Brunel University London

Department of Computer Science

CS3609 Cybersecurity

Cybersecurity Analytics:

Lab 5

Dr Natalie Clewley | Lecturer in Human Aspects of Cyber
Centre for Electronic Warfare, Information and Cyber
Cranfield University

Professor Panos Louvieris
Panos.Louvieris@brunel.ac.uk

27th October 2017

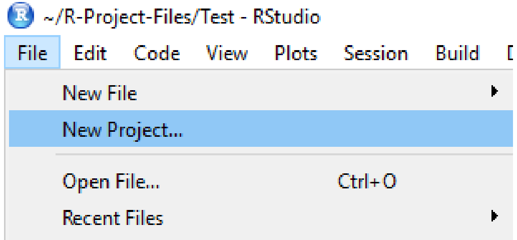
Overview

- Today's lab session will use a scenario (KDD Cup 1999) to allow you to explore two techniques used for intrusion detection mentioned in the lecture.
    - K-Means Clustering (Unsupervised)
    - Decision Tree Classification (Supervised)
- We will be using R and RStudio
- Work through the 2 activities as a class and discussion just before the end of the session.

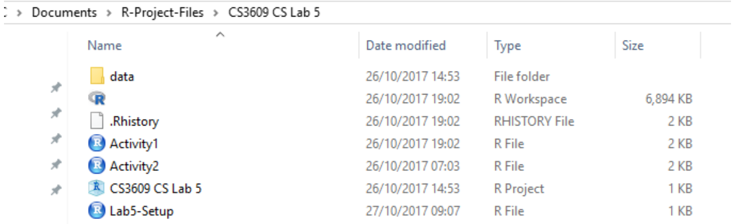
LAB SETUP

Lab – Setup Instructions

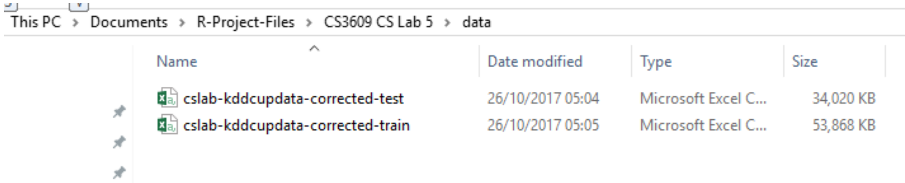
- Open **RStudio**
- Select **File → New Project**
- Select **New Directory**
- Select **Empty Project**
- Name it something like **CS3609 CS Lab 5**



- Download the lab zip files (**Lab5.zip**) from the module Blackboard page.
- In file manager, copy over files to the new folder keeping the same folder format.


Documents › R-Project-Files › CS3609 CS Lab 5

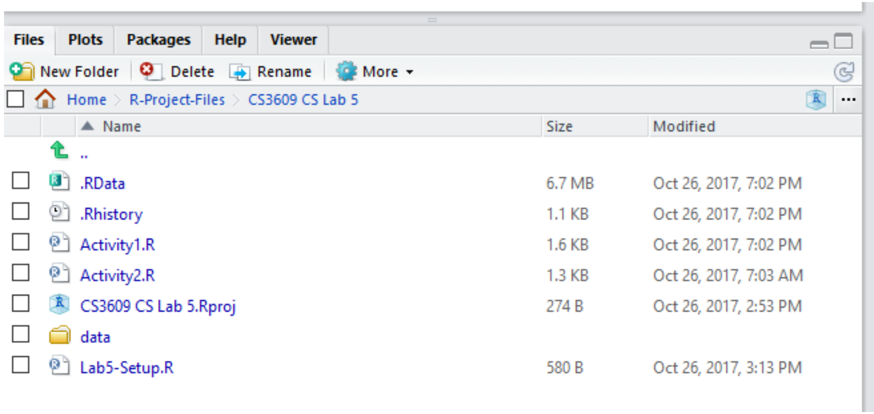
Name	Date modified	Type	Size
data	26/10/2017 14:53	File folder	
	26/10/2017 19:02	R Workspace	6,894 KB
.Rhistory	26/10/2017 19:02	RHISTORY File	2 KB
Activity1	26/10/2017 19:02	R File	2 KB
Activity2	26/10/2017 07:03	R File	2 KB
CS3609 CS Lab 5	26/10/2017 14:53	R Project	1 KB
Lab5-Setup	27/10/2017 09:07	R File	1 KB


