Lab-Assignment-4

Problem Statement:

The example dataset i.e. the App dataset consists of two classes (M/B) having numerous text files which corresponds to execution traces of applications in the Operating System. The attributes are different calls invoked by the application while executing each sample for fixed duration. In this experiment you are required to do the following tasks:

[1] Understand the representation of features which are system calls (Boolean, Occurrence of calls or TF-IDF(Term Frequency and Inverse document frequency)). Prepare a machine learning model using different kernels and report which feature representation is better and subsequently which model gives improved performance. The output must be graphical indicating the performance of different ML model as well on diverse representations of feature vectors. Additionally, you are required to plot the ROC and report the output in tabular format consisting of precision, recall, F1_measure and Accuracy. Estimate the training and test for diverse settings of the experiments. **(Due Date of Submission: 27-10-2020)**

[2] Create categories of attributes where the call must be represented in the form of sequence of two system calls/three calls. Sequences are considered in sliding window fashion. Repeat the experiments of question [1] on this revised dataset. **(Due Date of Submission: 27-10-2020)**

Examples of feature: Let us consider a training example

Open, read, connect, ioctl, ioctl, mmap, mprotect, mmap, clone

Features are:

2 sequence: open-read, read-connect, connect-ioctl, ioctl-ioctl…etc

3 sequence: open-read-connect, read-connect-ioctl, connect-ioctl-ioctl etc

[3] Understand feature selection approaches implemented as a part of Sklearn

Examples of feature: Let us consider a training example

Open, read, connect, ioctl, ioctl, mmap, mprotect, mmap, clone

Features are:

2 sequence: open-read, read-connect, connect-ioctl, ioctl-ioctl…etc

3 sequence: open-read-connect, read-connect-ioctl, connect-ioctl-ioctl etc

Report which set of attributes are having high score. Plot and report the results

(Due Date of Submission: 27-10-2020)

[4] Implement the clustering techniques on the same dataset and report which clustering algorithm studied in the course generate better quality of clusters. Report the metrics to understand the quality of clusters **(Due Date of Submission: 01-11-2020)**

****

Faculty InCharge

Dr.Vinod P.