Boyang Xia
Email: oy19982@bristol.ac.uk
Course: EMATM0044_2019_TB2: Introduction to Artificial Intelligence 2019
Instructor: Prof. Jonathan Lawry, Prof. Trevor Martin

Coursework - 1
Student ID: 1902964
Term: Spring 2020
Due Date: 24^{th} April, 2020

Problem 1

Determine how many distinct source IP addresses, destination IP addresses, and classifications there are in the dataset: coursework1.csv.

Solution:

- SourceIP: 98
- DestIP: 261
- Classification: 3

Problem 2

Visualise the number of records containing each source and destination IP address.

Solution:

Since there are too many distinct cases of IP adresses, the x-axis cannot display all IP addresses in high graphical quality, please run the code file *coursework_Q2_final.py* to check details.

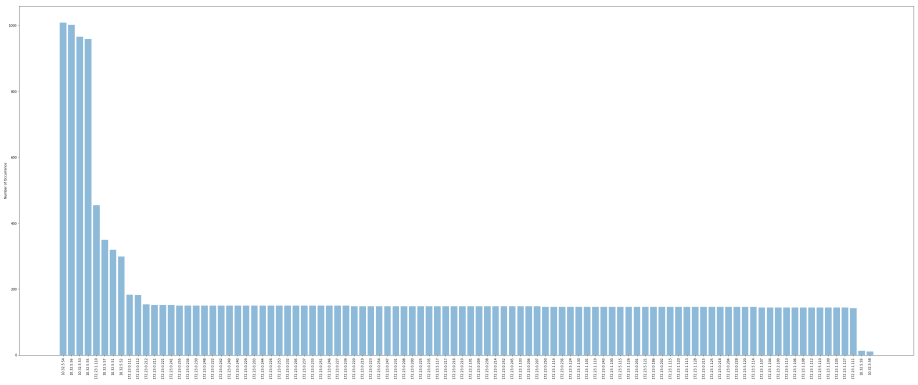


Figure 1: Histogram of SourceIP

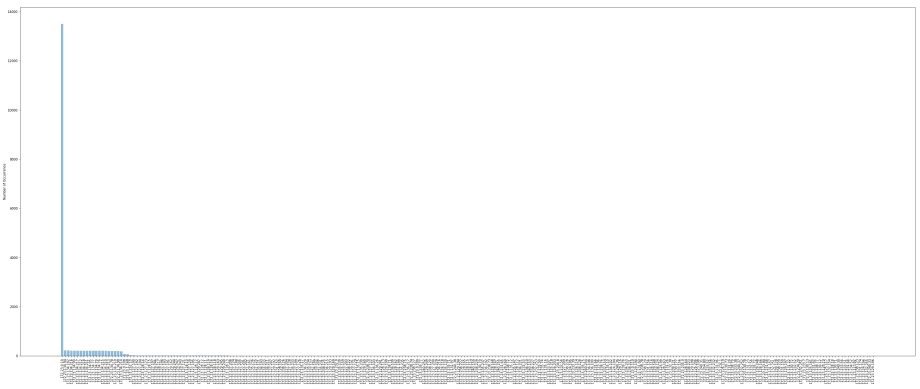


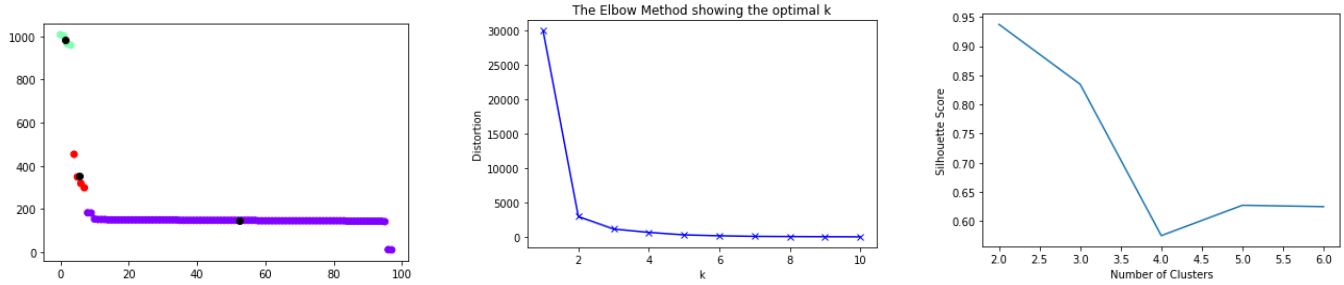
Figure 2: Histogram of DestIP

Problem 3

Use the values in Q2, cluster the source and destination IP addresses by the number of records they appear in. Try to find the obvious number of clusters, and explore using different clustering algorithms and tools to determine the number.

