Experiment 1: Basic Data Processing

Determine how many distinct source IP addresses, destination IP addresses, and classifications there are in the dataset coursework 1.

Experiment 2: Basic Data Analysis and Visualisation

Obviously, there is quite a lot of data here and we need to group the records in some way. One possibility is to group source IP addresses (and destination addresses) by the number of records they appear in. Write code to count the number of records containing each source IP address, and the number of records containing each destination IP address. Generate histograms to visualise results.

Experiment 3: Clustering

Using these values, cluster the source IP addresses by the number of records they appear in. Repeat for destination IP addresses. Is there an obvious number of clusters (and hence, an obvious split in the tally counts)? Explore using different clustering algorithms and different tools for determining the number of clusters in coursework 1 dataset.

Experiment 4: Finding Relationships

Using 4 clusters for source IP addresses (split them at up to 20 records, 21 − 200, 201 − 400, > 400) and 4 clusters for destination IP addresses (split them at up to 40 records, 41 − 100, 101 − 400, > 400), investigate the relation between source and destination - for example, does source-cluster 1 always contact a destination in destination cluster 3? Determine conditional probabilities and Illustrate this graphically?

Experiment 5: Decision Trees

Write code to learn a decision tree using the 2 features above (i.e. the source cluster and the destination cluster) to predict the classification field. Built in sklearn decision tree should not be used as this does not deal well with categorical features. Display the learnt decision tree. In how many cases does your learnt decision tree give an unambiguous answer (or a fairly certain answer)?

Experiment 6: Extension

Examine the dataset coursework2.csv. It contains similar data but has a few more IP addresses. Using the same clusters of IP addresses (plus sets of previously unseen source and destinations), are the patterns observed in Q4 still valid? How about the decision tree in Q5?