This PC › Documents › R-Project-Files › CS3609 CS Lab 5 › data

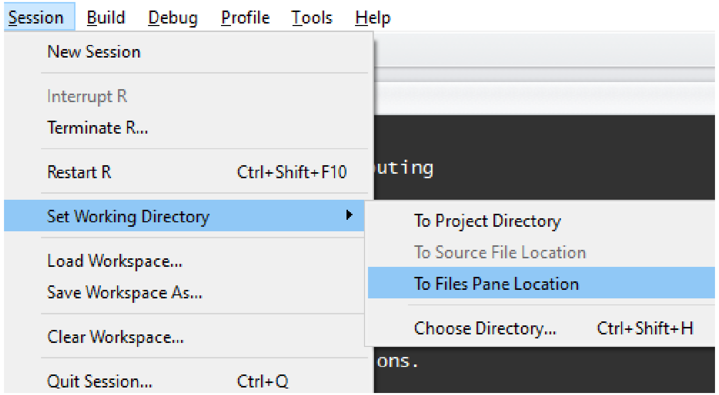
Name	Date modified	Type	Size
cslab-kddcupdata-corrected-test	26/10/2017 05:04	Microsoft Excel C...	34,020 KB
cslab-kddcupdata-corrected-train	26/10/2017 05:05	Microsoft Excel C...	53,868 KB

Lab – Setup Instructions (cont.)

- In RStudio, make sure the Files pane shows the lab files.
- Select Session → Set Working Directory → To Files Pane Location



Name	Size	Modified
..		
.RData	6.7 MB	Oct 26, 2017, 7:02 PM
.Rhistory	1.1 KB	Oct 26, 2017, 7:02 PM
Activity1.R	1.6 KB	Oct 26, 2017, 7:02 PM
Activity2.R	1.3 KB	Oct 26, 2017, 7:03 AM
CS3609 CS Lab 5.Rproj	274 B	Oct 26, 2017, 2:53 PM
data		
Lab5-Setup.R	580 B	Oct 26, 2017, 3:13 PM

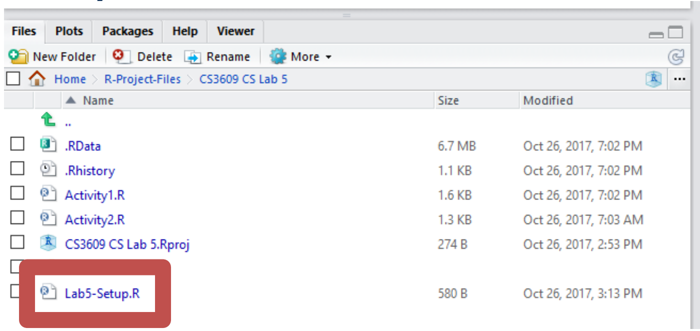


- Run any parts of a script, put the cursor on the line (or highlight a section) and press Ctrl + Enter.

Lab – Setup Instructions (cont.)

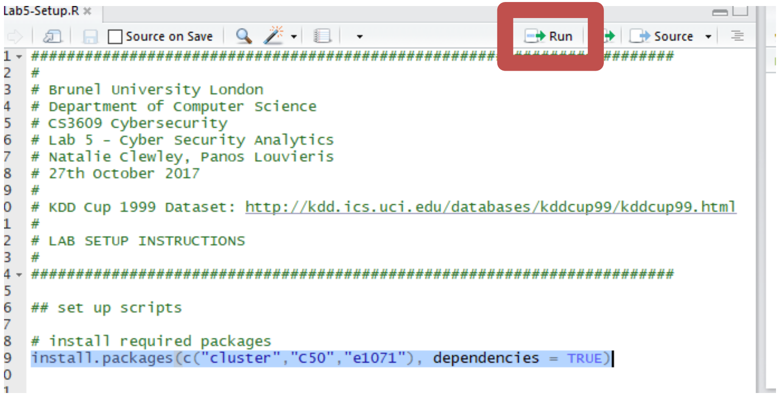
To install the required packages (libraries) for this lab:

- Select ‘Lab5-Setup.R’ from the Files Pane.



