Dr. Eick and Md Mahin

COSC 6335*“Data Mining”*

Problem Set2

Task 3 Fall 2024

Clustering

Not Peer Reviewed Individual Task

Last updated: September 27 at 10.30a

Deadline: Friday, November 1 end of the day in MS Teams

**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. Learning how to interpret unsupervised data mining results



Fig. 1: Earthquake Clusters in Nepal

In the project we will use the Earthquake[[1]](#footnote-1) dataset, EQ dataset for short and 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.

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 Earthquake 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 |

Houston\_Weather Dataset has the the following attributes:

DATE / nominal / Each record has a date starting from 01/01/2006 to 12/31/2021

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

3 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

**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. **Develop a visualization procedure to visualize EQ-dataset clusters on the map. A better visualization procedure will plot cluster boundaries. \*\*\*\***
2. Run K-means for k=5, 9, 13, 17 for the EQ dataset, using the (longitude,lattitude) attributes Additionally compute the asse value for each obtained clustering above with respect to the magnitude attribute and with respect to the depth attribute for each K (do not create new clusters for the given attributes, use obtained cluster using (longitude,lattitude) pair). Interpret the results you obtained while comparing asse values for different k. Visualize two clusterings one for k with best depth and one with best magnitude using the visualization techniques you developed in Task d. \*\*\*
3. Run K-means for k=3[[3]](#footnote-3)(check footnote 3) 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. \*\*\*
4. 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 e! \*\*\*
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. 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. \*\*\*\*\*\*

**Deliverables for Task 3:**

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