HW4 Anomaly Detection

Hsin-Ping Hsu (G01167652)

* Name registered on miner website: djmp
* Rank and score: 9/ 1.0
* Approach:

1. Read training data and split them for private testing: read normal data and reserve 100 normal data from different modes for private testing. Use the remaining normal data create a baseline.

Function: read\_sample()

1. Extract features via Fast Fourier transform: Extract features of baseline and test data through Fast Fourier transform from numpy library.

Function: fast\_fourier\_transform(baseline,test)

1. Implement LOF:
2. Compute and store the distance among each point. And get the distances of the Kth-nearest neighbor of each point (k\_dist).

Function: get\_k\_dist(data1, data2, K)

1. Store the sets of k nearest neighbors of each point into a 2D array, k\_neighbors.

Function: get\_k\_neighbors (data1,data2,dist,k\_dist)

1. Compute local reachability distance for each point, which is the inverse of the reachability distance of each point from its neighbors (through k\_dist and k\_neighbors).

Function: get\_lrd(data1,data2,k\_neighbors,dist)

1. Compute local density factor for each point with its lrd and the lrd of its k nearest neighbors.

Function: get\_lof(lrd\_baseline,lrd,k\_neighbors)

1. Compute and store LOF for each point in baseline.
2. Implement StrOUD:
3. Compute the LOF of each data point in test data with respect to the baseline.
4. Calculate the p value for each data point in test data by comparing the LOF of each of them with the LOF of baseline. ((b+1)/ (N+1))

Function: StrOUD(test,baseline,lof\_baseline,lrd\_baseline,k\_neighbors,dist\_test)

1. Compute the area of roc curve to determine the best K: use the roc\_curve library from sklearn and visualize the roc curves and calculate the areas under the curves for different values of K, then use the result to find the best K for this model.

**Results:**

By calculating the areas under the roc curves, I observed the best value of K is 8.

|  |  |  |
| --- | --- | --- |
| **一張含有 螢幕擷取畫面 的圖片  自動產生的描述** | **一張含有 螢幕擷取畫面 的圖片  自動產生的描述** | **一張含有 螢幕擷取畫面 的圖片  自動產生的描述** |
| Fig1. ROC curve (K=5~20) | | |

**References:**

1. <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html>
2. <https://docs.scipy.org/doc/numpy/reference/generated/numpy.fft.fft.html#numpy.fft.fft>
3. <https://plot.ly/matplotlib/fft/>
4. <https://cs.gmu.edu/~dbarbara/CS584/rtfp719-domeniconi.pdf>
5. <http://www.dbs.ifi.lmu.de/Publikationen/Papers/LOF.pdf>