Name	Size	Modified
..		
.RData	6.7 MB	Oct 26, 2017, 7:02 PM
.Rhistory	1.1 KB	Oct 26, 2017, 7:02 PM
Activity1.R	1.6 KB	Oct 26, 2017, 7:02 PM
Activity2.R	1.3 KB	Oct 26, 2017, 7:03 AM
CS3609 CS Lab 5.Rproj	274 B	Oct 26, 2017, 2:53 PM
Lab5-Setup.R	580 B	Oct 26, 2017, 3:13 PM

- Select the code you want to run and press Ctrl + R or click Run.



```
############################################################################
#
# Brunel University London
# Department of Computer Science
# CS3609 Cybersecurity
# Lab 5 - Cyber Security Analytics
# Natalie Clewley, Panos Louvieris
# 27th October 2017
#
# KDD Cup 1999 Dataset: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html
#
# LAB SETUP INSTRUCTIONS
#
############################################################################

## set up scripts

# install required packages
install.packages(c("cluster","C50","e1071"), dependencies = TRUE)
```

LAB SESSION SCENARIO

Lab Session Scenario: Overview

- Scenario: Typical U.S. Air Force Local Area Network (LAN)
- Duration: 9 weeks (7 weeks training, 2 weeks testing)
- Size of original dataset: 4GB of compressed binary TCP dump data = approx. 5 million connection records
    - A connection is a sequence of TCP packets starting and ending at some well defined times, between which data flows to and from a source IP address to a target IP address under some well defined protocol.
    - Each connection is labelled as either *normal*, or as an *attack*, with exactly one specific *attack type*.
    - Each connection record consists of about 100 bytes.

Source: KDD Cup 1999. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

Lab Session Scenario: Attacks

Attacks fall into four main categories:

1. **DOS**: denial-of-service, e.g. syn flood;
2. **R2L**: unauthorized access from a remote machine, e.g. guessing password;
3. **U2R**: unauthorized access to local superuser (root) privileges, e.g., various buffer overflow attacks;
4. **Probing**: surveillance and other probing, e.g., port scanning.

Challenge: identify the intrusions (cyber attacks) within the dataset

Lab Session Scenario: Features

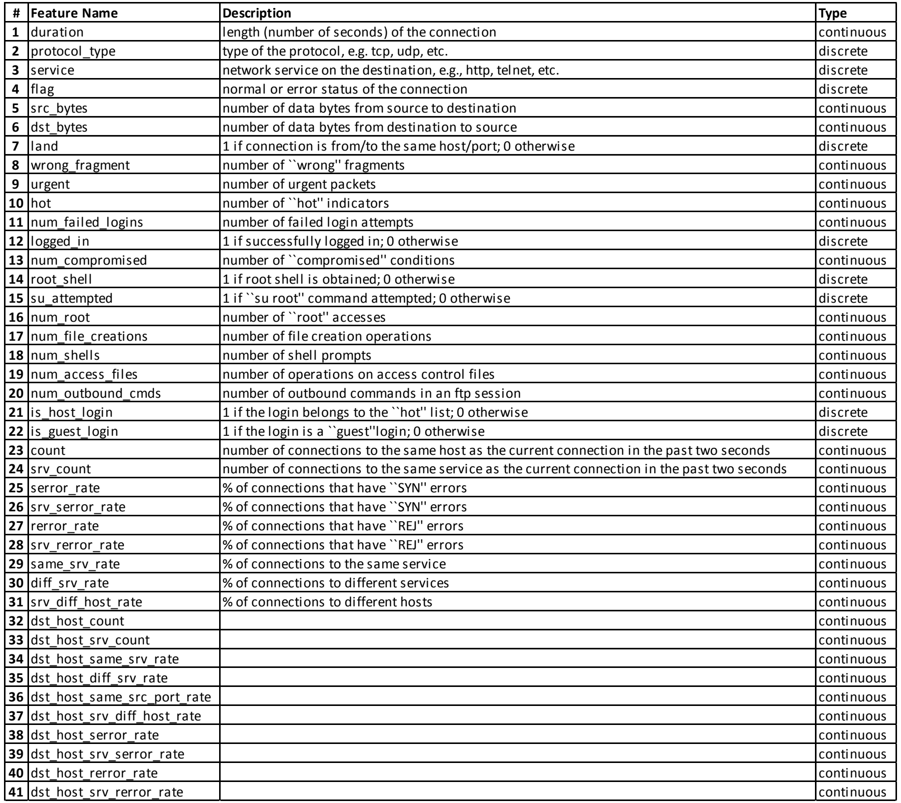
41 features in total

Contains two types of features:

