CIS6930 Fall 2017: Introduction to Data Mining
Project II: Clustering

October 30, 2017

Project Description

This project aims to make you familiar with the clustering techniques, available in R to do some easy data mining analysis on two given datasets. You need to apply several different clustering methods and submit:

- A detailed report showing:
  - Breif description of clustering techniques that you use
  - Detailed analysis of the required tasks for two datasets
  - Your conclusion and reference list
- A *Readme.txt* file, explaining how to run your script.
- Three .R script, for each section. By running this script, we should be able to get your reported results.

Submit your files as a .zip file to Canvas, with the name format as: **Firstname_Lastname_UFID.zip**.

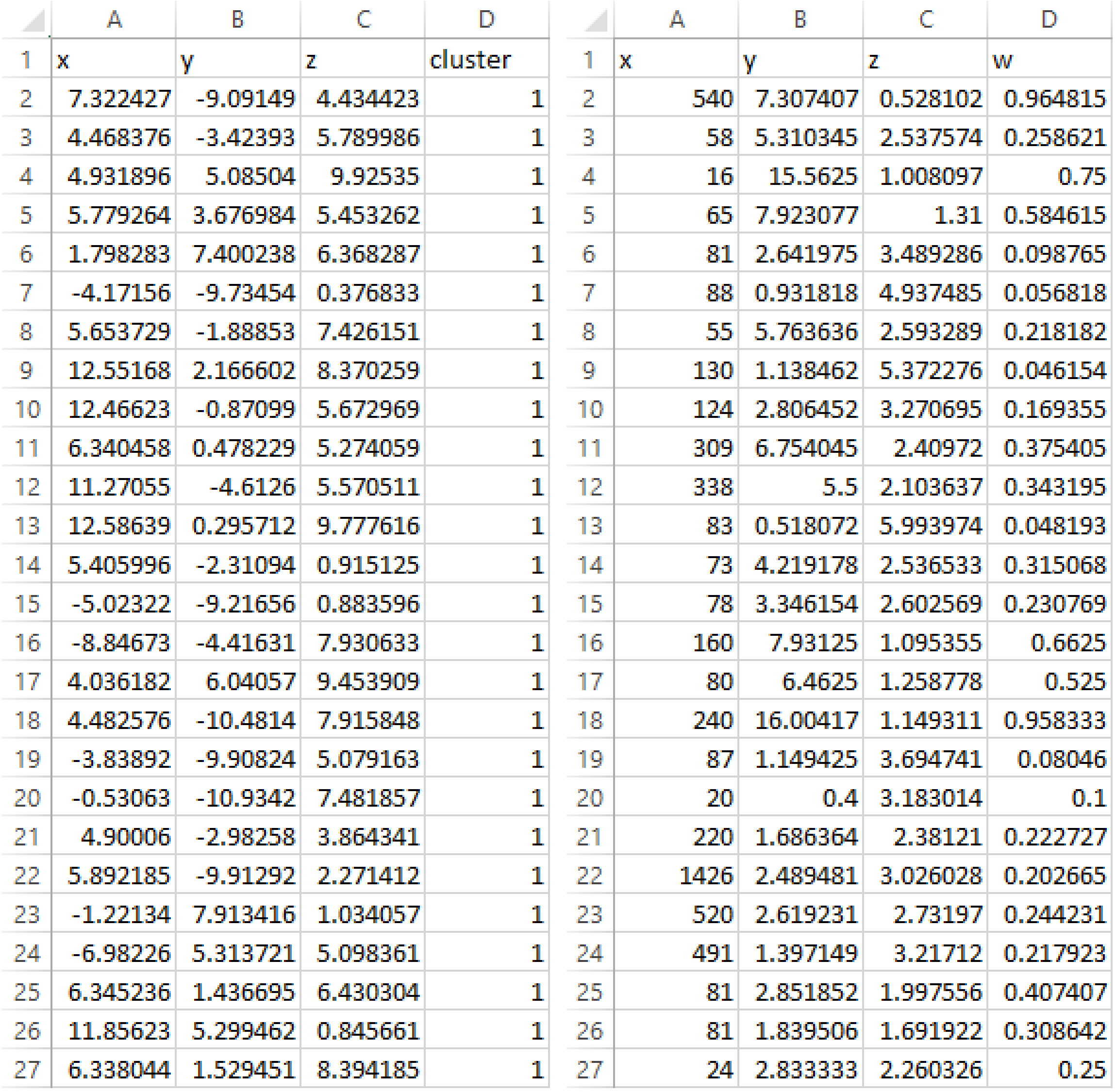
Dataset 1

The first dataset, named as *dataset1.csv*, contains 1000 data points of 8 clusters in 3D space (see Figure 1). The first 3 columns represent the corresponding coordinates in 3 dimensions and the 4th column gives the ground-truth cluster labels of each point. You need to apply the following clustering techniques that you have learned from the class, and plot the clustering results in 3D graphic figures (not necessarily required in R):

