NETW504 2022 Project

Task2

Bayesian classifier

In this stage will use Naïve Bayes estimation to calculate the probability of an attack for each connection (raw)

1- You will need to split the dataset that you have into two sets a training set and test set. The training set should contain 80% of the data rows and the test set should contain 20% of the data rows The split should be random to ensure no bias (it should contain all possibilities)

2- For each of the numerical columns of the training data set find the pdf type (e.g uniform , Pareto , Exponential,,.....) and pdf parameters that provides the best fit for the pdfs obtained in the first phase of this project. These pdfs will include the pdf of each column, the pdf of each column given the attack and the pdf of each column given no attack. The best fit is the fit that provides the least mean square error between the pdf obtained in the first assignment and the fitted pdf. An example code that find the best pdf fit for a set of data can be found in the following link

- https://stackoverflow.com/a/37616966
- Note in this step you will have to carefully choose the range for the calculation of the numerical pdf that you did in assignment 1 in order to obtain an accurate fit of the data
- The result of this step should be provided in the form of list that contains the field name, the fitted pdf , and the pdf parameters.

3- For each of the categorical data in the training data set store the PMFs that you calculated in the previous assignment in data format that enables you to refer to them. Similarly you will have to do that for the PMF for each column, and the conditional pdf for each column

4- The naïve bayes estimation works as follows according to the following equation $Pr(Attack\ |given\ one\ row\ of\ the\ data)$

$$= \frac{pdf_{A|Attack}\,(a) *\ pdf_{B|Attack}\,(b) *\ pdf_{C|Attack}\,(c) *\ \ldots\ .*\ PMF_{Q|Attack}\,(q) *\ PMF_{R|Attack}\,(r) *\ \ldots\ *\ Pr(Attack)}{pdf_A\,(a)\ *\ Pdf_B\,(b)\ *\ Pdf_C\,(c)\ * \ldots..Pmf_Q\,(q)\ *\ pmf_R\,(r)\ * \ldots.}$$

- The above equation is calculated for a given row in the test dataset (which represents a given customer) • Note that $pdf_{A|Attack}\,(a)$ refers to the fitted conditional pdf of column A evaluated the value of this column for this row (for a given connection), and $pdf_{B|Attack}\,(b)$ refers to the conditional pdf of another column for the same connection.
- You can use as many numerical columns as you find suitable
- Note that $PMF_{Q|Attack}\,(q)$ refers to the fitted conditional pmf of column Q evaluated the value of this column for this row (for a given connection), and $PMF_{R|Attack}\,(r)$ refers to the conditional PMF of another column for the same customer.
- You can use as many categorical and discrete columns as you find suitable
- Note that the above rule assumes that columns A, B ,C .... , Q,R are independent and hence the quality of the estimation would degrade if they are dependent.
- You may also use the same equation to calculate the probability of no attack Pr(noAttack} |row) to check the quality of your answer
- It is preferred if the code wil allow easy inclusion or removal of columns from the calculation

5- The equation used in step 4 is repeated for all the test dataset and the result is compared to the answer in the test data to check the accuracy of your approach. In this case you will have calculate the number of correct estimations, the false positive and false negatives

6- The above steps have to be repeated for the all the attach types