# CS 584 - Theory and Application of Data Mining Spring 2019
# HW 4 – Anomaly Detection

Name : Akshaya Damodaran (G01129364)
Registered Name on Miner: trytry77
Rank : 23
AUC: 1.0

Apr 21, 2019

## Problem

To implement the StrOUD algorithm, using LOF as the strangeness function for detecting outliers in the testdata that contains both anomalous and normal signal data obtained from a device that controls a centrifuge.

### Initial Steps

Firstly all baseline data from folders ModeA, ModeB, ModeC and ModeD) are fetched from the baseline dataset path, and collected in a DataFrame. Similarly, files from folder ModeM is are collected in another DataFrame.

### Dimensionality reduction using PCA

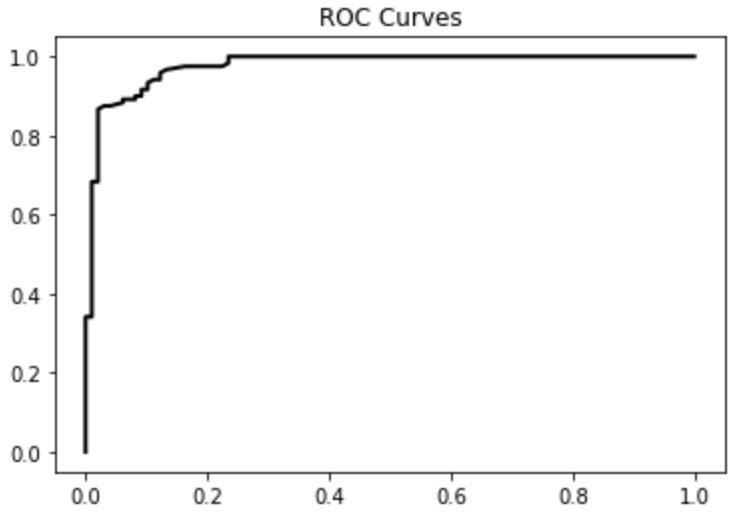
- For this problem, PCA is used for feature extraction to limit the number of features for performance reasons.

### Program Flow Outline

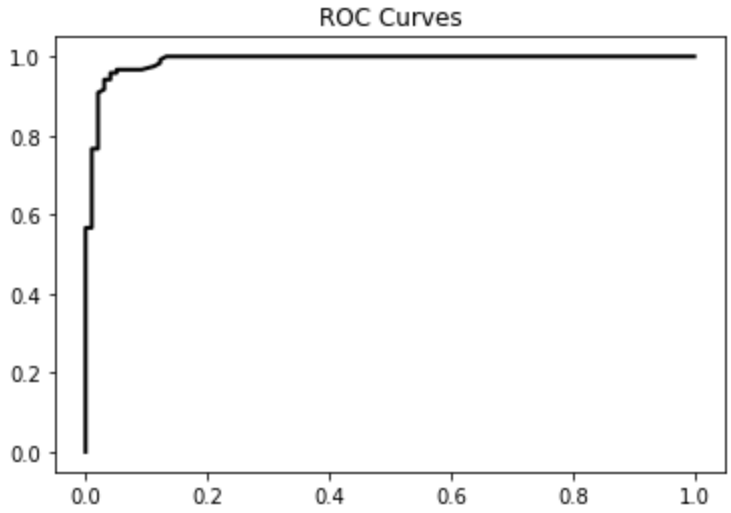
- The baseline data is split to get a portion that can be combined with the malicious data from ModeM.
- The portion of baseline data obtained from the previous step is then concatenated with the malicious data from ModeM and the result is put into a DataFrame. This forms our own test dataset.
- Distance matrices are constructed for the baseline and test dataset.
  The *k – distances* for each of the samples in the baseline and test datasets are computed. *K – distance* is the distance of the sample under consideration to the kth closest sample.
- Then the *local reachability densities* are then calculated for every sample in the baseline and test datasets. To get the *lrd* for a point *a*, we first calculate the *reachability distance* of *a* to all its *minPts* nearest neighbors and take the average of that number. The *lrd* is then simply the inverse of that average.
  This *reachability distance* measure is simply the maximum of the distance of two points and the *k-distance* of the second point.
- The LOF score for every sample in the baseline and test datasets is then computed as ratio of average local reachability density of the sample's k-nearest neighbors and local reachability density of the sample.
- Then the p-values are obtained as (b+1)/(N+1) for every sample in the test dataset by using the *lof* scores of the baseline where b is the number of samples in the baseline whose *lof* scores are greater than or equal to the *lof* score for the test sample under consideration and N is number of samples in the baseline.
- With a set confidence level, an ROC is constructed. Using the area under the curve, k is tuned to reach an area closer to 1.0
- Once we have tuned k, the above steps are performed on the actual test data and p-values for every sample is calculated.

## ROC for differenent values of k and minPts –

k = 12, minPts = 20




k = 10, minPts = 15




k = 7, minPts = 16