- **Raw** features
  - (e.g. flag, src_bytes)
- **Derived** ‘higher-level’ features
  - (e.g. serror_rate, count)

More information:

http://kdd.ics.uci.edu/databases/kddcup99/task.html



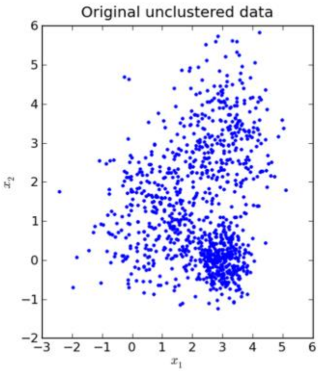
#	Feature Name	Description	Type
1	duration	length (number of seconds) of the connection	continuous
2	protocol_type	type of the protocol, e.g. tcp, udp, etc.	discrete
3	service	network service on the destination, e.g., http, telnet, etc.	discrete
4	flag	normal or error status of the connection	discrete
5	src_bytes	number of data bytes from source to destination	continuous
6	dst_bytes	number of data bytes from destination to source	continuous
7	land	1 if connection is from/to the same host/port; 0 otherwise	discrete
8	wrong_fragment	number of ``wrong'' fragments	continuous
9	urgent	number of urgent packets	continuous
10	hot	number of ``hot'' indicators	continuous
11	num_failed_logins	number of failed login attempts	continuous
12	logged_in	1 if successfully logged in; 0 otherwise	discrete
13	num_compromised	number of ``compromised'' conditions	continuous
14	root_shell	1 if root shell is obtained; 0 otherwise	discrete
15	su_attempted	1 if ``su root'' command attempted; 0 otherwise	discrete
16	num_root	number of ``root'' accesses	continuous
17	num_file_creations	number of file creation operations	continuous
18	num_shells	number of shell prompts	continuous
19	num_access_files	number of operations on access control files	continuous
20	num_outbound_cmds	number of outbound commands in an ftp session	continuous
21	is_host_login	1 if the login belongs to the ``hot'' list; 0 otherwise	discrete
22	is_guest_login	1 if the login is a ``guest''login; 0 otherwise	discrete
23	count	number of connections to the same host as the current connection in the past two seconds	continuous
24	srv_count	number of connections to the same service as the current connection in the past two seconds	continuous
25	serror_rate	% of connections that have ``SYN'' errors	continuous
26	srv_serror_rate	% of connections that have ``SYN'' errors	continuous
27	rerror_rate	% of connections that have ``REJ'' errors	continuous
28	srv_rerror_rate	% of connections that have ``REJ'' errors	continuous
29	same_srv_rate	% of connections to the same service	continuous
30	diff_srv_rate	% of connections to different services	continuous
31	srv_diff_host_rate	% of connections to different hosts	continuous
32	dst_host_count		continuous
33	dst_host_srv_count		continuous
34	dst_host_same_srv_rate		continuous
35	dst_host_diff_srv_rate		continuous
36	dst_host_same_src_port_rate		continuous
37	dst_host_srv_diff_host_rate		continuous
38	dst_host_serror_rate		continuous
39	dst_host_srv_serror_rate		continuous
40	dst_host_rerror_rate		continuous
41	dst_host_srv_rerror_rate		continuous

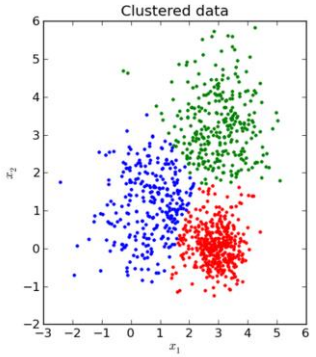
Exploratory Data Analysis with K-Means Clustering