- Hierarchical Clustering
- K-means Clustering
- Density-based Clustering
- Graph-based Clustering

Your final goal is to get 8 different clusters of the data points, so you need to feed the appropriate parameters to each clustering function (that you call in related R packages) and describe how these parameters can achieve 8 clusters. Since the ground-truth labels are available in the dataset, one of the measurement method is to find the best-fit mapping between the new 8 kinds of clustering label to the 8 kinds of ground-truth label. Compute the accuracies based on this mapping strategy and compare the performance of the above clustering techniques.

Note that if a clustering technique involves with randomly selecting the the initial centroids, you need to repeat the clustering experiments for multiple times and compute the average value of the accuracies.



	A	B	C	D
1	x	y	z	cluster
2	7.322427	-9.09149	4.434423	1
3	4.468376	-3.42393	5.789986	1
4	4.931896	5.08504	9.92535	1
5	5.779264	3.676984	5.453262	1
6	1.798283	7.400238	6.368287	1
7	-4.17156	-9.73454	0.376833	1
8	5.653729	-1.88853	7.426151	1
9	12.55168	2.166602	8.370259	1
10	12.46623	-0.87099	5.672969	1
11	6.340458	0.478229	5.274059	1
12	11.27055	-4.6126	5.570511	1
13	12.58639	0.295712	9.777616	1
14	5.405996	-2.31094	0.915125	1
15	-5.02322	-9.21656	0.883596	1
16	-8.84673	-4.41631	7.930633	1
17	4.036182	6.04057	9.453909	1
18	4.482576	-10.4814	7.915848	1
19	-3.83892	-9.90824	5.079163	1
20	-0.53063	-10.9342	7.481857	1
21	4.90006	-2.98258	3.864341	1
22	5.892185	-9.91292	2.271412	1
23	-1.22134	7.913416	1.034057	1
24	-6.98226	5.313721	5.098361	1
25	6.345236	1.436695	6.430304	1
26	11.85623	5.299462	0.845661	1
27	6.338044	1.529451	8.394185	1

	A	B	C	D
1	x	y	z	w
2	540	7.307407	0.528102	0.964815
3	58	5.310345	2.537574	0.258621
4	16	15.5625	1.008097	0.75
5	65	7.923077	1.31	0.584615
6	81	2.641975	3.489286	0.098765
7	88	0.931818	4.937485	0.056818
8	55	5.763636	2.593289	0.218182
9	130	1.138462	5.372276	0.046154
10	124	2.806452	3.270695	0.169355
11	309	6.754045	2.40972	0.375405
12	338	5.5	2.103637	0.343195
13	83	0.518072	5.993974	0.048193
14	73	4.219178	2.536533	0.315068
15	78	3.346154	2.602569	0.230769
16	160	7.93125	1.095355	0.6625
17	80	6.4625	1.258778	0.525
18	240	16.00417	1.149311	0.958333
19	87	1.149425	3.694741	0.08046
20	20	0.4	3.183014	0.1
21	220	1.686364	2.38121	0.222727
22	1426	2.489481	3.026028	0.202665
23	520	2.619231	2.73197	0.244231
24	491	1.397149	3.21712	0.217923
25	81	2.851852	1.997556	0.407407
26	81	1.839506	1.691922	0.308642
27	24	2.833333	2.260326	0.25

Figure 1: The partial screenshots of *dataset1.csv* (left) and *dataset2.csv* (right).

Dataset 2

The second dataset, named as *dataset2.csv*, contains more than 1 million data points with the coordinates in 4 dimensions (see Figure 1). You need to choose one of the above 4 clustering techniques and apply it on this dataset, by considering the following factors:

- How many clusters you think is the best to seek?
- Which methods can be applied for this case, and which one you think works best?

Note that, this dataset may be too large to apply some clustering methods and check the results. The decision you make for the clustering method is an important aspect of this part. After your clustering is done, evaluate the clustering method by the measures of similarity matrix, the sum of internal root-mean-square-deviation (RMSD), *etc.*