Homework 3

Matthew Chau
CS470
Emory University

February 28, 2020

Collaboration Statement

For this assignment, I have not consulted or asked for help from anyone for the completion of this assignment.

Dataset Description

For this homework assignment, I have used two data sets.

The first is '*iris.data*', containing 3 classes of 50 instances each, where each class refers to a type of iris plant. This data can be found on *http://archive.ics.uci.edu/ml/datasets/Iris*. According to the website, its attribute description is as follows:

1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class: —- Iris Setosa, Iris Versicolour, Iris Virginica

The second data set is *haberman.data*, containing cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer. This data can be found on *http://archive.ics.uci.edu/ml/datasets/Haberman%27s+Survival*. According to the website, its attribute description is as follows:

1. Age of patient at time of operation (numerical)
2. Patient's year of operation (year - 1900, numerical)
3. Number of positive auxiliary nodes detected (numerical)
4. Survival status (class attribute)

—- 1 = the patient survived 5 years or longer
—- 2 = the patient died within 5 year

Preprocessing

In order to make the data easier for numerical calculation, I have removed the nominal attribute on the last column for *iris.data*. In general, if the last column is not numerical type, it will be removed from the data set.
Initialization involves creating a data -¿ cluster map, where each key represents the index of each row of data in the data set, and each value represents the cluster index it belongs to. Also, a cluster -¿ data map will be created, where each key represents the index of each cluster, and the value represents a set of data indices that that belong to this cluster.

Results

Here are the results for running *iris.data*, with k = 3, 5, and 7. Each experiment is repeated 3 times.

when k = 3:

Trial	1	2	3
Time (s)	4.64	4.82	5.21
k	3		
size of cluster1	50	39	61
size of cluster2	39	61	39
size of cluster3	61	50	50
SSE	78.945	78.945	78.945
Silhouette Coeff.	0.551	0.551	0.551

when k = 5:

Trial	1	2	3
Time (s)	5.21	6.14	6.27
k	5		
size of cluster1	27	32	23
size of cluster2	23	27	28
size of cluster3	41	40	32
size of cluster4	27	28	40
size of cluster5	32	23	27
SSE	49.741	49.713	49.713
Silhouette Coeff.	0.376	0.376	0.376

when k = 7:

Trial	1	2	3
Time (s)	5.40	5.65	5.48
k	7		
size of cluster1	20	10	23
size of cluster2	12	47	24
size of cluster3	23	11	33
size of cluster4	39	30	27
size of cluster5	25	19	12
size of cluster6	24	23	27
size of cluster7	7	10	4
SSE	36.816	46.977	37.555
Silhouette Coeff.	0.350	0.324	0.337

Below are the results for running *haberman.txt* with the same configuration:
when k = 3:

Trial	1	2	3
Time (s)	23.0	21.4	21.4
k	3		
size of cluster1	123	124	124
size of cluster2	79	29	31
size of cluster3	104	153	151
SSE	23967.5	21167.2	21168.9
Silhouette Coeff.	0.324	0.432	0.431

when k = 5:

Trial	1	2	3
Time (s)	21.9	22.0	20.9
k	5		
size of cluster1	29	67	92
size of cluster2	89	74	73
size of cluster3	67	85	44
size of cluster4	70	29	28
size of cluster5	51	51	69
SSE	13790.8	13587.1	13779.5
Silhouette Coeff.	0.371	0.299	0.297

when k = 7:

Trial	1	2	3
Time (s)	24.6	23.3	23.1
k	7		
size of cluster1	22	41	4
size of cluster2	45	61	92
size of cluster3	66	53	40
size of cluster4	13	27	68
size of cluster5	26	50	12
size of cluster6	58	45	24
size of cluster7	76	29	66
SSE	10972.4	11759.9	10546.8
Silhouette Coeff.	0.272	0.248	0.314

Observations and Conclusion

From the data presented above we can have the following observations:

1. The larger k is, the smaller the clusters would be, and SSE would be smaller (since the size of the clusters are smaller and Silhouette Coeff. would be smaller.

2. Each trial the clusters might have different indexings (due to randomness while choosing first set of centroids), and each trial might have slight variations for the cluster sizes.

3. The larger k is, the more time it takes, but change is very subtle. However, if the data set is bigger, it takes a lot more time to run the k-means algorithm, and the time it takes is much sensitive to the size of the dataset.

From this experience, I have also reinforced my skillset with pandas and numpy as I have used lots of them in my code. Moreover, I have gained much experience on dealing with data structures if I draw them out explicitly, so that I have an overview of what and how I'm storing.
Also, while dealing with k-means algorithm, we have to test it multiple times with different k to find the best clustering (with the smallest SSE and greatest Silhouette Coefficient, because without k given, sometimes we do not know exactly how much clusters the dataset can be clustered into.