Janet Anagli, Md. Mahin and Christoph F. Eick

Data Science I Dr. Eick

Fall 2025

Task4: Outlier Detection for a Houston 2023 Weather Dataset

Individual Task



Figure 1: Some Unusual Weather

Last updated: September 26, 2025

Responsible TA: Janet Anagli

Due Date: Thusrday, October 30 end of the day

In this task you will be developing distance-based outlier detection technique for Houston Weather dataset HW2023—we already used in Task1—, which reports maximum temperature, humidity, visibility, cloudiness, wind speed and cloudiness and rain for the 365 days of the year 2023; the objective of this task 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

**cloudiness** in percent

**rain** / continuous / inch / Amount of rainfall of the day; no entry means ‘no rain’

**max\_temp** / continuous / Fahrenheit / maximum temperature of that day

**wind\_speed** / continuous / mile per hour / wind speed at 3pm/ from 0 to 29

**visibility** / continuous / in meter

**humidity** / continuous / % / relative humidity at 3pm/ from 0 to 100

Subtasks:

1. Design a “good” distance function for HW2023!
2. Design and implement a distance-based outlier detection technique for HW2023! The technique if applied to the HW2023 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.
3. Apply the outlier detection technique to the HW2023 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 HW2023 datasets with the OLS column added.
4. 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 example in each augmented dataset; try to explain why this example was rated to be “most normal” ones.
5. Briefly assess how well your outlier detection technique worked.
6. 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 of the tuples in the dataset. Next, describe the outlier detection you developed in task b. Finally, discuss summarize you findings with respect to tasks c through e.

**Deliverables for Task 4:**

1. Report
2. File witch contains the software you developed for Task4.

A car driving on a road with clouds

Description automatically generated

Figure 2: More Unusual Weather

**Task 4 Submission Guidelines:**

1. Name your python/R files to **COSC3337F25-PS2T4-Firstname-Lastname.ipynb** or any other appropriate extension.
2. Name the pdf copy of your report **COSC3337F25-PS2T4-Report-Firstname-Lastname.pdf** carefully.
3. Create a folder and name it **COSC3337F25-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**.