CLUSTERING ALGORITHMS(Problem 1)

Nitin K (2017csb1093)

30.10.2020
CSE FINAL YEAR, UG
IIT ROPAR

INTRODUCTION

This a report for two clustering algorithms(K-Means, DBScan). We are going to analyse how these algorithms perform on two different dataset(**Iris**, **Spiral**).

Algorithms Details:

I have Taken All the Algorithms from the books. That is why I am providing the reference only as the algorithm has already been discussed in the class.

K-Means Clustering

Reference book for the Algorithm is **Data Mining and Analysis: Fundamental Concepts and Algorithms, Mohammed J. Zaki and Wagner Meira Jr.**

DBScan

Reference book for the Algorithm is **Data Mining and Analysis: Fundamental Concepts and Algorithms, Mohammed J. Zaki and Wagner Meira Jr.**

DATA

DataSet Descriptions

Here We have Provided two Different DataSet:

- **Iris DataSet (https://archive.ics.uci.edu/ml/datasets/iris)**

	sepal_length	sepal_width	petal_length	petal_width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

- **Spiral DataSet(https://github.com/milaan9/Clustering-Datasets)**

	x1	x2	label
count	312.000000	312.000000	312.000000
mean	18.408173	16.344712	2.016026
std	7.299923	6.867232	0.815682
min	3.000000	2.900000	1.000000
25%	12.912500	11.337500	1.000000
50%	18.325000	16.050000	2.000000
75%	23.400000	21.362500	3.000000
max	31.950000	31.650000	3.000000

RESULTS

2. K-Means

DataSet 1: Iris

Table 1.1

Clusters(K)	SSE	Misclassification	RunTime
1	1834.027083	100	0.03686404228
2	959.1855172	71	0.09820246696
3	88.63365581	17	0.1986601353
4	74.50320241	17	0.4794006348
5	60.70936552	21	0.2423560619
6	47.19391621	6	0.438573122
7	61.5951715	14	0.4988059998
8	48.05052359	18	0.4835071564
9	84.27642952	19	0.2536296844

