Then project to PC2, i.e. vertical axis and see if there is separation between the malware and non-malware, and whether it is better than the projection to PC1).
- the key features in this dimension that can drive this process (Hint: based on your decision above, examine the loadings from part (iii) of your chosen PC and choose those whose absolute loading (i.e. disregard the sign) is greater than 0.3).
-

What to Submit

1. A single report (**not exceeding 5 pages, does not include cover page, contents page and reference page, if there is any**) containing:
    a. summary tables of all the variables in the dataset;
    b. a list of data issues (if any) and how you have dealt with them in the data cleaning process;
    c. your implementation of PCA and interpretation of the results, i.e. variances explained, scree plot, and the contribution of each feature for PC1 and PC2;
    d. biplot and its interpretation;
    e. your explanation of selection and contribution of the features with respect to possible identification of malicious files.

    If you use any references in your analysis or discussion outside of the notes provided in the unit, you must cite your sources.

2. The dataset containing your sub-sample of 500 observations, i.e., **mydata**.
3. A copy of your R code.

The report must be submitted through **TURNITIN** and checked for originality. The R code and data file are to be submitted separately via a Blackboard submission link.

Note that no marks will be given if the results you have provided cannot be confirmed by your code.

Marking Criteria

Criterion	**Contribution to assignment mark**
Correct implementation of descriptive analysis, data cleaning and PCA in R	20%
Correct explanation and justification in the treatment of missing and/or invalid observations in the data cleaning process	10%
Accurate specification and interpretation of the contribution of principal components and its loading coefficients.	15%
Accurate scree plot, with appropriate interpretation.	5%
Accurate biplot, with appropriate interpretation presented	25%
Appropriate selection of dimension for classification and features that contribute to the identification malicious files with justification	10%
Communications skills – Tables and figures are well presented. Report, analysis and overall narrative is well-articulated and communicated using language appropriate for a non-mathematical audience	15%
Total	100%

Academic Misconduct

Edith Cowan University regards academic misconduct of any form as unacceptable. Academic misconduct, which includes but is not limited to, plagiarism; unauthorised collaboration; cheating in examinations; theft of other student's work; collusion; inadequate and incorrect referencing; will be dealt with in accordance with the ECU Rule 40 Academic Misconduct (including Plagiarism) Policy. Ensure that you are familiar with the Academic Misconduct Rules.

Assignment Extensions

Applications for extensions must be completed using the ECU Application for Extension form, which can be accessed online.

Before applying for an extension, please check out the ECU Guidelines for Extensions which details circumstances that can and cannot be used to gain an extension. For example, normal work commitments, family commitments and extra-curricular activities are not accepted as grounds for granting you an extension of time because you are expected to plan ahead for your assessment due dates.
Please submit applications for extensions via email to both your tutor and the Unit Coordinator.

Where the assignment is submitted no more than 7 days late, the penalty shall, for each day that it is late, be 5% of the maximum assessment available for the assignment. Where the assignment is more than 7 days late, a mark of zero shall be awarded.