Network Security

Project 2: Attack Classification

Instructor: Shiuhpyng Shieh
TA: Su-Xin Chong, Wan-Yu Chen, Tsung-Hung Wu, Pei-Hsuan Hung, Zhong-Hao Liao
Email: TA@dsns.css.nctu.edu.tw

Due date: 23:55, May 30, 2021

1. Project Description

The goal of this project is to practice cyber attack log analysis. This is a personal project, each of you will be provided with 5 sets of data collected when different attacks are carried out. You have to implement a classifier to classify different attack scenarios. You can either observe the logs and then write a rule-based model or use machine learning method to tune a best-result model. You are encouraged to use ELK stack in project1 to perform the analysis.

2. Project Guide

1. Project Target:
   The goal of this project is to practice analyzing logs. Log can be analyzed in many different ways for different purposes. For this project, you are practicing to classify attack types with simple attack cases. In this project, you need to first observe the logs, then pre-process the logs, and then finally construct a classification model.**You can use rule-based or ML-based models in any language you prefered .**
2. Keywords:
   (a) Log Analysis:
   Log analysis is the process of making sense of computer-generated log messages, also known as log events, audit trail records, or simply logs. Log analysis provides useful metrics that paint a clear picture of what has happened across the infrastructure. You can use this data to improve or solve performance issues within an application or infrastructure. Looking at the bigger picture, companies analyze logs to proactively and reactively mitigate risks, comply with security policies, audits, and regulations, and understand online user behavior.

   (b) Machine Learning:
   Machine learning (ML) is the study of computer algorithms that improve automatically through experience. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as ”training data”, in order to make predictions or decisions without being explicitly programmed to do so. Machine learning algorithms are used in a wide variety of applications, such as email filtering and computer vision, where it is difficult or infeasible to develop conventional algorithms to perform the needed tasks.

   (c) Cyber Attack:
   In computers and computer networks an attack is any attempt to expose, alter, disable, destroy, steal or gain unauthorized access to or make unauthorized use of an asset.[1] A cyber attack is any type of offensive maneuver that targets computer information systems, infrastructures, computer networks, or personal computer devices. An attacker is a person or process that attempts to access data, functions or other restricted areas of the system without authorization, potentially with malicious intent.

3. Attack Scenarios

1. Attack Scenario Description:

- Port Scan:
  The logs are collected from a host that was attacked by port scan.
  A port scan is an attack that sends client requests to a range of server port addresses on a host, with the goal of finding an active port and exploiting a known vulnerability of that service. Scanning, as a method for discovering exploitable communication channels, has been around for ages.

- IP Scan:
  The logs are collected from a host that launched an IP scan attack.
  A IP scan is an attack that sends requests to a range of IP addresses, with the goal of finding the active IP addresses.

- RDP Brute-Force attack:
  The logs are collected from a RDP server that was attacked by brute-force attack. The attacker attempted to guess the username and password. You can observe the attacker attempting to connect to the RDP port.The RDP service is running on the default port 3389.

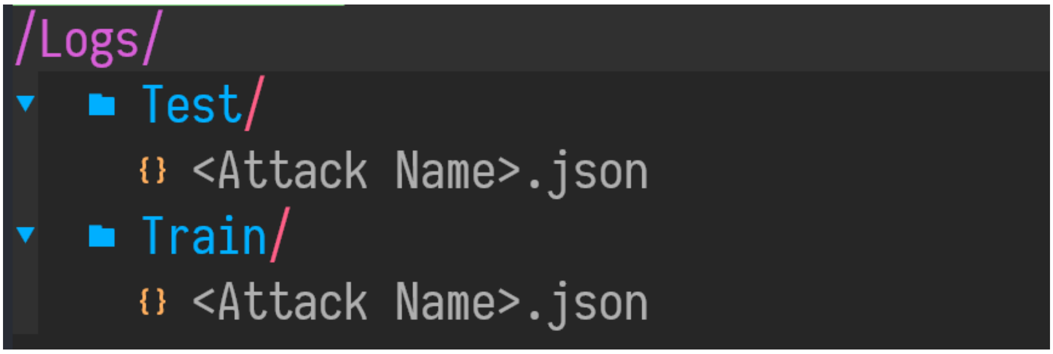
- DDoS:
  The logs are collected from a ssh server that was attacked by DDoS. In other words, the ssh service was overwhelmed by numerous requests. The ssh service is running on the default port 22.
  Distributed Denial of Service(DDoS) is typically accomplished by flooding the targeted machine or resource with superfluous requests in an attempt to overload systems and prevent some or all legitimate requests from being fulfilled.