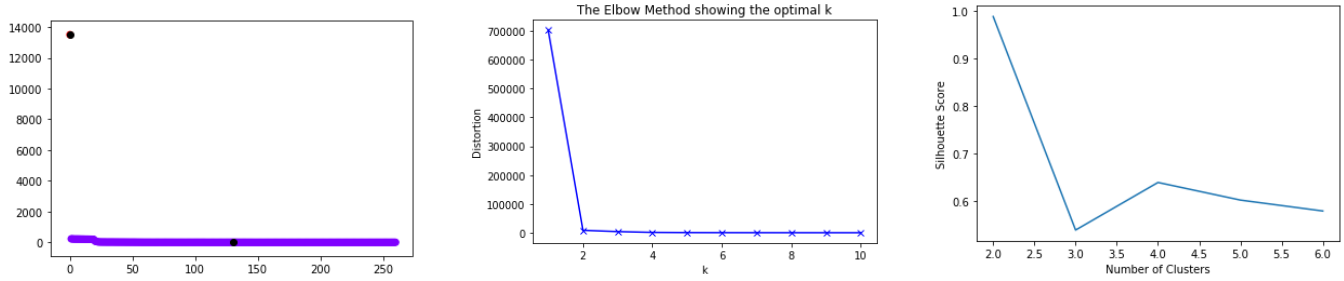
Solution:

In the case of well displaying, the IP addresses are replaced with index numbers, but the order is identical to that of Q2. In this section, Kmeans and Hierarchial Clustering algorithms are used to cluster, with the help of tools: Elbow Curve, Silhouette Curve and Hierarchical Analysis, to determine the number of clusters. The optimals are **2, 3 or 4 for SourceIP**, **2 for DestIP**.



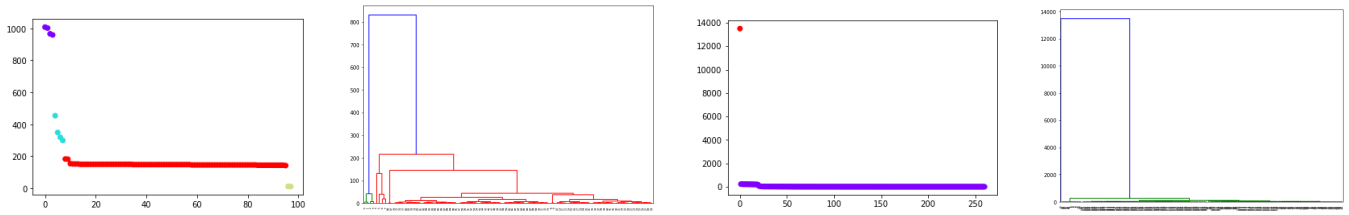
(a) Kmeans to SourceIP. (b) Elbow Curve to SourceIP. (c) Silhouette Curve of SourceIP.

Figure 3: Kmeans, Elbow Curve and Silhouette Curve to SourceIP.



(a) Kmeans to DestIP. (b) Elbow Curve to DestIP. (c) Silhouette Curve of DestIP.

Figure 4: Kmeans, Elbow Curve and Silhouette Curve to DestIP.



(a) Hierarchical Clustering of SourceIP. (b) Hierarchical Clustering Measurement of SourceIP. (c) Hierarchical Clustering of DestIP. (d) Hierarchical Clustering Measurement of DestIP.

Figure 5: Hierarchical Clustering and Measurements.

