C S 488/508 Introduction to Data Mining
Homework 8: Anomaly detection

Objective

In this **individual** homework, you will do exercises to write program to utilize anomaly detection algorithms.

Requirements

This assignment uses a data set called Breast Cancer Wisconsin Data Set from the UCI machine learning repository. Its csv file can be downloaded from Canvas (breast-cancer-wisconsin.csv). For this dataset, we want to detect anomaly class (i.e., maglinant). Refer to the data set description at `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)` for more details about its attribute information and instances.

Q1. (40 points) (**Local Outlier Factor**) Remove rows with missing values. Perform unsupervised outlier detection using Local Outlier Factor (LOF) with number of neighbors = 10 and metric = Euclidean distance. Use default values for other parameters. Plot the ROC curve. Report the processing time and AUC. (Reference: `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html`)

Q2. (40 points) (**Density-based Approach**) Remove rows with missing values. Perform unsupervised outlier detection using DBSCAN. Consider noisy samples with cluster -1 as outliers. Properly choose values for parameters *eps* and *min_samples*. Use default values for other parameters. Plot the ROC curve. Report the processing time and AUC. (Reference: `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html`)

Q3. (20 points) (**Comparison**) Which method is more accurate and faster?

Submission instructions

A zipped file `hw-lastname.zip` consisting of all the source code and the PDF files containing discussions and figures.

Grading criteria

(1) The score allocation has been put beside the questions.

(2) Please make sure that you test your code **thoroughly**.

(3) FIVE points will be deducted if files are not submitted in the required format.