Dr. Eick

COSC 4335*“Data Mining”* Assignment 4 Fall 2018

*Design, Implementation and Comparison of Two Outlier Detection Techniques*

*for a Spatial Dataset*

Group Project (Groups of 2)

Due date: Wednesday, Dec. 5 at noon; submissions that are received by Monday, December 3 at noon receive a 3% bonus!

Last updated: November 13, 9pm!



Complex8\_RN15 available at: <https://drive.google.com/open?id=1_geQDIMQUNHhc3d7zRxfOrOreTT3HBhU>

The goal of the project is to design and implement two[[1]](#footnote-1) bivariate, spatial outlier detection techniques of your own preference—each student develops a different one. You will apply the technique of your choice to the 2-dimensional dataset called *Complex8\_RN15*[[2]](#footnote-2), is the variation of the Complex8 dataset with 15% gaussian noise added to the original Complex8 dataset. Finally, you will compare the results with those of your team mate and write a summary report. Your outlier detection techniques should take the dataset and create a copy of the dataset that contains an additional column/attribute called *ols (“outlier score”)* which contains numbers that indicates how much your outlier detection method believes that the particular object is an outlier—the smaller the value of the ols attribute the less likely the object is believed to be an outlier. For the project, you can use any R-library or any other software library to accomplish the project tasks; just acknowledge what external software you used in your project report.

The Complex8\_RN15 dataset has attributes x, y, class; for example, after applying your outlier detection technique to 5 examples of the dataset the result produced by the method you are supposed to implement could look as follows:

563.225,56.748,0,0.24

564.887,58.119,0,0.41

565.434,68.061,0,0.12

565.926,79.953,0,0.33

566.762,69.405,0,0.11

This depicted result indicates that the second example is the most likely outlier, the fourth example is the second most likely outlier,…, and that the last example is the least likely outlier.

**Assignment4 Tasks:**

Task 1: Visualize the Complex8\_RN15 dataset; visualize the third attribute using different colors, like supervised scatterplots we used in Assignments 1/2. 3 pts

Task 2: Design and implement 2 outlier detection techniques—one per student—for the Complex8\_RN15 dataset. 10 pts (5 for each technique)

As explained earlier your implementation should add a column/attribute *ols* to the dataset and fill this column with numbers.

Task 3: Evaluation (one for each technique)

a. Apply our outlier detection to the Complex8\_RN15 dataset obtaining a new file X; your outlier detection method is only applied to attributes x and y of the dataset, and ignores the attribute named class and adds an ols-column to the dataset. 3 pts

b. Sort X in descending order based on the values of attribute ols (the example with the highest ols value/the example that is the most likely outlier should be the first entry in X)! 3 pts

c. Visualize the first 7% of the observations in X, just displaying their x and y value and the class using a different color for each class, in a display and the remaining 93% of the observations in a second display. In general, the first display visualizes the outliers and the second display visualizes the normal observations in the dataset. 3 pts

d. Visualize the first 14% of the observations in X, just displaying the x and y value and the class using a different color for each class and the remaining 86% of the observations in a second display. 3 pts

e. Visualize the first 21% of the observations in X, just displaying the x and y value and class using a different color for each class, in a display and the remaining 79% of the observations in a second display. 3 pts

f. Interpret the 6 displays you generated in steps c-e; particularly, assess how well does your outlier detection method work—intuitively observations that are quite far a way of the 8 natural clusters of the original Complex8 dataset should be outliers. Also try to characterize which points are picked as outliers as the first 7%, 14%, and 21%, respectively 9 pts

g. Create a histogram for the ols values of the top 21% entries in file X. Briefly, interpret the obtained histogram! 5 pts

Task 4: Write 2-5 paragraphs (for each technique), explaining each outlier detection technique works and how it was implemented. If you enhanced your approach based on feedback to get better results also describe how you enhanced your technique. If your outlier detection technique needs the selection of parameter values before it can be run, describe how you selected those parameter values. Moreover, mention in an additional paragraph what (if any) external software packages your used in the project! 10 pts

Task 5: Compare the two outlier detection techniques you developed in 2-4 paragraphs. Try to assess which technique worked better. Finally, write 1-2 paragraphs, describing what you leant from the project and if you enjoyed it or not. A significant amount of points will be allocated to this comparison; therefore, try to do a good job in comparing and evaluating your two techniques. 15 pts.

There will also be up to 6 bonus pts for very sophisticated approaches.

Submit the code of the implementation of your outlier detection techniques in a separate file! Finally, paste all project results for the different tasks into a single file and also submit this file at the due date! We only want one project report (either in Word or pdf format) per group! Please upload your submission on Blackboard and name the submission file as HW4\_Group\_# (# denotes the group number).

1. Groups of 3 implement 3 different outlier detection techniques! [↑](#footnote-ref-1)
2. The outlier detection method is only applied to the first 2 numerical attributes and the third attribute is just used for visualization purposes; for that reason, it is called a spatial outlier detection method; e.g. it can detect outliers for object locations described by (longitude, latitude)-pairs. [↑](#footnote-ref-2)