Exercise 1 – Machine learning for Communication Networks and Systems

1. You are given a set of rules an a file. Each row describe a different rule. The structure of the a rule is as follows (the @ sign is the first character of each rule):
   @SourceIP/sourceMask DestinationIP/destinationMask sourcePort destinationPort protocol

   - **sourceIP -** a 32 bit numbers. The **sourceMask** (a number in the range of 0 to 32) determined the number of relevant bits (starting from the most significant bit to the lowest significant bit). All the remaining bits are considered as "wild cards" (don't cares)
   - destination**IP -** a 32 bit numbers. The **destinationMask** (a number in the range of 0 to 32) determined the number of relevant bits (starting from the most significant bit to the lowest significant bit). All the remaining bits are considered as "wild cards" (don't cares)
   - **sourcePort** - a 16 bit range of numbers in the format of **low:high**
   - **detinationPort** - a 16 bit range of numbers in the format of **low:high**
   - **Protocol(s)** - an 8 bit numbers, each in the range of 0 to 255, separated with a "/" sign

   The file is given in a TSV format (i.e., the separator between any two fields of a rule is a Tab character ("\t") and the end of line with "\n"
   Example1 (here tabs are replaced by "\t"):
   @85.52.83.77/32\t25.184.48.129/32\t0:65535\t1490:1490\t0x06/0xFF\n
   Source IP address is equal to (all 32 bit count) 01010101001101000101001101001101
   Destination IP address is equal to (all 32 bit count) 00011001101110000011000010000001

   Example2 (the * denotes a wildcard)
   @85.52.83.126/31\t 25.184.52.0/22\t0:65535 0:65535\t0x06/0xFF\n
   Source IP address is equal to (only 31 bits) 01010101001101000101001101111111*
   Destination IP address is equal to (22 bit) 00011001101110000011010**********

1.1. After the above explanation, the file contains a set of n rules $R = \{r_1, r_2, \ldots, r_n\}$. The rules need to be classified using the technique of Decision Tree, using only bits from source IP and destination IP fields

1.2. Using the criteria of maximum information gain, as described in the lecture, find the analytical expression of the $IG(R_i, b_j)$, that is, the information gain about rule R_i, $i = 1,\ldots,n$, given a bit $b_j, j = 1,\ldots,64$ from the bits of the source ip ($j \in \{1,\ldots,32\}$) or from the destination IP ($j \in \{33,\ldots,64\}$). You have to consider that
   1.2.1. Number of rules is known
   1.2.2. Each address is 32 bits long
   1.2.3. There is a mask of known length for each rule
   1.2.4. You have to count for the effect of the wild cards
   What is the cost of the wildcards

1.3. Using the criteria developed in 1.2, write a code (python, Mathematica or Swift), that will compute the appropriate decision tree for all the versions given below

1.3.1. For each branch in the tree you take the best bit, so different branches may use different sequences of bits

1.3.2.Branches at the same level will use the bit that provides the best IG among all the tested bits, so all paths will use the same order of bits

1.3.3.For 1.3.1 and 1.3.2 above, the stoping point may be defined as getting a groups of rules all below a predefined length. We define the sizes of 16, 32, 64, 96 and 128. This means, that if a group contains up to the given size number of rules, it need not be split more (i.e., becomes a decision node).

2. Compare you results in question when the criteria is changed so the bit with the highest entropy is selected (in place of the highest IG): consider both cases – the best bit for each branch and the best bit for ll branches. How would you consider the effect of wildcards in this case?
3. In this question you may use a predefined function (if you find one, there is in Mathematica, I assume you could find one in Python as well) to classify the rules using the methods of Random Forest.
4. Now you are given DATA (packets) in a file with 2M packets. Each row of the file defines a packet with the following format (this is a TSV file, no headers are included, fields, as before, are "\t" and end of the line is "\n")

Source IP	Destination IP	Source Port	Destination Port	Protocol	Rule Number
3031143309	1518948592	46629	7125	6	511
3031143422	1439115900	42438	21	17	44

The rule number serves for training and testing. The Rule number (right most column) is the number of the line in the file of the rules (numbered from 0 to N-1). Suggest a way of training a classifier so the packets in the file are classified to the different rules). Use 80% percent of the data (1.6M packets) for classifications and 20% of the data (400k packets) for testing. Use at least two methods that were presented in the lectures. How good is your classification???