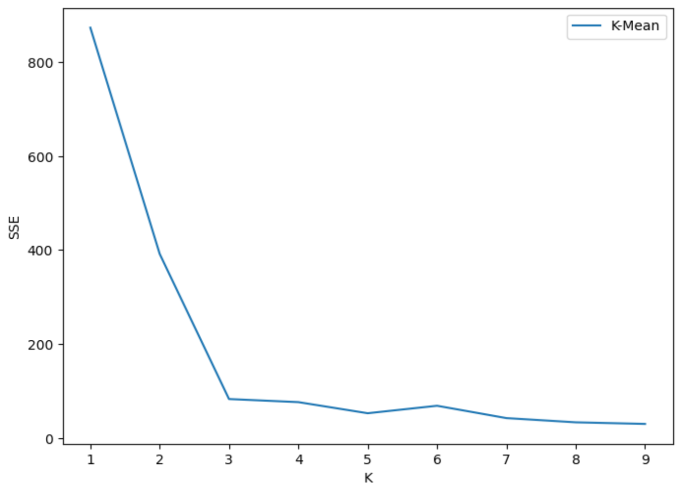


Fig:1.1

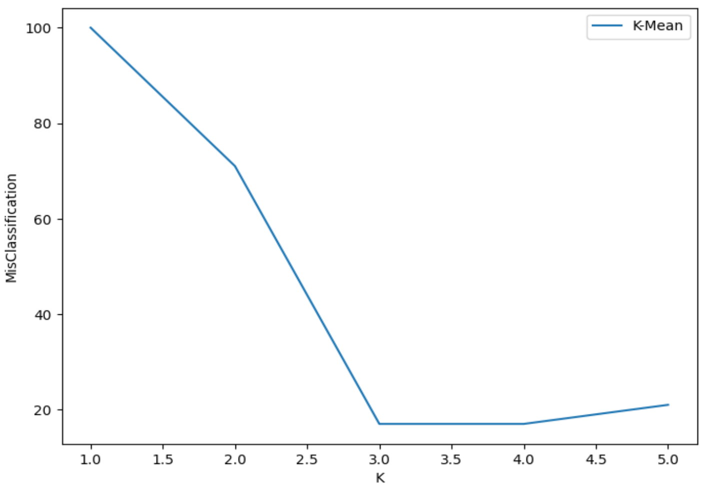


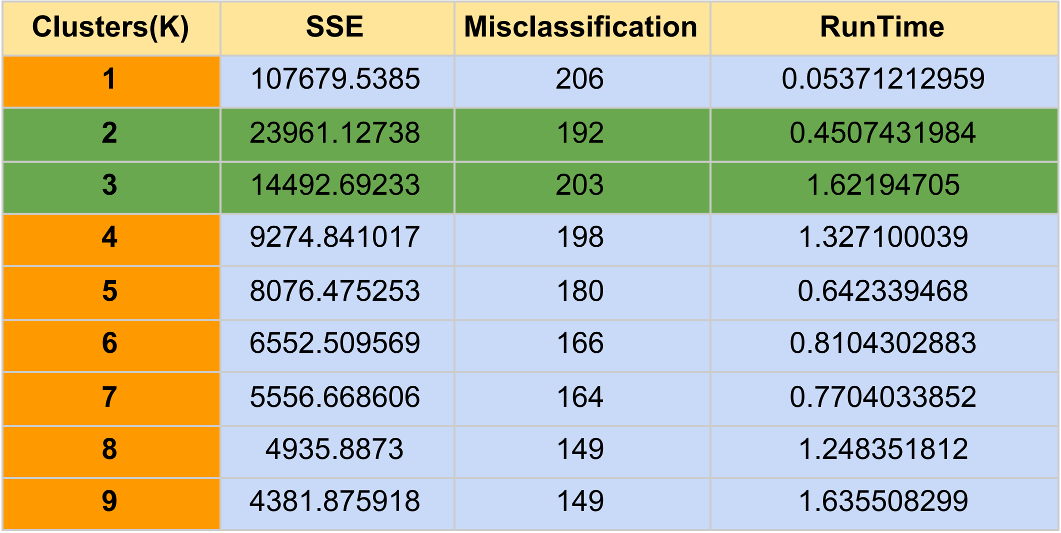
Fig: 1.2

From **Table 1.1** and **Fig:1.1** We can see how **K-Means** behave on the given **Iris DataSet**. SSE value decreases rapidly with increase in K to some extent but after some value of K it decreases slowly. Here till **K=3,** SSE decreases rapidly. After this It decreases slowly . Similarly we can see from **Fig: 1.2** how misclassifications depend on K.
Miss-Classifications for K=3 is 17 out 150 data points.

So From above observations, we can take **K=3** as the best representation of the dataset.

DataSet 2: Spiral

Table 2.1



Clusters(K)	SSE	Misclassification	RunTime
1	107679.5385	206	0.05371212959
2	23961.12738	192	0.4507431984
3	14492.69233	203	1.62194705
4	9274.841017	198	1.327100039
5	8076.475253	180	0.642339468
6	6552.509569	166	0.8104302883
7	5556.668606	164	0.7704033852
8	4935.8873	149	1.248351812
9	4381.875918	149	1.635508299

