Dr. Eick and Md Mahin

COSC 6335*“Data Mining”*

ProblemSet2 Fall 2023

Clustering and Outlier Detection

Last updated: November 2, 9a

**Learning Objectives**:

1. Learn to use popular clustering algorithms, namely K-means, DBSCAN
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. Learn how to design and implement outlier dection techniques
5. Learn how to interpret outlier detection results
6. Learn how to design density functions
7. Learn how to design distance functions
8. Learning how to interpret unsupervised data mining results

Task 3: Clustering

Group Task (groups of 3)

Not Peer Reviewed

First Draft



Fig. 1: Earthquake Clusters in Nepal

Deadline: November 30, 2023

In the project we will use the Earthquake[[1]](#footnote-1) dataset, EQ dataset for short and the Basel Weather Dataset, or BWD for short. The first and last attribute of the BWD 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.

The Earthquake Dataset available at:Teams/ H\_20233\_COSC\_6335\_18523/Files/Data/EarthQuack.csv

Earthquake Dataset has the the following attributes:

time / nominal / Each record has a date starting from 09/05/2022 to 10/05/2022

latitudes / continuous / -- / latitudes for continuous use in between 28.50 to 48.96

longitudes / continuous / %/ longitudes for continuous use in between -124.613 to -67.62

depth / continuous / km / Earthquake origination depth

mag / continuous / / Earthquake magnitude

3 Examples in the Weather Prediction Dataset:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 2022-10-05T23:25:08.960Z | 33.1785 | -116.411 | 12.6 | 0.63 |
| 2022-10-05T22:59:07.030Z | 33.931 | -116.358 | 6.74 | 2.22 |
| 2022-10-05T22:42:30.890Z | 38.79867 | -122.75 | 1.6 | 0.95 |

BWD available at:Teams/ H\_20233\_COSC\_6335\_18523/Files/Data/Basel\_Weather.csv

Basel\_Weather Dataset has the the following attributes:

DATE / nominal / Each record has a date starting from 01/01/2000 to 01/01/2010

MONTH / nominal / -- / 1 to 12

cloud\_cover / nominal / %/ 0 to 10, The fraction or percentage of the sky that is covered by clouds at a specific location and time

global\_rediation / continuous / W/m2/ The solar radiation that reaches the Earth's surface

precipitation / continuous / mm / Amount of rainfall

subshine / continuous / hours / The amount of time during which the sun is visible and unobscured by clouds or other atmospheric obstructions.

temp\_mean / continuous / celcicus / Average temperture of the day

temp\_min / continuous / celcicus / Minimum temperture of the day

temp\_max / continuous / celcicus / Maximum temperture of the day

humidity / continuous / % /

humidity\_class / ordinal/ ---/ Low, Mid, and High (derived from Humidity attribute; see below)

3 Examples in the Weather Prediction Dataset:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 20000101 | 1 | 7 | 0.82 | 1.34 | 0 | -15.2 | -17 | -13.4 | 0.89 | Mid |
| 20000102 | 1 | 5 | 0.6 | 0.39 | 2.8 | -13.7 | -15 | -12.3 | 0.86 | Low |
| 20000110 | 1 | 8 | 0.25 | 0.38 | 0 | -13.3 | -15 | -11.6 | 0.97 | High |

**Task 3 Subtasks:**

1. Write an function[[2]](#footnote-2) purity(a,b,outliers=FALSE) that computes the purity of a clustering result based on an apriori given set of class lables, where *a* gives the assignment of objects in O to clusters, and *b* is the “ground truth”. \* Purity is defined as follows: Let

O be a dataset

X={C1,…,Ck} be a clustering of O with Ci ⊆O (for i=1,…,k), C1∪…∪Ck ⊆O and Ci∩Cj=∅ (for i≠ j)

 1

If the used clustering algorithm supports outliers, outliers should be ignored in purity computations; if you use R-clustering algorithms, you can assume that cluster 0 contains all the outliers, and clusters 1,2,…,k represent “true” clusters. If the parameter outliers is set to FALSE, the function just returns a floting point number of the observed purity, if parameter outliers is set to T the function returns a vector: (<purity>,<percentage\_of\_outliers); e.g. if the function returns (0.98, 0.2) this would indicate that the purity is 98%, but 20% of the objects in dataset O have been classified as outliers.\*

1. Write an function asse(a,b) that computes the average sum of square error of a clustering result based on the apriori given values of a numical attribute, where *a* gives the assignment of objects in O to clusters, and *b* gives the value of the numerical attribute”. \* The average sum of the square error is defined as follows: Let

O be a dataset

X={C1,…,Ck} be a clustering of O with Ci ⊆O (for i=1,…,k), C1∪…∪Ck ⊆O and Ci∩Cj=∅ (for i≠ j)

 (2)

In equation 2, he sum is calculated for all objects in O which belong to clusters, excluding outliers. If the used clustering algorithm supports outliers, outliers should be ignored in asse computations; if you use R-clustering algorithms, you can assume that cluster 0 contains all the outliers, and clusters 1,2,…,k represent “true” clusters. \*

1. Let O be a dataset

X={C1,…,Ck} be a clustering of O with Ci ⊆O (for i=1,…,k), C1∪…∪Ck ⊆O and Ci∩Cj=∅ (for i≠ j).

