

## Improvements to malware detection and classification

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

Computational techniques for analysing characteristics of objects can help to identify patterns and attributes which can then be used to classify these objects 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

### N00BIoT

N00BIoT is a new computer security start-up. They have developed a secure appliance which has had great success in detecting and preventing network intrusion from Internet of Things (IoT) based threats. Their award winning N00BIoT shield is a small box that connects to the network, it isolates IoT devices from other devices.

### N00BIoT Email Sentry

In early 2019, N00BIoT has received investor capital to further improve their N00BIoT shield by adding email and data protection features. The software development team is experienced with internet security concepts and very capable of developing successful applications.

N00BIoT released version 1 of their N00BIoT Email Sentry in July 2019. The goal of the product was to add email threat detection to their N00BIoT shield. The product worked by scanning all emails for characteristics familiar to cyber security experts. When the system detected a certain match of characteristics, it takes action to quarantine emails to prevent the spread of malware. Unfortunately, the product was not successful; N00BIoT Email Sentry suffered from poor threat detection rates and excessive false positives, which resulted in user inconvenience and delayed email delivery. In September 2019 (after significant bad press) N00BIoT Email Sentry was scrapped and removed from the N00BIoT shield environment. As a result, in December 2019, the

**Malware Identification – PCA and Visualisation**  
**Assignment 1 – Part B (15%)**

N00BIoT board voted to appoint a new CEO to see the company through a new phase of product development.

The new CEO, Steve Falken has a background in AI and machine learning. He has overseen projects in the military and has experience with cyber security and security incident response management.

Steve has brought on a new team and intends on restarting research and development in to the failed N00BIoT Email Sentry with the aim delivering an improved Email Sentry 2.0 product by the end of 2020. Steve has identified a lack of expertise and specialists skills in the software development team with respect to machine learning.

## SCENARIO

You have been brought on as part of a data analysis team to help 'fix' the problems of the Email Sentry product.

The existing code for Email Sentry relied on assumptions made by developers to detect malicious software. These assumptions were guided by experience and 'gut feel' and have no statistical basis.

Data from the previous Email Sentry product has been extracted (MalwareDataSet.XLS) and provided to the development team. The initial goals of the Email Sentry 2.0 team are:

- Identify whether PCA of the MalwareDataSet could be used to identify and discriminate between emails.
- Create a brief report to the rest of the research team that will describe whether PCA could be used to effectively identify malicious emails.

## 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.

```
#You may need to change/include the path of your working directory
dat <- read.csv("Malware.csv",na.strings="") #Import the dataset into R Studio.

set.seed(Enter your student ID here)

#Randomly select 350 rows
selected.rows <- sample(1:nrow(dat),size=350,replace=FALSE)
```

**Malware Identification – PCA and Visualisation**  
**Assignment 1 – Part B (15%)**

```
#Your sub-sample of 350 observations and excluding the 1st column  
mydata <- dat[selected.rows,2:11]  
  
dim(mydata) #check the dimension of your sub-sample
```

You are to perform a principle component analysis (PCA) on the data in the MalwareDataSet.XLS. All analyses are to be done using R. You will report on your findings.

