Birla Institute of Technology & Science, Pilani
Hyderabad Campus

SECOND SEMESTER 2017-2018
CS F415: DATA MINING
Assignment 3

Submission Date & Time: 22/04/18, 2359 hrs Maximum Marks: 25

Very often, there exist data objects that do not comply with the general behaviour or model of the data. Such data objects, which are grossly different from or inconsistent with the remaining set of data, are called outliers or anomalies. Outlier detection can be classified into four categories: *statistical*, *distance-based*, *density-based*, and *deviation-based* approach. Statistical and distance-based outlier detection both depend on the overall or global distribution of the given set of data points. However, data are usually not uniformly distributed. These methods encounter difficulties when analysing data with rather different density distributions, so in this assignment you will apply **density-based outlier detection to find frauds in credit card transactions**. For this, you can use any of the below mentioned techniques:

- Local Outlier Factor (LOF) technique
- Relative Density based Outlier detection - Section 10.4 in Introduction to Data Mining – Pang Ning Tan
- DBSCAN – Clustering the points using DBSCAN and declaring all points not assigned to any cluster as outliers - Section 10.4 in Introduction to Data Mining – Pang Ning Tan

Dataset: Credit Card Fraud Detection

Programming Languages: Python, Java, C/C++

Team Size: 3

Report:

- Name and ID of team members.
- Brief description of algorithm used for outlier detection
- Pre-processing done on the data (if any).
- 2D Plot of the dataset for different parameters (epsilon for DBSCAN or k for LOF and Relative Density based outlier detection) highlighting outliers (**NOTE:** For plotting, you can use dimension reduction libraries to reduce to two dimensions and then plot the new dataset)
- Accuracy (last attribute in each row of the dataset is the actual classification as outlier or not)

Submission Files:

- Source code files
- Image files of the dendrogram plot
- Report in PDF format
- README

Remarks:

- All submission documents should be zipped together and submitted to CMS through one of the group member's account before deadline. Name of the file should be DM_ASSN3_201x0xxx_201x0xxx_201x0xxx.zip
- All source codes will be checked for PLAGIARISM on Moss (for a Measure of Software Similarity). Any kind of plagiarism will not be entertained.
- You are expected to demo your code and present your results as per the schedule that will be made available on CMS later.

Evaluation:

- Code & comments (15 marks)
- Report (5 marks)
- Viva (5 marks)

Please contact following teaching assistants for any queries:

1. Keval Morabia (f20150143@hyderabad.bits-pilani.ac.in)