Assignment 10

Applied Machine Learning

In this assignment, we will develop a ML model for cybersecurity intrusion detection. Please visit the website: https://www.unb.ca/cic/datasets/ids-2017.html and look around to see the problem space and the evaluation datasets to be used for ML model development.

This dataset is collected by cyber experts during experimentation that was carried on for 5 days long. The description of the experiments also informs the experimental ground truth. (Suggested: GeneratedLabelledFlows.zip, note that it is already pre-processed by someone)

1. [10 pts] Download the labeled dataset, if you like use a dummy email address for registration. There must be 8 data files, each representing a particular cyber-attack type and it's day, and it's collected pcap (packet capture) data.

2. [10 pts] Pick one of the data files, call it **Dataset 1**, and examine its features. Make sure it has more than one class value for its label.

3. [10 pts] For the **Dataset 1**, pick a machine learning methodology and justify your choice.

4. [10 pts] Process the class feature/category as binary classes for supervised learning, assign BENIGN to value 0 and the rest to value 1. Check its balance for the **Dataset 1.**

5. [10 pts] Explore **Dataset 1** features with respect to the class. (Hint: features `Source Port` and `Destination Port` are very useful, research and find out important networking port numbers and one-hot-encode them. Unimportant port numbers or source port numbers can be assigned to a feature called '`other ports`')

6. [10 pts] Display some histograms and anything you deem fit to pick independent **Dataset 1** features. (Hint: source/destination bytes, packets, ports and the duration features)

7. [10 pts] Attempt a few classifier models and report their 10-fold CV performance.

8. [10 pts] Convert your code to be used for the remaining 7 datasets, i.e. Datasets 2-8.

9. [10 pts] Pick a classifier algorithm and report its evaluation for the remaining 7 datasets. Note that one dataset has a single class, which might need an unsupervised learning.

10. [10 pts] Briefly write up your thoughts about developing a machine learning model where you are not a subject matter expert, such as, developing a cybersecurity intrusion detection pipeline as in this assignment.