Dr. Eick and Md Mahin

COSC 6335*“Data Mining”*

ProblemSet2 Fall 2024

Clustering and Outlier Detection

Task2: Outlier Detection for a Houston Weather Dataset

Individual Task

Peer Reviewed

First Draft



Fig. 2: Some Unusual Weather

Last updated; Sept. 26, 10a

Deadline: **October 15**, 2024

In this task you will be developing outlier detection techniques for a Houston Weather Dataset; the objective is to find “*unusual weather days*” in this dataset.

A day can be unusual if it's much hotter or colder than usual (temperature), windier or calmer than usual (wind speed), more humid or less humid than usual (humidity), or wetter or drier than usual (rainfall). Each of these things can affect our daily lives. For example, a very hot day in winter or a very cold day in summer would be unusual. Or, if it rains a lot more or a lot less than normal, that could also be unusual. To know if a day is unusual, we need to compare it to what's typical for the location.

In this task, we will use a dataset called the Houston weather dataset. It contains daily weather data for Houston in the year 2021, with attributes like date, min\_temp, max\_temp, rainfall, wind\_speed9am, wind\_speed3pm, humidity9am, humidity3pm, pressure9am, pressure3pm, cloud9am, cloud3pm, temp9am, temp3pm, rain\_today, and rain\_tomorrow. However, for this task, we will focus on a subset of the dataset called RHOUSTONW. This subset includes the following attributes: Date, min\_temp, max\_temp, rainfall, wind\_speed, humidity, and cloud. In the dataset, wind\_speed and humidity refer to wind\_speed3pm and humidity3pm, while cloud is the numerical conversion of cloud3pm from the original dataset.

Houston\_Weather Dataset has the the following attributes:

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

**cloud** / categorical / %/ 17 different types of cloud cover. Categories are 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 / farenhit / Minimum temperture at 3pm / from 34 to 83

**max\_temp** / continuous / farenhit / Maximum temperture at 3pm/ from 46 to 98

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

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

3 Examples in the Weather Dataset:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| date | min\_temp | max\_temp | rainfall | wind\_speed | humidity | cloud |
| 1/1/2021 | 41 | 55 | 0 | 8 | 51 | Mostly Cloudy |
| 1/2/2021 | 41 | 59 | 0 | 7 | 42 | Fair |
| 1/3/2021 | 43 | 68 | 0 | 13 | 37 | Fair |

Subtasks:

1. Design and implement a distance-based and a model/density-based object outlier detection technique for the Houston Weather Dataset. The technique if applied to the Houston Weather 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. The challenge for the first task will be the development of a “good” distance function for the RHOUSTONW dataset; the challenge for the second task will be to develop a “good” density function for the RHOUSTONW dataset. \*\*\*\*\*\*\*\*\*\*\*
	* You must design a multivariate distance function and a multivariate density function that has been tailored to the dataset. You can also use clustering algorithms, but in such case marks related to density function and distance function would be zero.
	* Please provide clear definition of the distance and density function you designed and describe and justify your design choices.
2. Apply the two outlier detection techniques to the RHOUSTONW dataset; if your methods involves hyper parameters, apply the methods 3 times to the dataset using 3 different hyper parameter settings. \*\*\*\*
3. Sort the obtained augmented RHOUSTONW Datasets using the OLS attribute. Discuss the top 3 examples of each augmented dataset; explain why you believe the particular examples were viewed as likely outlier. Also discuss the bottom example in each augmented dataset: try to explain why twere rated to be “most normal”.\*\*\*\*
4. Based on the results you obtained in the previous steps evaluate and compare the two outlier detection techniques you developed. \*\*
5. If necessary, enhance your two outlier detection techniques and redo steps d, e, and f!

 **Deliverables for Task 2:**

1. Indivdual task solutions will be submitted via Kritik.
2. Your submission should include both the code and the analysis, and it should be structured in a way that is easy for your peers to review.

Rubrics:

**a Q**: Design and implement a distance-based and a model/density-based object outlier detection technique for the Houston Weather Dataset. The technique if applied to the Houston Weather 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. The challenge for the first task will be the development of a “good” distance function for the RHOUSTONW dataset; the challenge for the second task will be to develop a “good” density function for the RHOUSTONW dataset. \*\*\*\*\*\*\*\*\*\*\*

* + You must design a multivariate distance function and a multivariate density function that has been tailored to the dataset.
	+ Please provide clear definition of the distance and density function you designed and describe and justify your design choices.