ACTIVITY 1

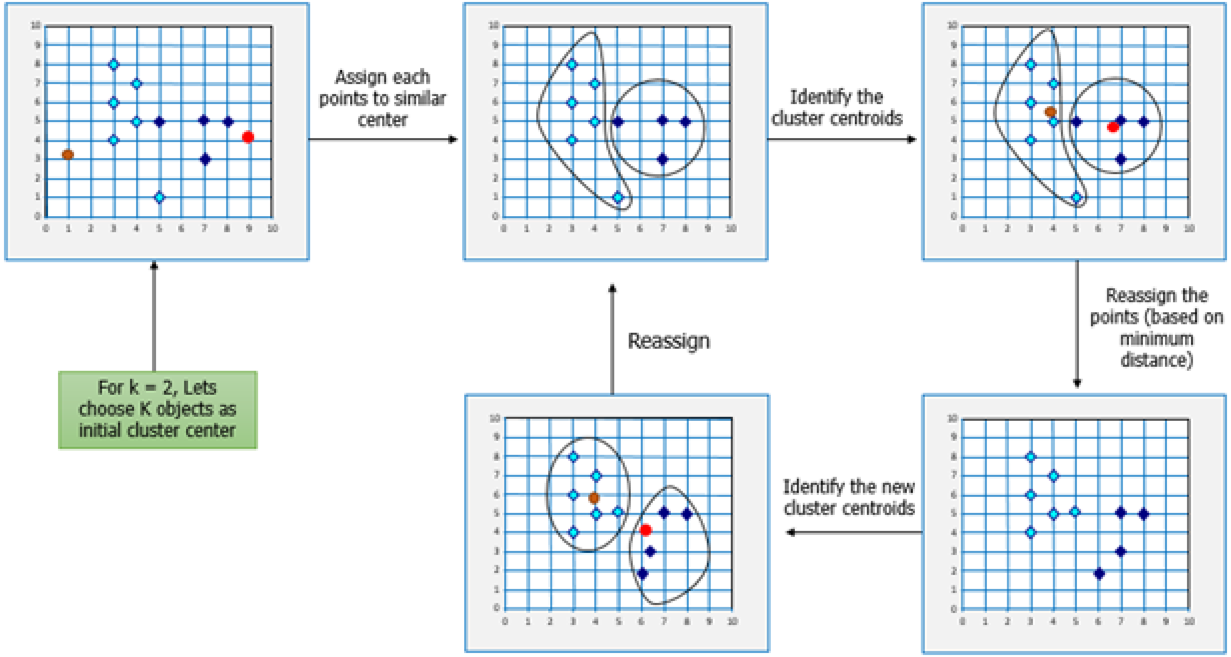
K-Means Clustering

- Clustering is a form of **unsupervised learning**, which is employed in the analysis of unlabelled data that is not categorised or grouped or when not a lot is known about the data.
- K-Means (MacQueen, 1967) is a **partition-based** clustering algorithm commonly used in intrusion detection because it provides transparent analysis of clustered data.
- **K-Means** groups the data into ***k*** groups (clusters) based on a similarity measure. Based on the features, it iteratively assigns points to each cluster so that each point is similar to those within the cluster and dissimilar to those outside the cluster.
- K-Means outputs:
  - The centroids (centre points) of each cluster;
  - The clustering assignments where each point is assigned to exactly one cluster;