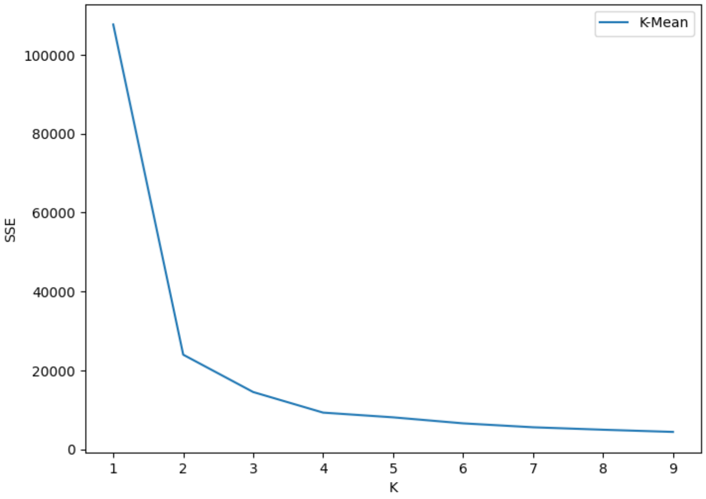


Fig:2.1

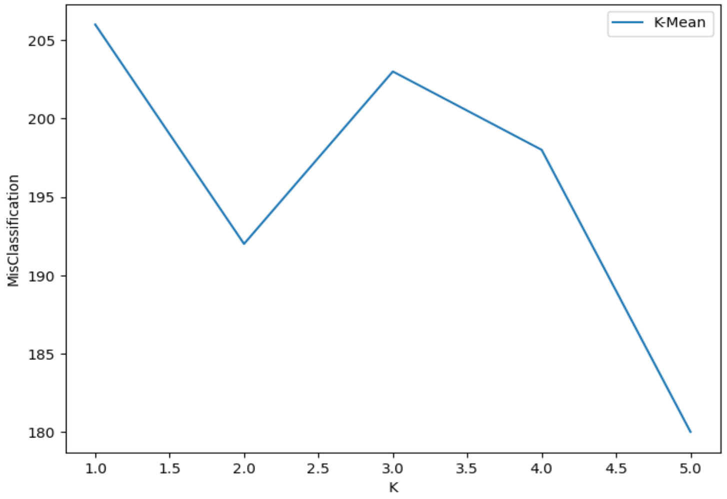


Fig:2.2

From **Table 2.1** and **Fig:2.1** We can see how **K-Means** behave on the given **Spiral DataSet**. SSE value decreases rapidly with increase in K to some extent but after some value of K it decreases slowly. Here till **K=2,** SSE decreases rapidly. After this It decreases slowly . However SSE is very high for all K.

Similarly we can see from **Fig: 2.2** how misclassifications depend on K.

So From above observations, we can take **K=2 or K=3** as the best representation of the dataset. But From **Table 2.1 and Fig:2.2** we can see that misclassifications are very high for all the K. As we have considered **K=2 or K=3** as the best representation from the observation but it also has a very high misclassifications as we have a total **312 label points** out of which it has **192 and 203** misclassifications respectively.

It's because **K-Means works better on only convex dataset** because of means calculation.So It is **not providing good clusters representation** of the Spiral dataset.

Hence we would not be able to find the best Suited K for Spiral data from K-Means.

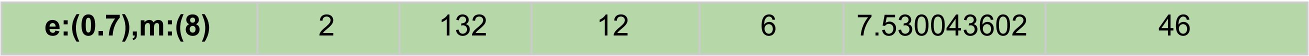
But for this case K=2 represents best.

2. DBScan

DataSet 1: Iris

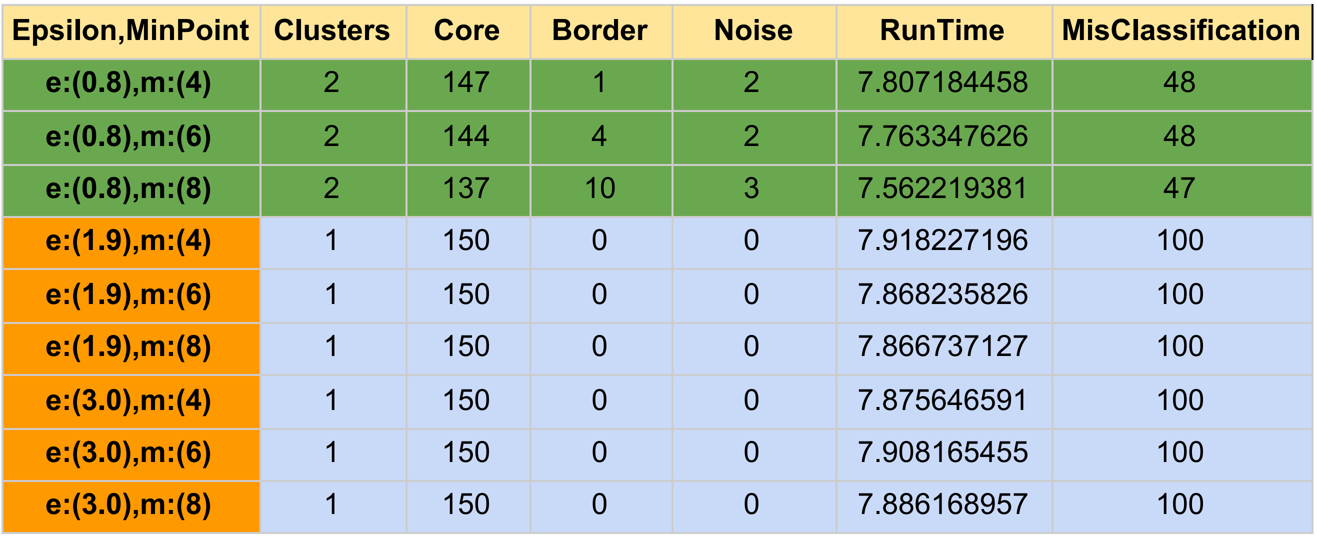
Table 2.1.1

Epsilon,MinPoint	Clusters	Core	Border	Noise	RunTime	MisClassification
e:(0.2),m:(4)	3	16	6	128	4.453220606	0
e:(0.2),m:(6)	2	8	9	133	4.20665288	0
e:(0.2),m:(8)	1	2	6	142	4.087703228	0
e:(0.39),m:(4)	4	104	20	26	6.735896587	3
e:(0.39),m:(6)	4	77	39	34	6.093718767	3
e:(0.39),m:(8)	3	41	33	76	5.12462306	0
e:(0.7),m:(4)	2	143	4	3	7.904423475	47
e:(0.7),m:(6)	2	135	10	5	7.589650869	47