Deliverable:

1. Properly commented code. [Add comments above each block. Make variable and function names big enough to understand their purpose. And Add a doc section at beginning of each module describing their inputs, outputs, and briefly mention what they will do and how they will do ]
2. Explanation containing
	* 1. Algorithm/Psudocode that explain your detection mechanism
		2. Explanation how the algorithm works
		3. Example input and output and discussion of input/output

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| Quality of the Distance function  | No Distance function is presented | The Distance function is not very sophisticated/incorrect and will produce wrong outputs in most cases | The Distance function is modestly sophisticated/incorrect and will produce wrong outputs in some cases | The Distance function is very good | 4 |
| Distance-based outlier detection technique Quality | No distance-based outlier detection technique is presented | The distance-based outlier detection technique is not very sophisticated/incorrect and will produce wrong outputs in most cases | The distance-based outlier detection technique is modestly sophisticated/incorrect and will produce wrong outputs in some cases | The distance-based outlier detection technique is very good | 4 |
| Quality of the Density function  | No Density function is presented | The Density function is not very sophisticated/incorrect and will produce wrong outputs in most cases | The Density function is modestly sophisticated/incorrect and will produce wrong outputs in some cases | The Density function is very good | 4 |
| Model/density -based outlier detection technique Quality | No Model/density -based outlier detection technique is presented | The Model/density -based outlier detection technique is not very sophisticated/incorrect and will produce wrong outputs in most cases | The Model/density -based outlier detection technique is modestly sophisticated/incorrect and will produce wrong outputs in some cases | The Model/density -based outlier detection technique is very good | 4 |

**b Q**: Apply the two outlier detection techniques to the RHOUSTONW dataset; if your methods involves hyper parameters, apply the methods 3 times to the dataset using 3 different hyper parameter settings.
Deliverable:

1. Properly commented code. [Add comments above each block. Make variable and function names big enough to understand their purpose. And Add a doc section at beginning of each module describing their inputs, outputs, and briefly mention what they will do and how they will do ]
2. Explanation containing
	* 1. Example input and output of each iteration
		2. Discussion of input/output of each iteration

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| Input/ Output from the three runs Quality  | Input, outputs and their discussions are not written in the report | One out of three runs are done and/ or Input, outputs and their discussions are poorly written in the report and has many mistakes | Two out of three runs are done and/ or Input, outputs and their discussions are modestly written in the report and has some mistakes | All runs are done properly, Input, outputs and their discussions are very good | 3 |

**c Q**: Sort the obtained augmented RHOUSTONW Datasets using the OLS attribute. Discuss the top 3 examples of each augmented dataset; explain why you believe the particular examples were viewed as likely outlier. Also discuss the bottom example in each augmented dataset: try to explain why twere rated to be “most normal
Deliverable:

1. Code showing sorts using OLS attribute
2. A report containing
	* 1. The top 3 examples of each augmented dataset
		2. Discussion of why they viewed as likely outlier candidates
		3. The bottom 1 examples in the augmented dataset
		4. Discussion of why rated to be “most normal”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| Presentation of first 3 and bottom 1 samples  | No samples are presented | Presented samples from both sides are wrong | Presented samples from at least one side is wrong | Presented samples from both sides are correct | 3 |
| Discussion of the samples | No discussion given | Discussion is wrong with lots of erroneous claims  | Discussion is modest with some of erroneous claims | Discussion is very good | 4 |

 **Q**: Based on the results you obtained in the previous steps evaluate and compare the two outlier detection techniques you developed.

Deliverable:

 A report containing the discussion

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
|  |  |  |  |  |  |
| Comparison of the two outlier detection techniques | No discussion given | Discussion is wrong with lots of erroneous claims  | Discussion is modest with some of erroneous claims | Discussion is very good | 4 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Report Quality  | No report is given | The report is poorly written with lots of mistakes and contains many redundant comments and bad organization | The report quality is moderate with some mistakes and contains a few redundant comments and okay organization | The report is very well written with no redundancy and good organization | 2 |