Coursework for Introduction to AI

Download the following dataset from blackboard;

`coursework1.csv`

This is a logfile from an intrusion detection system (IDS). It contains records in the format: "time","sourceIP", "sourcePort", "destIP", "destPort", "classification", "priority", "label", "packet info", "packet info cont'd", "xref". We will be concerned with the sourceIP, destIP and classification fields. You can load this .csv into python as a Panda data frame.

```
import matplotlib.pyplot as plt
import pandas as pd

data = pd.read_csv("coursework1.csv")
```

Pandas data frame are a bit like spreadsheets and use a straightforward indexing method. So for example,

```
print(data['sourceIP'][0])
```

prints the entry under column sourceIP for row 0, and

```
print(data['destIP'][1086])
```

prints the entry in the destIP column for row 1086. There are then other commands which give summaries of the data. For example,

```
print(data.head)
```

prints the first and last 5 rows of the data and

```
print(data.shape)
```

prints the $rows \times columns$ dimensions of the data

You might also find the tolist() command useful to convert from data frames to lists. For example,

```
allsourceip = data['sourceIP'].tolist()
alldestinationip=data['destIP'].tolist()
```

creates lists of all the source IP addresses and the destination IP addresses.

For this coursework you should write a report of no more than 6 pages which answers the following questions. You should upload this to blackboard together with a commented version of your code. These should both be uploaded by 24th April 2020.

Q1: Basic Data Processing

Determine how many distinct source IP addresses, destination IP addresses, and classifications there are in this dataset.
(10 marks)

Q2: 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 your results.
(10 marks)

Q3: 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)? You should explore using different clustering algorithms and different tools for determining the number of clusters. You might find some of the content of worksheet 1 useful in this context.
(15 marks)

Q4: 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? Can you determine conditional probabilities? Can you illustrate this graphically?
(20 marks)

Q5: Decision Trees

Write some of your own 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. You should not use the built in sklearn decision tree as this does not deal well with categorical features.

Display your learnt decision tree. In how many cases does your learnt decision tree give an unambiguous answer (or a fairly certain answer)?
(30 marks)

Q6 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?
(15 marks)