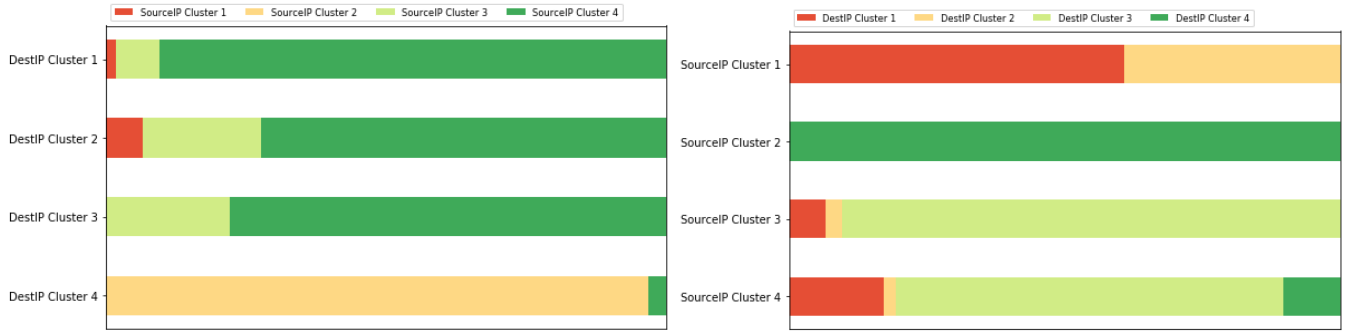
Problem 4

Using 4 clusters for source and destination IP addresses, respectively, investigate the relation between them, determine conditional probabilities and illustrate graphically.

Solution:

The solution of Question 4 is divided into several parts:

1. Clustering the IP addresses based on the number of occurrence;
2. For each IP in each cluster, i.e., X in SourceIP Cluster 1, check the status of its pairing IP: X DestIP, which DestIP Cluster does it belong to, record;
3. Calculate the probability distribution, then plot.



(a) DestIP to SourceIP.

(b) SourceIP to DestIP.

Figure 6: Conditional Probability.

Figure 6a and Figure 6b illustrate the conditional probability of the relation between DestIP and SourceIP clusters. The vertical labels represent the cluster needed to be investigated, the horizontal bars represent the proportion that how many IPs in the current cluster have contact with the other types of IP clusters. For example, there are 1.7% cases in DestIP Cluster 1, whose pairing IPs are in the SourceIP Cluster 1.

Table 1: Conditional Probability

P(Source|Dest)

	SC1	SC2	SC3	SC4
DC1	0.017	0	0.078	0.905
DC2	0.066	0	0.212	0.722
DC3	0	0	0.221	0.779
DC4	0	0.996	0	0.034

P(Dest|Source)

	DC1	DC2	DC3	DC4
SC1	0.610	0.390	0	0
SC2	0	0	0	1
SC3	0.067	0.030	0.903	0
SC4	0.172	0.023	0.702	0.103

Problem 5

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

Solution:

On the basis of Q4, the name of each IP is replaced with their relevant name of clusters in the data pre-processing, which are utilised as the attributes in the decision tree building process.

In this section, a binary decision tree based on CART algorithm is generated. In CART algorithm, the core function are GINI Impurity, Information Gain, which help find the best split. Listing 1 demonstrates its structure in the text form, the graphical form is illustrated in Q6 with that of dataset2 together.

The dataset1 is split into two parts: training set and test set, which are split randomly with ratio of 9:1. After the tree is built, the accuracy of training set and test set classification is calculated:

- The accuracy of training data is *97.98%*
- The accuracy of testing data is *97.88%*

In the result of classifying the test set, there are **7 unambiguous answers** in total, with the classification accuracy threshold of 1.

Explanation: The leaf of the decision tree represent the classification result and confidence. For instance, the second node on the true branch asks: *Is sourceIP = SourceIP Cluster 2?*

- **If true**, then there is a 97.2% that the label is **Generic Protocol Command Decode**, and 2.8% that the label is **Potential Corporate Privacy Violation**;
- **If false**, there is a 99.5% that the label is **Generic Protocol Command Decode**, and 0.5% that the label is **Potential Corporate Privacy Violation**.

```
Is destIP == DestIP Cluster 4?
--> True:
  Is sourceIP == SourceIP Cluster 2?
  --> True:
    Predict {' Generic Protocol Command Decode': 0.972,
             ' Potential Corporate Privacy Violation': 0.028}
  --> False:
    Predict {' Generic Protocol Command Decode': 0.995,
             ' Potential Corporate Privacy Violation': 0.005}
--> False:
  Predict {' Misc activity': 1.0}
```

Listing 1: Decision Tree

Problem 6
Examine the dataset coursework2.csv. Using the same clusters of IP addresses, are the patterns observed in Q4 still valid? How about the decision tree in Q5?