**Part 1 – Data analysis using PCA and report on findings**

- a) You must clean and standardised the data to make it usable in “R”.
- To run PCA on this data set,
    - (1) First you will need to address the NAs appropriate prior to PCA. **That is, you will need to replace the missing values for `outside.network` and `Verified.as.Malware` with the appropriate value.**
    - (2) Then, you will need to convert the category variables into something usable noting that PCA will only accept numerical values. Hint: Google the terms “dummy variables”. **Beware of functions that generate the two dummy columns for each variable. You should remove one of them prior to running PCA and be consistent with your removal across all the variables. Note that retaining the redundant dummy column does not affect the final decision from your PCA analysis, it is more that your biplot will not be as readable.**
  - Briefly report on the data manipulation, **i.e. from (1) and (2)**, that was required to make the data usable in R.
  - Export your “cleaned” data as follows.  
*#Write to a csv file. This will need to be submitted*  
`write.csv(mydata, "mydata.csv")`
- b) Perform PCA on the 9 features (1<sup>st</sup> 9 columns in **mydata only, excluding `Verified.as.Malware`**) using `prcomp(.)` in R.
- Outline the individual and cumulative proportions of variance explained by each of the first 3 components.
  - Outline the coefficients (or loadings) for PC1 to PC3.
- c) Create a scree plot **and together with part b)** outline how many principal components do you believe are adequate to explain at least 50% of the variability in your data.
- d) Create a biplot with the PC1 and PC2 to help visualise the results of your PCA in the first two dimensions. Colour code the points with the variable **`Verified.as.Malware`**. Write a short paragraph to explain what your biplot is showing. **That is, comment on the PCA plot, the loading plot and then both combined (see Slides 28-29 of Module 3 notes)**

**Malware Identification – PCA and Visualisation**  
**Assignment 1 – Part B (15%)**

- e) Based on the results from parts b) to d), describe
- (1) which dimension(s) (if any) 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 malware and non-malware. 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).
  - (2) what are the key features in this dimension that can drive this process (Hint: based on your decision in part e)(1) above, examine the loadings from part b) of your chosen PC and choose those whose absolute loading (i.e. disregard the sign) is greater than 0.3).
- f) Based on the features that you have identified in part e),
- Cross-tabulate each of your categorical feature(s) (if any) against the malware variable, i.e. **Verified.as.Malware**, using the **table(.)** function (see **Workshop 1 notes for examples**) and determine the count and percentage, i.e. N (%), relative to each level of the Malware variable. An example is given below.

		Verified as Malware	
		Yes	No
Categorical Feature	Yes	60 (60%)	20 (10%)
	No	40 (40%)	180 (90%)
Total		100	200

- Determine the median (IQR) of your continuous feature(s) (if any) across each level of the Malware variable using the **aggregate(.)** function (see **Workshop 1 notes for examples**). An example is given below.

	Verified as Malware	
	Yes	No
Continuous Feature	115 (52)	85 (44)

Comment on the differences between malware and non-malware as it relates to the features that you have identified in part e).

- g) Write a short paragraph to explaining issues you encountered with the data. Were the data appropriate for PCA? Identify any issues with the data and explain if these issues are likely to affect the reliability of your findings.

## What to report

1. Submit a single report (**not exceeding 5 pages**) containing:
  - a. explanation of data preparation steps
  - b. your implementation of PCA and interpretation of the contribution from each principal component
  - c. scree plot and its interpretation
  - d. biplot and its interpretation
  - e. your explanation of selection and contribution of the factors with respect to possible malware identification (attachment includes executable, sender outside network, email URL count)
  - f. cross-table tabulation of relevant features to malware status, with appropriate interpretation of results
  - g. any additional analysis and any discussion on data issues
  
2. If you use any references in your analysis or discussion outside of the notes provided in the unit, you must cite your sources.

## Marking Criteria

Criterion	Contribution to assignment mark
Correct explanation of data cleaning and preparation techniques.	10%
Accurate PCA results and correct specification of the contribution of principal components.	10%
Accurate scree plot, with appropriate interpretation presented in the report.	10%
Accurate biplot, with appropriate interpretation presented in the report.	15%
Accurate Interpretation of PCA results and features which contribute to the identification malicious email.	15%
Accurate cross-tabulation of relevant features with malware status with appropriate interpretation.	20%
Analysis and discussion on data issues.	10%
Communications skills – report, analysis and overall narrative is well-articulated and communicated using language appropriate for a non-mathematical audience.	10%

## Submission Instructions:

Your submission must include the following:

- Your report (5 pages or less)
- A copy of your R code
- The dataset containing your sub-sample of 350 observations

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.**

## 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.