Course ID: CSE-506
Course Name: Data Mining (DMG)

Assignment 3

October 20, 2023

Instructions

1. The assignment is to be attempted in groups.
2. Programming Language: Python
3. For Plagiarism, institute policy will be followed
4. You need to submit the report.pdf, readme.pdf, Code files and Images.
5. Deliverables should be submitted in Google Classroom in a zip folder with the name 'A3_Rollnumbers separated by underscore.zip'.
6. You **can use any library** for pre-processing, training, doing experiments and post-processing in all questions.
7. One member should submit on Google Classroom while other members can mark turn in without the attachment.
8. In case of doubts, please comment on the Google Classroom.
9. **You should be well aware of theory of the algorithms mentioned and chosen by you. You will end up losing sufficient no. of marks in case you haven't prepared well**.

Total Marks: 200 (will be scaled later on)

1. Choose any 3 **real world datasets** of your choice and provide the corresponding hyperlinks in the report but the selection should be done in such a manner that:
   (a) **(5 points)** Density based clustering is most suitable for the 1st dataset.
   (b) **(5 points)** Hierarchical based clustering is most suitable for the 2nd dataset.
   (c) **(5 points)** Prototype based clustering is most suitable for the 3rd dataset.
   (d) Marks will be awarded based on both theoretical explanations and data visualization. You may use t-SNE plots. This section encourages you to analyze and visualize the data first rather than directly jumping onto the algorithms. You may want to have a look at this.
2. **Density based clustering** as the name suggests works by identifying areas where data points are concentrated and where they are separated by areas that are sparse. In this section we will learn and implement one such technique.
   (a) DBSCAN is probably the most popular density based clustering algorithm and is also a recipient of KDD Test of Time Award.
   (b) **(2+2 points)** Mention two advantages and disadvantages of DBSCAN.
   (c) **(2 points)** Choose any one density based clustering algorithm other than DBSCAN and state its 2 advantages over DBSCAN.
   (d) **(10+10+10 points)** Implement the chosen algorithm using any library on all the 3 datasets and show the best results obtained respectively.
   (e) **(10 points)** Play with the input parameters or some other knobs that your chosen algorithm offers. Visualize and report the clustering performance comparison on 1st dataset only. So essentially we are performing relative cluster validation.
   (f) **(5 points)** Confirm your selection hypothesis in section 1(a) with the results obtained. In case of any discrepancy, draw inference.
   (g) **(5+5+5 points)** It is a wise practice to check whether the results obtained after clustering are valid, i.e., they are significantly different from random data. You need to come up with a suitable way to confirm it for all the 3 datasets.
3. Given n points in a d-dimensional space, the goal of **hierarchical clustering** is to create a sequence of nested partitions, which can be conveniently visualized via a dendogram. Agglomerative clustering

is a hierarchical clustering technique where we begin with each of the n data points in a separate cluster and repeatedly merge the two closest clusters until all points are members of the same cluster accompanied with a stopping criteria.

(a) **(1+1+1 points)** Agglomerative clustering belongs to one of the two main categories of hierarchical clustering. Name the categories and provide the reason for wide usage of one category over the other.

(b) The main step in agglomerative clustering is to determine the closest pair of clusters. Computing cluster distances ultimately translates to computing distance between two points which is generally taken to be Euclidean distance. But there are different measures that are generally employed for representing the cluster distance, known as linkage.

(c) **(10+10+10 points)** Run agglomerative clustering using the following linkages {single, complete, group average, minimum variance}. Compare the clustering performance both visually and empirically on the 3 datasets. Report the best results.

(d) **(5 points)** Confirm your selection hypothesis in section 1(b) with the results obtained. In case of any discrepancy, draw inference.

(e) **(5+5+5 points)** It is a wise practice to check whether the results obtained after clustering are valid, i.e., they are significantly different from random data. You need to come up with a suitable way to confirm it for all the 3 datasets.

4. In **prototype-based clustering**, also known as representative-based clustering, a cluster is a group of objects in which some object is nearer to the prototype that represents the cluster than to the prototype of some other cluster.

(a) **(1+1 points)** k-means clustering is such a profound and simple prototype-based clustering algorithm that needs the centroid of the elements in a cluster as the prototype of the cluster. But it has some limitations. Two such limitations are (i) clustering outliers (ii) scaling with number of dimensions. Explain these two limitations.

(b) **(2+2 points)** Suggest two existing algorithms that use some technique to mitigate these limitations respectively. *i.e.* 1 algorithm for 1 limitation. Marks will not be awarded without explaining the techniques incorporated.

(c) **(10+10+10 points)** Run these two algorithms on the 3 datasets and report the best results obtained. Compare the results with that of k-means and confirm if they are performing better than k-means.

(d) **(4+4+4 points)** Report and visualize the hyperparameter tuning for the two algorithms required to achieve the best results obtained on the 3 datasets.

(e) **(3 points)** Confirm your selection hypothesis in section 1(c) with the results obtained. In case of any discrepancy, draw inference.

(f) **(5+5+5 points)** It is a wise practice to check whether the results obtained after clustering are valid, i.e., they are significantly different from random data. You need to come up with a suitable way to confirm it for all the 3 datasets.

5. **FYI** – For reporting the empirical results for clustering, two distinct type of measures are employed (i) Intrinsic (ii) Extrinsic. Commonly used intrinsic measures are {Sum of squared errors, Silhouette Coefficient, Davies Bouldin index, Dunn index} whereas for extrinsic measures, {Purity, Normalized Mutual Information, Rand Index, F1-score} are the commonly used ones. You need to choose any 2 extrinsic measures mentioned above. Choose any 2 intrinsic measures from above or based on what your chosen algorithm offers.