COMP9071 Fraud and Anomaly Detection
Assignment 1

February 18, 2022

1 Introduction

This assignment is worth 50% of your overall grade. It requires you to perform an analysis of classification models to detect anomalies. Standard methodologies should be applied, and the performance should be evaluated.

2 Submission

The assignment pdf must be submitted through Canvas by **10 p.m. Sunday 13th March 2022**
You must email me separately (through Canvas mail) the python file(s) containing all the code used and a Readme.txt with instructions for running the code at the same time.
You MUST name your main python file (to be run from command line) and your report with your name and student id in the format LastName_FirstName_Sid, e.g. grimes_diarmuid_123456.py, grimes_diarmuid_123456.pdf.
As per CIT regulations, submitting within 7 days of the deadline will result in a 10% penalty, between 7 and 14 days late will result in a 20% penalty, and later than 14 days after the due date will result in a 100% penalty applied.

Academic Integrity

This is an individual assignment. The work you submit must be your own. In no way, shape or form should you submit work as if it were your own when some or all of it is not.
Collusion: Given how much freedom there is in the assignment, everybody's work will be different. It will be obvious if there is collusion. All parties to collusion will be penalized.
Deliberate plagiarism: You must not plagiarise the programs, results, writings or other efforts of another student or any other third-party. Plagiarism will meet with severe penalties, which can include exclusion from the University.

Inadvertent plagiarism: In reporting your exploration of the research literature be careful to avoid inadvertent plagiarism (e.g where "paraphrases" of the source material are too close to the original).
Falsification and fabrication: The experimental results reported must come from the experiments that you have run. Do not falsify or fabricate results.
Your report will be checked for signs of collusion, plagiarism, falsification and fabrication. You may be called to discuss your submission and implementation with me and this will inform the grading, any penalties and any disciplinary actions.

Asking Questions

Please post questions in the discussion forum on Canvas so that everyone can benefit from the responses/clarifications.

3 Project Brief: Analysis of Anomaly Detection Algorithm

You work for a large Enterprise that wants to know more about anomaly-based network intrusion detection system. You have been tasked with providing an analysis of an anomaly detection method in order to find exceptional patterns in network traffic that do not conform to the expected normal behaviour. As a case study, you are given a dataset. You need to perform some pre-processing over the dataset and then identify anomalies in the given dataset using an anomaly detection classification method of your choice. You are also required to evaluate and compare the performance of your anomaly detection techniques. You must implement your anomaly detection algorithm in Python (you can use any built-in functions or libraries such as Numpy, Pandas, or Sklearn, etc).

You are provided with the dataset "Assignment1dataset22.csv" which is an adapted version of data from the NSL-KDD dataset [1]. This contains 14 attributes and a binary class target: 0 for Normal, 1 for Anomaly (i.e. an attack). Note some of the attributes have been added to the original dataset. You must perform the following actions:

1. Statistical analysis: use boxplots to identify outliers for the different features and assess the accuracy of such an approach on this dataset for identifying outliers.

   [15%]

2. Preprocessing and feature selection: Analyse the data and select 10 of the features that you will then use for your classification. You should indicate where possible why you chose / didn't choose certain features. Divide the dataset into training and testing sets.

   [15%]

3. The anomalous instances in the dataset are less than 5% of the dataset. Artificially generate a *balanced* dataset: Use under sampling and oversampling to produce a balanced dataset.

   [15%]

4. Run a classifier on both the original and the balanced datasets. You should run a controlled scientific experiment, and care should also be given to the metrics used.

   [15%]

[1] `https://www.unb.ca/cic/datasets/nsl.html`

5. Provide in-depth analysis of the behaviour of the classifier, including discussion of underfitting and overfitting. Feel free to use tables, diagrams, charts and graphs to make the presentation of your work more vivid. Interpret the results where applicable.

[30%]

For parts 3 and 4, give precise, concise explanations of what is being compared, what is being measured, and your experimental methodology.

In each case, you are being assessed based on your understanding, so for each of the four steps you should discuss what you did, why you did it, and what the outcome was. The final 10% of marks are based on demonstration of general understanding of the relevant course content.

Reproducable results

For running experiments, you must *seed* the random number generator (as shown below using the `seed` function of the `random` package), and the seed value you must use is your student id number (integer only removing the R and trailing 0's to the left, so R000012345 becomes 12345). This is what will be used to reproduce your results, the same number should be passed as argument to any *random_state* inputs in functions used. If being set in a function that is being looped over than Your results should be presented in a report, together with the code used.

```
import numpy as np
np.random.seed(123456789)
```

4 Report Format

The format for your report should be a pdf comprising the following:

1. A heading that includes your name and student id
2. One or more sections and subsections presenting your discussion of what was implemented and why, what was tested, the experimental design, and your experimental results
3. A final section that offers conclusions and
4. A list of references, i.e. sources cited in the body of your notebook.

Citations should be given using the Harvard referencing style. Note that the references section is not a list of things you've read. It is a list of things you've cited in the body of your report. References should be exact, "Wikipedia" is not an exact reference, nor is "GitHub", etc. You must give the exact url if referencing an online resource that was used.