Here,

Where:

* n is the number of dimensions (attributes) in the data.
* Ci​,Cj is the i-th and j-th cluster in the clustering X
* ∣Ci​∣ is the number of data points in the i-th cluster
* d(xi​,xj​) is the distance between the data points xi​ and xj​

You can compute the silhouette score when requested in the subsequent tasks using existing libraries.

1. **Develop a visualization procedure to visualize the clusters in the EQ clusters on the map. A better visualization procedure will plot cluster boundaries. \*\*\*\***
2. Run K-means for k=3 to k=20 for the EQ dataset, using the (longitude,lattitude) attributes Additionally compute the asse value for each obtained clustering with respect to the magnitude attribute and with respect to the depth attribute for each K. Interpret the results you obtained. Visualize the obtaining two clusterings using the visualization techniques you developed in Task d. \*\*\*
3. Run K-means for k=3[[3]](#footnote-3) for the BWD dataset excluding the Date, Month and Class Attributes. Using the function you developed in step a, compute the purity of the obtained clustering results; next, create box plots for attributes 2 to 9 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. Finally, report the silhouette score and purity for the clustering result and interepret it. \*\*\*
4. Try to obtain a DBSCAN clustering for the BWD dataset exclusing the Date, Month and class attribute, having between 2 and 15 clusters with less than 20% outliers. Report its purity and Silhouette score. Compare the result with the K-means result you obtained in task f! \*\*\*
5. Try to find a “good DBSCAN clustering” using 15 or less clusters and less than 20% outliers for the EQ dataset using the earthquake longitude and lattitude minimizing the asse score for the earthquake depth attribute. Devise a search procedure for finding MINPOINTS and epsilon parameter values which minimize the asse measure for the earthquake depth attribute. You have to develop a method that find optimal MINPOINTS and epsilon parameters which give the best asse score for the earthquake depth attribute. **Visualize the best clustering you obtained and report its asse score and also the average asse value for each cluster.** Compare this clustering result with the K-means clustering you obtained in task e. Also describe the search procedure you appled to find your “best” DBSCAN clustering was obtained. \*\*\*\*
6. The clustering algorithms you applied in steps d and h to the EQ dataset only obtains spatial clusters based on earthquake locations, but ignores other earthquake attributes, such as magintude and depth of the earthquake. Try to develop a clustering procedure which finds spatial clusters with similar earthquake depth; e.g. the asse value of the obtained clustering with respect to the depth attribute should be as low as possible. You have to develop a similar method like the one in task h and find find MINPOINTS and epsilon gives lowest **asse value.** Describe how your developed procedure works and report its best result and include a visusalization of this result. \*\*\*\*\*

**Deliverables for Task 3:**

1. A Report[[4]](#footnote-4) which contains all deliverables for the 9 subtasks of Task 3.
2. Properly commented softwar/code you developed as part of Task 3.
3. Readme to write the code

Task 4: Outlier Detection

Individual Task

Peer Reviewed

Second Draft



Fig. 2: Some Unusual Weather

Deadline: **November 6**, 2023

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

However, in this task we restrict ourselves to a subset of the dataset, called RBASELW, focusing on the following attributes: Date, global\_radiation, precipitation, sunshine, temp\_mean, humidity.

Subtasks:

1. Develop a single attribute outlier detection technique! \*\*\*
2. Apply the technique you developed to the global\_radiation and temp\_mean attribute each; report and interpret your obtained results. \*\*
3. Design and implement a distance-based and a model/density-based object outlier detection technique for the Basel Weather Dataset. The technique if applied to the Basel 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 RBASELW dataset; the challenge for the second task will be to develop a “good” density function for the RBASELW dataset. \*\*\*\*\*\*\*\*\*
4. Apply the two outlier detection techniques to the RBASWLW Dataset; if your methods involves hyper parameters, apply the methods 3 times to the dataset using 3 different hyper parameter settings. \*\*\*\*
5. Sort the obtained augmented RBASELW Datasets using the OLS attribute. Discuss the top 4 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 these examples were rated to be “most normal”.\*\*\*\*
6. Based on the results you obtained in e evaluate and compare the two outlier detection techniques you developed. \*\*
7. If necessary, enhance your two outlier detection techniques and redo steps d, e, and f!