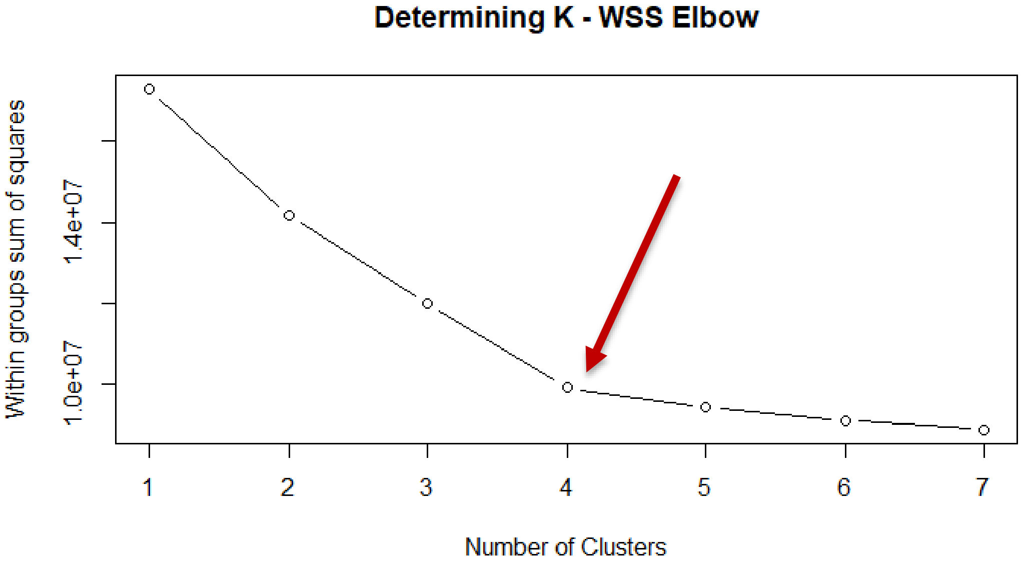
K-Means Example


For k = 2, Lets choose K objects as initial cluster center
Assign each points to similar center
Identify the cluster centroids
Reassign the points (based on minimum distance)
Identify the new cluster centroids
Reassign

Source: https://www.edureka.co/blog/k-means-clustering

Activity 1

1. Identify k through analysing the within groups sum of squares (the mean distance between the points and their centroids) elbow graph.



Activity 1

2. Perform k-means clustering with identified k value.

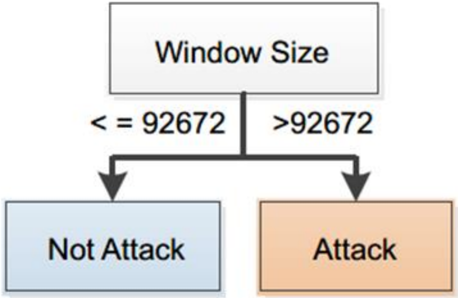
3. Can you identify any relationships between clusters and attacks?

Decision Tree Classification

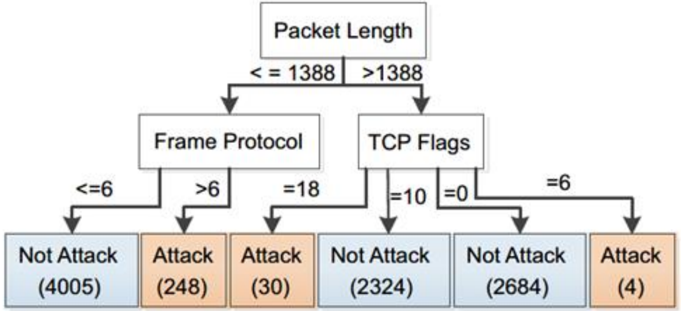
ACTIVITY 2

Decision Tree Classification

- Decision trees construct **tree-like structures** through a series of Boolean functions (i.e. "yes" or "no" questions based on the characteristics of a set of variables) until no more relevant branches can be derived.
- New data items can then be classified by starting at the root node and moving down through the branches until a **leaf node** is reached and a classification obtained.
- **Binary trees** or **multi-branch trees**.
- Quinlan's C4.5/C5.0 algorithm is commonly employed.



Binary Tree



Multi-Branch Tree

Activity 2

1. Build a C5.0 decision tree model on the training data.
2. What issues do you notice with the tree model?
3. Test the model on the test data. How does it perform?
4. Select subsets of features and rerun the classification.

Final Thoughts

- Number of features
- Visualisation
- Scaling variables
- Types of variables
- Normality
- Interpretation requires domain knowledge

References

- MacQueen, J. (1967) ‘*Some methods for classification and analysis of multivariate observations*’, In Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics and Probability, University of California, Vol.1, pp. 281–297. Available at: https://projecteuclid.org/euclid.bsmsp/1200512992.

Useful links:

- Good starting resource if you get stuck is https://www.statmethods.net