CST8390 - Lab 5
Outlier Detection

Due Date: Check Brightspace for due dates.

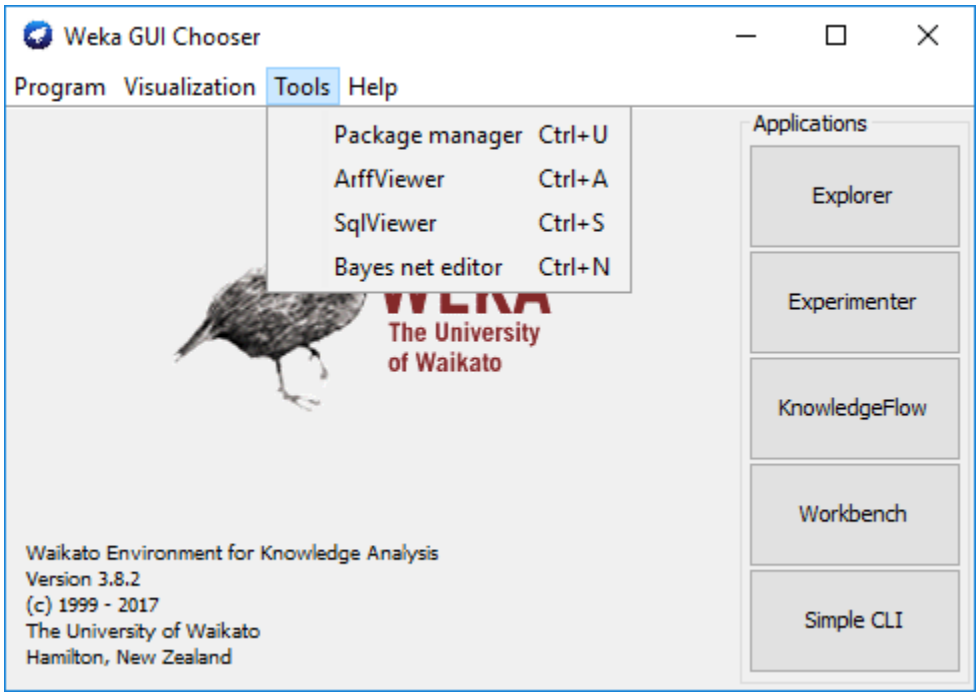
Introduction

The goal of this lab is to perform outlier detection on Salary File using Local Outlier Factor and Isolation Forest.

Steps:

Local Outlier Factor

1. With Weka 3.8, outlier detection methods like Local Outlier Factor and Isolation Forest are not included. But they are available as packages to be installed using Package Manager of Weka.

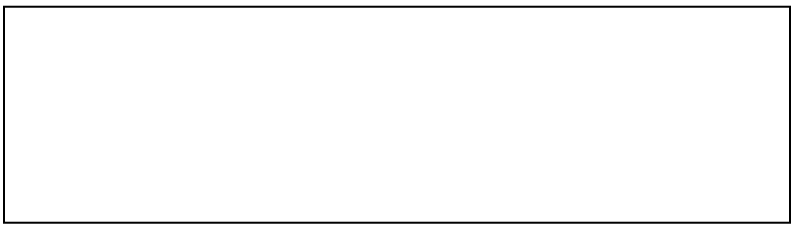


   From the package manager, install localOutlierFactor and isolationForest (find the package in the big list, select it, and hit install).

2. Now, download EmployeesSalaryOriginalOutlier.csv file from Brightspace and load it into Weka explorer. If everything worked well, you should be able to see Local Outlier Factor and Isolation Forest listed as classifiers under weka → classifiers → misc on Classify tab.

3. Make sure that all attributes are loaded with right data types. If not, apply filters to convert them. Save the file as EmployeesSalaryOriginalOutlier.arff.
   (Expectation: ID, first_name, last_name, email, Address - String,
   Country, Branch and Currency, Outlier - Nominal and
   salary - Numeric).

4. We are going to perform outlier detection on this file. There are some attributes that are not relevant for outlier detection. Identify and remove those attributes. List the names of **removed** attributes below:



5. Run addID filter to create an ID column.

6. Implementation of outlier detection methods in Weka needs a class attribute. So, we will use Outlier as the class attribute. In order to detect outliers using Local Outlier Factor, you need to select it from weka → classfiers → misc on Classify tab. You need to select 10-fold cross validation and Outlier as the class attribute.

7. Right click on the result in the result pane and click on "Visualize classifier errors" and save the file as LOF_Results.arff.

8. Now, open another explorer and open LOF_Results.arff. Two more attributes are created by LOF, namely prediction margin and predicted outlier. You have a few instances predicted as outliers. Hit Edit to open Viewer. Sort Predicted Outlier and see how many of actual outliers are predicted as outliers.

Isolation Forest

9. Open **another** explorer and load EmployeesSalaryOriginalOutlier.arff from step 3. Remove all irrelevant attributes. Make sure you have the right data types.

10. Convert all nominal attributes except Outlier to binary using filter.

11. Run addID filter to create an ID column.

12. Run Isolation Forest by setting "Use training set" as the test option and Outlier as the class attribute.

13. Right click on the result in the result pane and click on "Visualize classifier errors" and save the file as ISF_Results.arff.

14. Now, open another explorer and open ISF_Results.arff. Two more attributes are created by LOF, namely prediction margin and predicted outlier. You have a few instances predicted as outliers. Hit Edit to open Viewer. Sort Predicted Outlier and see how many of actual outliers are predicted as outliers.

15. **Combine Results**

16. Open EmployeesSalaryOriginalOutlier.csv and save it as Results.xlsx.

17. Open both results file in Notepad++. Copy results from LOF_Results into another sheet. Use text to columns to convert data into columns. Add header row based on the header info in the arff file. Give LOF prefix for the new columns created. Sort it based on the ID column. Copy and paste new columns into the first sheet of Results.xlsx.

18. Next, copy results from ISF_Results from arff file into another sheet and do the same as in step 13. Give ISF prefix for new columns created. Copy and paste new columns into the first sheet of Results.xlsx.

19. Now you have both results along with the data in one sheet. Replace all Yes with 1 and No with 0 (use find & replace).

20. Create a new column named Ensemble and apply formula that calculates the sum of LOF: predicted Outlier and ISF: predicted Outlier for this column.

21. Select the sheet and sort it from Largest to Smallest based on Ensemble column. Your header row of combined sheet should look like:

Id	first_ name	last_ name	email	Address	Country	Branch	Currency	Salary	Outlier	LOF:prediction margin	LOF:predicted Outlier	ISF:prediction margin	ISF:predicted Outlier	Ensemble

22. Create a new column named Reason and **based on your judgement,** record the reason for the instances that are predicted as outlier by both methods.

In order to get grades,

1. You should be ready with your explorers for LOF, ISF, LOF results, ISF results and the Results excel file.
2. Submit a screenshot of the sorted (based on ensemble) Results file (first 20 instances should be visible) to Brightspace. (submit just an image file and not a zipped folder)