Machine Learning

Final Project Description

Manar Mohaisen, Department of Computer Science

Introduction:

Intrusion detection/prevention systems (IDSs/IPSs) are essential perimeter security tools. They are considered the frontline defense tools against various cyberattacks. In this project, the CIC-IDS2017 dataset is used. The original dataset contains benign traffic and 14 attack traffic. The traffic is captured over five days. The dataset is available at https://www.unb.ca/cic/datasets/ids-2017.html.

The dataset provides five large pcap files, one file per day. The pcap files include information about each network packet exchanged. As a preprocessing step, the pcap files are processed by the CICFlowMeter tool, which generates flows (think of flows as sessions with unique attributes) with 78 features. Note that an internet traffic flow is identified by a unique combination of source and destination Internet Protocol numbers (IPs), source and destination port numbers, a timestamp, and transport protocol (TCP or UDP). These features can be used as input to machine learning algorithms.

In this project, you will use the knowledge you acquired in this course to build classifiers that can efficiently distinguish denial of service (DoS) attacks from benign traffic. In the dataset, only Wednesday traffic includes DoS attacks and benign traffic. The Wednesday traffic includes:

- Benign
- DoS GoldenEye
- DoS Hulk
- DoS slowhttptest
- DoS slowloris
- Heartbleed

The last attack is not a denial-of-service attack, so the corresponding flows should be removed from the dataset. The DoS attacks should be grouped into a class labeled ‘attack’ for binary classification.

Data Sources:

You are provided with the dataset in the following formats:

- Several CSV files
- Several JSON files
- Several parquet files

The last column in the dataset is the class label.

Submission:

1. Team member names and student IDs must appear on top of every submitted document.
2. The completed architectural decision records in PDF format
3. The Python notebook for the entire project in ipynb format
    a. You should include comments and discussions at every stage, mainly when making critical decisions or evaluating model performances.
4. A snapshot of the first 20 rows of the loaded document in step 10 of the ADR document
5. A presentation file to present the content of the ADR document (You may use the ADR and Python notebook for presentation)