- C&C:
  Command and Control, C&C, is a tactic to control a compromised device during a campaign. Generally, the victim device will connect to a C&C server periodically to receive new commands from the attacker or to exfiltrate data to the attacker. The logs are collected from a victim.

2. Given data:
   You are given 5 attacks' cases as training data to design your model. A case is a type of attack from the list above and there are no repetitive attacks in the cases. Each case contains a packetbeat log file dumped from elasticsearch..

4. Dataset Structure and I/O Rules

1. Log Structure:



```
/Logs/
▼ ■ Test/
    {} <Attack Name>.json
▼ ■ Train/
    {} <Attack Name>.json
```

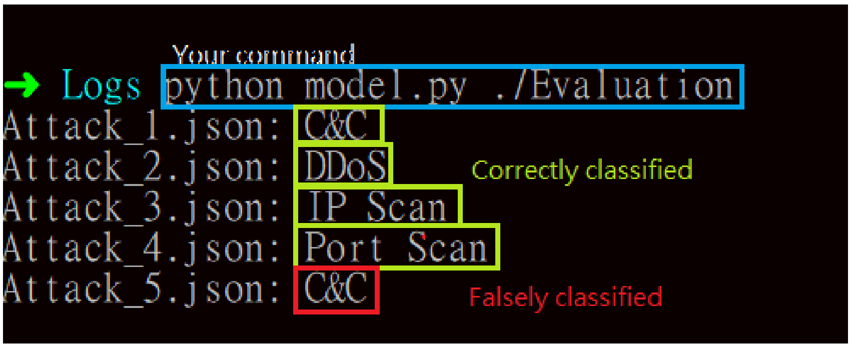
2. Code Input/Output:

   (a) Input: Your code should accept folder path as an argument, and read files under that folder. The input format should be in *python3 <project2_name.py>* *<dataset path>* e.g: python3 project2.py ./Train

   (b) Output: Testing data is stored like the above structure, you need to read in **all the files in the**

directory to do the testing part. The result for each test case should be printed in *<filename: category>* e.g: DDoS.json: DDoS

(c) Example:



```
➜ Logs python model.py ./Evaluation
Attack_1.json: C&C
Attack_2.json: DDoS
Attack_3.json: IP Scan
Attack_4.json: Port Scan
Attack_5.json: C&C
```

3. WARNING: You'll be deducted for 10 points if you don't follow the coding regulations of output.

5. Submission and Scoring

Part A: (25%)

A report in PDF format that contains:

1. What model or algorithm do you use?
2. What features/rules you used for your model?
3. Why do you select them? Please describe as much detail as possible
4. Anything interesting you find or problems you encounter in the whole process.

A model folder that contains:

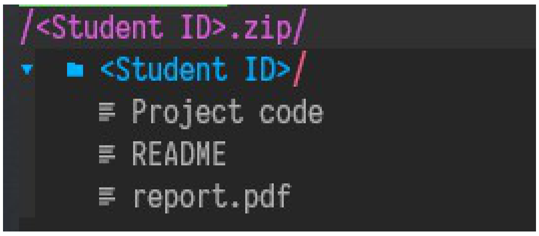
I. Source code of your model.
II. A README file in which you explain how to execute your model.

Part B: (75%)

- Demo period will be announced on E3 one week before the deadline, please pay attention to our announcement and fill in the demo period table.
- Please bring a computer with you to demo your project.
- We'll test your model with new test cases that we didn't provide
- Demo held at:
    - 5/31(Mon.)
    - 6/1(Tue.)
    - 6/4(Fri.)
- 10 points will be deducted for handing in the wrong file format.

6. Submission Rules

- Compress your report PDF and project code (model folder) into a zip file. Upload the zip file as <Student_ID>.zip " to E3 platform.



```
/<Student ID>.zip/
▾ <Student ID>/
    Project code
    README
    report.pdf
```

- Late submission penalty is 10% per day.