e:(0.7),m:(8)	2	132	12	6	7.530043602	46

Table 2.1.2



Epsilon,MinPoint	Clusters	Core	Border	Noise	RunTime	MisClassification
e:(0.8),m:(4)	2	147	1	2	7.807184458	48
e:(0.8),m:(6)	2	144	4	2	7.763347626	48
e:(0.8),m:(8)	2	137	10	3	7.562219381	47
e:(1.9),m:(4)	1	150	0	0	7.918227196	100
e:(1.9),m:(6)	1	150	0	0	7.868235826	100
e:(1.9),m:(8)	1	150	0	0	7.866737127	100
e:(3.0),m:(4)	1	150	0	0	7.875646591	100
e:(3.0),m:(6)	1	150	0	0	7.908165455	100
e:(3.0),m:(8)	1	150	0	0	7.886168957	100

Table 2.1.1 and Table 2.1.2 both are observations for Iris dataset on different values of **Epsilon and MinPoints.** We can see from both the tables that if we increase minPoints core points decrease and noise and border points increase. Because for the same radius if we would increase the number of points inside the circle is going to decrease or it may be that we would not be able to find any points. From Both Table We can say that **K =2** is best suited but here comes different conditions for it like in **Table 2.1.1** misclassifications for **K=2** where we can see that for other K's misclassifications is less but most of the points become noise because of low value of radius.

But from **Table 2.1.2** we can clearly see that for a high value of radius we get less misclassifications for **K=2** but however it is not good because **47 out of 150** misclassifications is not going to represent our data best as compared to K-Means algorithms. But for this case K=2 represents best..

If we increase radius all the points would fall in the same cluster which results in high misclassifications.

DataSet 2: Spiral

Table 2.2.1

Epsilon,MinPoint	Clusters	Core	Border	Noise	RunTime	MisClassification
e:(0.8),m:(4)	4	92	5	215	23.56340027	0
e:(0.8),m:(6)	4	38	16	258	19.07578135	0
e:(0.8),m:(8)	2	23	11	278	17.753896	0
e:(1.9),m:(4)	3	298	5	9	32.2869277	0
e:(1.9),m:(6)	3	180	9	123	26.03496718	0
e:(1.9),m:(8)	3	117	10	185	22.64236522	0
e:(3.0),m:(4)	3	312	0	0	33.07237601	0
e:(3.0),m:(6)	3	306	6	0	32.70800185	0
e:(3.0),m:(8)	3	227	12	73	28.82156372	0

Table 2.2.1 is observations for the **Spiral dataset** on different values of **Epsilon and MinPoints.** We can see from the table that if we increase minPoints core points decrease and noise and border points increase. Because for the same radius if we would increase the number of points inside the circle is going to decrease or it may be that we would not be able to find any points within the radius .

Here K=3 is best suited because most of the points are either core or border and very less points are noise. The algorithm has been run on different values of Epsilon and Min Points. Here we can see that misclassifications is zero for all the provided observations but does not mean that algorithm is going to work on every combination of epsilon and min points as from the table we can see that for many such pairs the algorithm gives misclassification 0 but most of the points are noise.

Hence for the given K=3 we have found that it represents the dataset best because of less noise points and misclassifications.

The behaviour of the algorithm is because it mainly depends on the density of the dataset.

CONCLUSION

Best Suited K for both DataSet:

Iris DataSet

Table 3.1

Algorithms	Misclassification	RunTime(s)	Cluster(K)
K-Means	17	0.1986601353	3
DBScan	46	7.530043602	2

Spiral DataSet

Table 3.2

Algorithms	Misclassification	RunTime(s)	Cluster(K)
K-Means	192	0.4507431984	2
DBScan	0	32.2869277	3

From both the Table 3.1 and 3.2 we can clearly say that **K-Mean takes less time to execute than DBScan.**

I have chosen misclassification to compare between both the algorithms.

Misclassifications for Iris data for K-Mean is less than DBScan and for Spiral data for K-Mean is higher than DBScan.

Hence from misclassification observations it is clear that K-Mean best represents Iris dataset and DBScan best represents Spiral dataset.

Hence K=3 for by considering the algo on both the dataset is best suited.

Here some overall observations that I found.

DBScan is Robust to outliers. It can find arbitrarily shaped clusters.

DBScan can not cluster a dataset with large differences in densities.

KMeans clustering does not do a good job when data is not spherically or convex distributed.

DBScan is very sensitive to the MinPoints requirements and maximum radius.