INFS 5102 – Unsupervised Methods in Analytics

Homework: Anomaly Detection

Objective:

To consolidate the understanding of anomaly/outlier score and anomaly detection methods

Answer the following questions

1. Assume that you are given a dataset with N one-dimensional objects (scalar values). Using an equal-depth histogram, design a way to assign an object an outlier score.

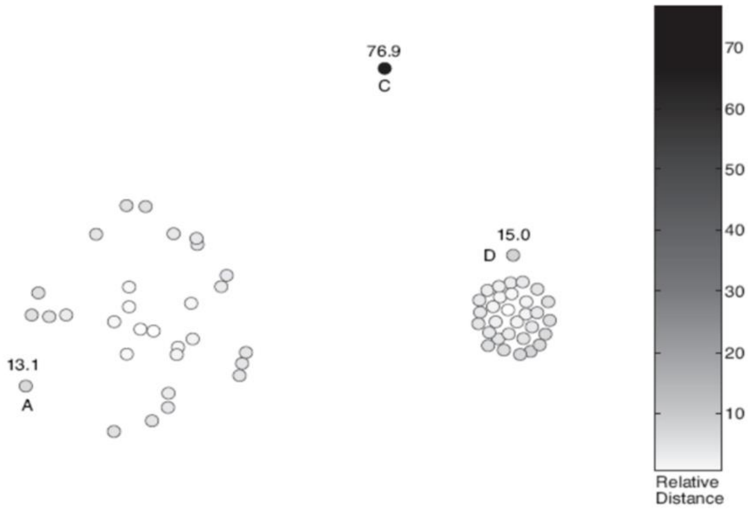
   (**Hints**: An outlier/anomaly score is expected to measure the outlierness/anomalousness of an object. Refer to Module 2 (Data) slides for details of equal-depth (frequency) histogram/binning.)

2. Consider a K-means scheme for outlier detection which defines the outlier score of a point as the relative distance of the point from its closest centroid, where the relative distance is the ratio of the point’s distance from its closest centroid to the median distance of all points in the cluster from the centroid.

   (a) The points at the bottom of the compact cluster shown in the figure below have a somewhat higher outlier score than those points at the top of the compact cluster. Why? (**Note**: k=2 was used in the clustering)

   (b) Suppose that we choose the number of clusters to be much larger, e.g., 10. Would the proposed technique still be effective in finding the most extreme outlier at the top of the figure? Why or why not?

   (c) The use of relative distance adjusts for differences in density. Give an example of where such an approach might lead to the wrong conclusion.



3. Many statistical tests for outliers were developed in an environment in which a few hundred observations was a large data set. With this question, we discuss the limitations of such approaches.

   (a) For a set of 1,000,000 values, how many outliers we would have according to the test that says a value is an outlier if it is more than three standard deviations from the average? (Assume a normal distribution.)

   (b) Does the approach that states an outlier is an object of unusually low probability need to be adjusted when dealing with large data sets? If so, how?