 **Deliverables for Task 4:**

1. Indivdual task solutions will be submitted via Kritik.
2. Submit a separte file which contains the software/code you developed as part of Task

Rubrics:

1. **Q**: Develop a single attribute outlier detection technique!
Deliverable:
	1. A properly commented code file. [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. A readme file how to run the code
	3. A report 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 |
| 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 |
| Outlier Detection Method Correctness | No outlier detection method is presented | The outlier detection method is fundamentally wrong and producing erroneous outputs | The outlier detection method is okay but can be improved | The outlier detection method is fundamentally correct | 5 |
| Outlier Detection Method Quality | No outlier detection method is presented | The outlier detection method is not very sophisticated and will produce wrong outputs in most cases | The outlier detection method is modestly sophisticated and will produce wrong outputs in some cases | The outlier detection method is very good | 4 |

**b Q**: Apply the technique you developed to the global\_radiation and temp\_mean attribute each; report and interpret your obtained results.
Deliverable:

* 1. A properly commented code file. [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. A readme file how to run the code
	3. A report containing
		1. Example input and output
		2. Discussion of input/output

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| 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 |
| Input/ Output Quality  | Input, outputs and their 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 | 3 |

**c Q**: Design and implement a distance-based and a model/density-based object outlier detection technique for the Basel Weather Dataset. The technique if applied to the Basel 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 RBASELW dataset; the challenge for the second task will be to develop a “good” density function for the RBASELW dataset.
Deliverable:

1. **Q**: Develop a single attribute outlier detection technique!
Deliverable:
	1. A properly commented code file. [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. A readme file how to run the code
	3. A report 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 |
| Correctness of the Distance function  | No Distance function is given | The Distance function is fundamentally wrong and producing erroneous outputs | The Distance function is okay but can be improved | The Distance function is fundamentally correct | 5 |
| Quality of the Distance function  | No Distance function is presented | The Distance function is not very sophisticated and will produce wrong outputs in most cases | The Distance function is modestly sophisticated and will produce wrong outputs in some cases | The Distance function is very good | 4 |
| Distance-based outlier detection technique Correctness | No distance-based outlier detection method is presented | The distance-based outlier detection technique is fundamentally wrong and producing erroneous outputs | The distance-based outlier detection technique is okay but can be improved | The distance-based outlier detection technique is fundamentally correct | 5 |
| Distance-based outlier detection technique Quality | No distance-based outlier detection technique is presented | The distance-based outlier detection technique is not very sophisticated and will produce wrong outputs in most cases | The distance-based outlier detection technique is modestly sophisticated and will produce wrong outputs in some cases | The distance-based outlier detection technique is very good | 4 |
| Correctness of the Density function  | No Density function is given | The Density function is fundamentally wrong and producing erroneous outputs | The Density function is okay but can be improved | The Density function is fundamentally correct | 5 |
| Quality of the Density function  | No Density function is presented | The Density function is not very sophisticated and will produce wrong outputs in most cases | The Density function is modestly sophisticated and will produce wrong outputs in some cases | The Density function is very good | 4 |
| Model/density -based outlier detection technique Correctness | No Model/density -based outlier detection method is presented | The Model/density -based outlier detection technique is fundamentally wrong and producing erroneous outputs | The Model/density -based outlier detection technique is okay but can be improved | The Model/density -based outlier detection technique is fundamentally correct | 5 |
| 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 and will produce wrong outputs in most cases | The Model/density -based outlier detection technique is modestly sophisticated and will produce wrong outputs in some cases | The Model/density -based outlier detection technique 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 | 4 |

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

1. A properly commented code file. [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. A readme file how to run the code
3. A report 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 |
| 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 |
| Input/ Output Quality  | Input, outputs and their 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 | 3 |

**e Q**: Sort the obtained augmented RBASELW Datasets using the OLS attribute. Discuss the top 6 examples of each augmented dataset; explain why you believe the particular examples were viewed as likely outlier candidates. Also discuss the bottom 2 examples in the augmented dataset; try to explain why these two examples were rated to be “most normal”.\*\*\*\*

Deliverable:

1. A code file showing sorts using OLS attribute
2. A report containing
	* 1. The top 6 examples of each augmented dataset
		2. Discussion of why they viewed as likely outlier candidates
		3. The bottom 2 examples in the augmented dataset
		4. Discussion of why rated to be “most normal”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| 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 |
| Presentation of first 4 and bottom 4 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 first 4 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 |
| Discussion of bottom 4 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 |

**f Q**: If necessary, enhance your two outlier detection techniques and redo steps d, e, and f!

Deliverable:

 A report containing the discussion

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Level 0 | Level 1 | Level 2 | Level 3 | Weight |
| 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 |
| 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 |

1. The earthquake dataset has the following attributes: (time, longitude, latitude, magnitude, depth) [↑](#footnote-ref-1)
2. 2 This function could be an R-function, a Python function or any other function. You might find some implementation of this function online; it is okay to use those implementations, as long as you acknowledge in your report what you use, and not all software you find on the internet is running properly. [↑](#footnote-ref-2)
3. Actually run it 10 times but then analyze only the (single) clustering with the lowest SSE further. [↑](#footnote-ref-3)
4. Single-spaced; please use an 11-point or 12-point font! [↑](#footnote-ref-4)