Programming Assignment #2

COEN 281 Pattern Recognition and Data Mining
Department of Computer Engineering
Santa Clara University

Dr. Ming-Hwa Wang	Summer Quarter 2017
Phone: (408) 805-4175	Email address: m1wang@scu.edu
Course website:	http://www.cse.scu.edu/~mwang2/mining/
Office Hours:	Wednesday & Friday 9:00am-9:30am

Due date: Midnight August 2, 2017

Outlier Detection in Streaming Data Using LOF Scores (200 points)

Please implement outlier detection in streaming data using local outlier factor (LOF) score. You should implement it in C, C++ or Java.

Your program should take input from stdin. The input contains a window size w (32-bit integer), a <host>:<port> pair where you receive an input stream of fixed-d-dimensional data points in a comma separated values (CSV) format. Your program should output outliers starting from the $(w+1)^{th}$ input data until end of input stream. Your program should handle concept drift too.

The LOF approach is a normalized distance-based approach. It adjusts for local variations in cluster density by normalizing distances with the average point-specific distances in a data locality. For a given data point $\boldsymbol{x}$, let $v_k(\boldsymbol{x})$ be the distance to its k-nearest neighbor, and let $L_k(\boldsymbol{x})$ be the set of points within the k-nearest neighbor distance of $\boldsymbol{x}$. $|L_k(\boldsymbol{x})| \geq k$ because of ties in the distance. Then, the asymmetric reachability distance $r_k(\boldsymbol{x}, \boldsymbol{y})$ of object $\boldsymbol{x}$ with respect to $\boldsymbol{y}$ is defined as $r_k(\boldsymbol{x}, \boldsymbol{y}) = \max\{\text{Dist}(\boldsymbol{x}, \boldsymbol{y}), v_k(\boldsymbol{y})\}$. When $\boldsymbol{y}$ is in a dense region and the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is large, $r_k(\boldsymbol{x}, \boldsymbol{y})$ is equal to the true distance $\text{Dist}(\boldsymbol{x}, \boldsymbol{y})$. When the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ is small, then $r_k(\boldsymbol{x}, \boldsymbol{y})$ is smoothed out by the k-nearest neighbor distance of $\boldsymbol{y}$. The larger the value of k, the greater the smoothing. The average reachability distance $ar_k(\boldsymbol{x}) = \text{MEAN}_{\boldsymbol{y} \in L_k(\boldsymbol{x})} r_k(\boldsymbol{x}, \boldsymbol{y})$, and $\text{LOF}_k(\boldsymbol{x}) = \text{MEAN}_{\boldsymbol{y} \in L_k(\boldsymbol{x})}(ar_k(\boldsymbol{x}) \,/\, ar_k(\boldsymbol{y}))$. The maximum value of $\text{LOF}_k(\boldsymbol{x})$ over a range of different values of k is used as the outlier score to determine the best size of the neighborhood. To handle streaming data with sliding window, we extend LOF to incremental scenarios: 1) the statistic of the newly inserted data points are computed, 2) only the LOF scores of the affected data points by the newly inserted data point in the existing data points in the window are updated, and 3) similarly updated the deleted data points.

The data point with LOF score greater than a threshold t will be reported as outlier.

Student Name:

SSN/ID:

Score:

Correctness and boundary condition (60%):

Compiling without warnings (5%):

Error Handling (5%):

Modular design, file/directory organizing, showing input, documentation, coding standards (25%):

Automation (5%):

Subtotal:

Late penalty (20% per day):

Special service penalty (5%):

Total score: