Dr. Eick

COSC 6335*“Data Mining”*

ProblemSet2 Fall 2022

Deadlines; November 8, 2022

Last updated: October 6 at noon

Task 3: Clustering with K-Means, and DBSCAN

Group Task (groups of 2 or 3)

First Draft

**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

**Datasets**: Links: [UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29)

Download: Both datasets are available in Teams/Files/Datasets in Team DM2022.

In the project we will use the Earthquake[[1]](#footnote-1) dataset, EQ dataset for short, and a subset[[2]](#footnote-2) of the Wisconsin-breast-cancer dataset, WBC for short. The last attribute of the WBC dataset denotes a class variable which should be ignored when clustering this data set—however, the class variable will be used in the post analysis of the clusters generated by running K-means, and DBSCAN.

**Task 3 Subtasks:**

1. Write an function[[3]](#footnote-3) 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)

purity(X)= (number\_of\_majority\_class\_examples(X)/(total\_number\_examples\_in\_clusters(X))

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)

asse(X)= Σ(sqared difference of the object’s numerical attribute value to the cluster-mean of the numerical attribute for the cluster the object belong to)/(total\_number\_examples\_in\_clusters(X))

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. Run K-means for k=8 and k=12 twice for the EQ dataset, using the longitude and lattitude attributes. Also compute the asse value for each obtained clustering with respect to the magnitude attribute and with respect to the depth attribute. Compare the obtained four clusterings! \*\*\*
2. Run K-means for k=3[[4]](#footnote-4) for the WBC dataset. Using the function you developed in step a, compute the purity of the obtained clustering results; next, create box plots for attributes 2 to 10 of the obtained 3 clusters for each clustering and report their centroids, means. Compare the two clusterting results. Finally, summarize based on the obtained boxplots and centroids/cluster means what kind of objects each of 3 clusters contains. \*\*\*\*\*
3. Try to obtain a DBSCAN clustering for the WBC dataset, having between 2 and 15 clusters with less than 20% outliers. Report its purity! \*\* Compare the result with the K-means result you obtained in task d!
4. Develop a visualization procedure which puts the spatial clusters you obtained for the EQ dataset on an US map; visualize the 4 clusterings you obtained for task c! \*\*\*\*
5. Try to find a “good clustering” for the EQ dataset using the DBSCAN algorithm just based on earthquake longitude and lattitude by selecting proper values for MINPOINTS and epsilon. Use feedback from the visualization procedure you developed in task to obtain a “good” clustering. Report the best clustering you obtained, its asse value with respect to the magnitude and severity attributes, and compare it with the 4 K-means clustering you obtained in task c. Also describe how your “best” DBSCAN clustering was obtained. \*\*\*\*
6. The clustering algorithms you applied in steps c and g 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. 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[[5]](#footnote-5) which contains all deliverables for the 8 subtasks of Task 3.
2. An Appendix which contains the 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. It is the original dataset with 16 examples removed from the original dataset which contain missing values; that is, WBC is a numerical dataset which does not contain any missing values. [↑](#footnote-ref-2)
3. 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-3)
4. Actually run it 10 times but then analyze only the (single) clustering with the lowest SSE further. [↑](#footnote-ref-4)
5. Single-spaced; please use an 11-point or 12-point font! [↑](#footnote-ref-5)