Arthur Dunbar and Christoph F. Eick

Task4: Outlier Detection for a Houston Weather Dataset

Individual Task

Second Draft

Due Date: October 22nd, 2024

Figure 2: Some Unusual Weather

Last updated: Oct. 8, 10a

In this task you will be developing distance-based outlier detection technique for Houston Weather dataset HW2021, which reports minimum temperature, humidity, rainfall, wind speed and cloud cover for the 365 days of the year 2021; the objective is to find "*unusual weather days*" in this dataset.

Dataset Description:

Houston_Weather Dataset has the following attributes:

DATE / nominal / Each record has a date starting from 01/01/2021 to 12/31/2021

cloudcover / categorical / %/ 17 different types of cloud cover. Categories are: "Fair", "Fair / Windy", "Partly Cloudy", "Partly Cloudy / Windy", "Cloudy", "Cloudy / Windy","Mostly Cloudy","Mostly Cloudy / Windy","Fog","Haze", "Light Rain", "Light Rain with Thunder", "Thunder", "Rain" "Thunder / Windy" "Heavy T-Storm", "Thunder in the Vicinity", "T-Storm"

rainfall / continuous / inch / Amount of rainfall of the day/ from 0 to 5

min_temp / continuous / Fahrenheit / Minimum temperature at 3pm / from 34 to 83

wind_speed / continuous / mile per hour / wind speed at 3pm/ from 0 to 29

humidity / continuous / % / Humidity at 3pm/ from 0 to 100

Subtasks:

a. Design a "good" distance function for HW2021!
b. Design and implement a distance-based outlier detection technique for HW2021! The technique if applied to the HW2021 dataset should add a column to the examples in the dataset named OLS (Outlier Score) which contains a single number which measures the strength of our belief that the particular example is an outlier.
c. Apply the outlier detection technique to the HW2021 dataset. Since distance-based outlier detection techniques use hyper parameters: apply your technique 3 times to the

dataset using 3 different hyper parameter settings, obtaining three different augmented HW2021datasets with the OLS column added.

d. Sort the three obtained augmented datasets using the OLS attribute. Discuss the top 4 examples of each augmented dataset; explain why you believe these particular examples were viewed as likely outlier candidates. Also discuss the bottom 2 examples in each augmented dataset; try to explain why these two examples were rated to be “most normal” ones.
e. Briefly assess how well your outlier detection technique worked.
f. Write a 2-page single spaced report which summarizes the main findings of Task 4. Clearly describe the distance function you designed in task a; in particular discuss how you assess similarity for the ordinal cloud cover attribute and the numerical for other attributes for the dataset. Next describe the outlier detection you developed in task b. Finally, discuss your findings with respect to tasks c through e.

Deliverables for Task 4:

A. Report
B. File witch contains the software you developed for Task4.

Figure 2: More Unusual Weather

Task 4 Submission Guidelines:

1. Name your python/R files to **COSC3337F24-PS2T4-Firstname-Lastname.ipynb** or any other appropriate extension.
2. Name the pdf copy of your report **COSC3337F24-PS2T4-Report-Firstname-Lastname.pdf** carefully.
3. Create a folder and name it **COSC3337F24-PS2T4-Firstname-Lastname**.The folder should contain both python/R file and pdf copy of your report named correctly. Compress (zip) the folder and submit it to MS TEAMS.

4. Upload the zipped folder to the Assignment tab in MS Teams **before the deadline**.