"Data Mining"

Clustering and outlier detection

Learning Objectives:

1. Learn to use popular clustering algorithms, namely K-means, DBSCAN and detect outliers
2. Learn how to summarize and interpret clustering results
3. Learn to write analysis and evaluation functions which operate on the top of clustering algorithms and clustering results
4. Learning how to interpret unsupervised data mining results

Task-1

In the project you will use the Houston Weather Dataset, or HWD for short. The first and last attribute of the HWD should be ignored when clustering this data set; the last attributes denotes a class variable which will be used in the post analysis of the clusters generated by running K-means, and DBSCAN.

Houston_Weather Dataset has the the following attributes:

DATE / nominal / Each record has a date starting from 01/01/2006 to 12/31/2021. You may download this dataset from https://www.kaggle.com/datasets/alejandrochapa/houston-weather-data.
cloud_cover / categorical / %/ 0 to 16, each number represents a category
rainfall / continuous / inch / Amount of rainfall of the day
min_temp / continuous / farenhit / Minimum temperture of the day
max_temp / continuous / farenhit / Maximum temperture of the day
wind_speed/ continuous / mile per hour / wind speed at 3pm
pressure/ continuous / pai / atmospheric pressure at 3pm
humidity / continuous / % / relative humidity at 3pm
class/ / categorical / %/ H, M, L repregenting High, Midium and Low humidity

Examples in the Weather Prediction Dataset:

Date	min_temp	max_temp	rainfall	wind_speed	humidity	pressure	cloud	Class
1/1/2021	41	55	0	8	51	29.95	4	M
1/2/2021	41	59	0	7	42	30.09	3	L
1/3/2021	43	68	0	13	37	30.01	3	L
1/4/2021	49	75	0	3	43	29.99	0	L

a. Run K-means for k=3[1](check footnote) for the HWD dataset excluding the Date and Class attribute. Using the function you developed in step a, compute the purity of the obtained clustering results; next, create box plots for attributes temp_max, temp_min, rainfall, humidity,wind_speed, pressure of the obtained 3 clusters for each clustering and report their centroids, means. Finally, summarize based on the obtained boxplots and centroids/cluster means what kind of objects each of 3 clusters contains (you need to compare attributes in terms of their clusters). Finally, report the purity for the clustering result and interepret it. ***

b. Try to obtain a DBSCAN clustering for the HWD dataset exclusing the Date, and class attribute, having between 2 and 15 clusters with less than 20% outliers. Report its purity score. Compare the result with the K-means result you obtained in task 1! ***

Deliverables:

A. A Report[2] which contains all deliverables for the subtasks of Task 1.
B. Properly commented software/code you developed as part of Task 1.

Task-2

In this task you will be developing outlier detection techniques for a HWD Dataset as provided in Task-1; the objective is to find "*unusual weather days*" in this dataset.

[1] Actually run it 10 times but then analyze only the (single) clustering with the lowest SSE further.
[2] Single-spaced; please use an 11-point or 12-point font!

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, you will use a dataset called the HWD 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. **********

a) 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.

b) 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 were 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!

Deliverable:

a) 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]
b) Explanation containing

i. Algorithm/Psudocode that explain your detection mechanism
ii. Explanation how the algorithm works
iii. Example input and output and discussion of input/output

	Level 0	Level 1	Level 2	Level 3	Weig ht
Quality of the Distance function	No Distance function is presented	The Distance function is not very sophisticated/inc orrect and will produce wrong outputs in most cases	The Distance function is modestly sophisticated/inc orrect and will produce wrong outputs in some cases	The Distance function is very good	4
Distance-based	No distance-based	The distance-based outlier	The distance-based outlier	The distance-based	4

outlier detection technique Quality	outlier detection technique is presented	detection technique is not very sophisticated/incorrect and will produce wrong outputs in most cases	detection technique is modestly sophisticated/incorrect and will produce wrong outputs in some cases	outlier detection technique is very good	
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

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

    a) Example input and output of each iteration
    b) Discussion of input/output of each iteration

	Level 0	Level 1	Level 2	Level 3	Weight
Input/ Output	Input, outputs and their	One out of three runs are done and/ or	Two out of three runs are done and/ or	All runs are done properly,	3

from the three runs Quality	discussions are not written in the report	Input, outputs and their discussions are poorly written in the report and has many mistakes	Input, outputs and their discussions are modestly written in the report and has some mistakes	Input, outputs and their discussions are very good	

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:

3. Code showing sorts using OLS attribute
4. A report containing

   iv. The top 3 examples of each augmented dataset
   v. Discussion of why they viewed as likely outlier candidates
   vi. The bottom 1 examples in the augmented dataset
   vii. 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

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