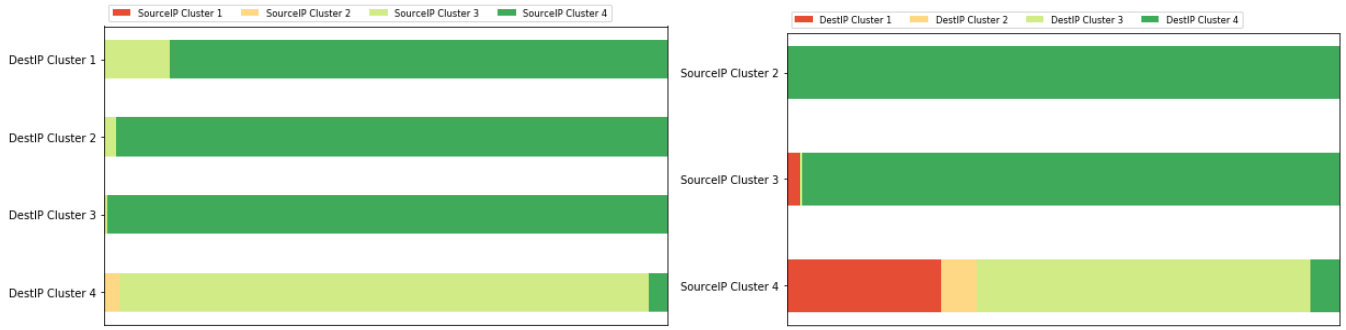
Solution: **Patterns** in coursework2.csv **vary** from that in coursework1.csv. **SourceIP Cluster 1 has no relation with other DestIP clusters.** As illustrated in Figure 7a and Figure 7b.

The decision tree generated on the basis of coursework2.csv is different with that in Q5, **which has 3 levels of nodes**.

The dataset2 is split into two parts: training set and test set, which are split randomly with ratio of 9:1. After the tree is built, the accuracy of training set and test set classification is calculated:

- The accuracy of training data is *99.25%*
- The accuracy of testing data is *99.38%*

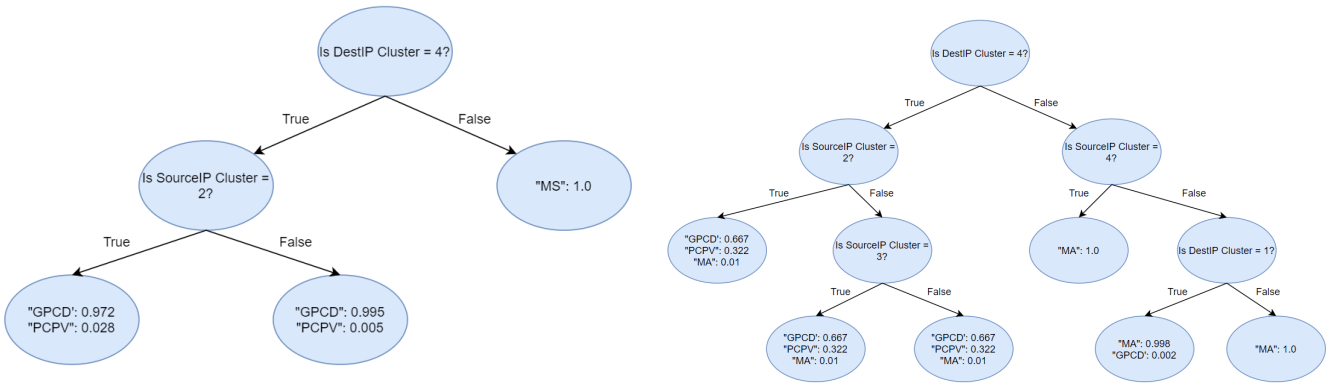
In the result of classifying the test set, there are **9 unambiguous answers** in total, with the classification accuracy threshold of 0.8.



(a) DestIP to SourceIP.

(b) SourceIP to DestIP.

Figure 7: Conditional Probability.



(a) Decision Tree of Dataset1.

(b) Decision Tree of Dataset2.

Figure 8: Decision Trees.