Homework3: Clustering

Mingyi Xue[*]

June 7, 2018

1 Problem 1. K-means clustering

1.1 Algorithm

k-means objective,

$$J = \sum_{k=1}^{K} \sum_{\boldsymbol{x}_n \in C_k} \|\boldsymbol{x}_n - \boldsymbol{m}_k\|_2^2 \tag{1}$$

Firstly, initial centers are randomly selected and stored in a list. Secondly, in each iteration, we compute distances from the current point to each center in the list. The shortest distance will be added directly to objective described above and the index of current point will be appended to the corresponding value list in a cluster dictionary. Finally, the list of centers will be updated with infomation of cluster dictionary.

Table 1: Data Structure

name	structure	element
X	numpy.ndarray	shape of (N,m)
centers	list	numpy.ndarray, shape of (1,m)
dit	dictionary	key : index of centers, value : list of row indices in X
J	list	real number
time_lst	list	time.time()

Table 2: Default parameters

name	value
# of centers	10
# of iteration	40

[*] GSP student from Nanjing University

1.2 Result

Table 3: Result of dense dataset

iteration	objective function
10	55737.842681
20	55685.979060
30	55685.808653
40	55685.808653
total time(sec)	345.9897

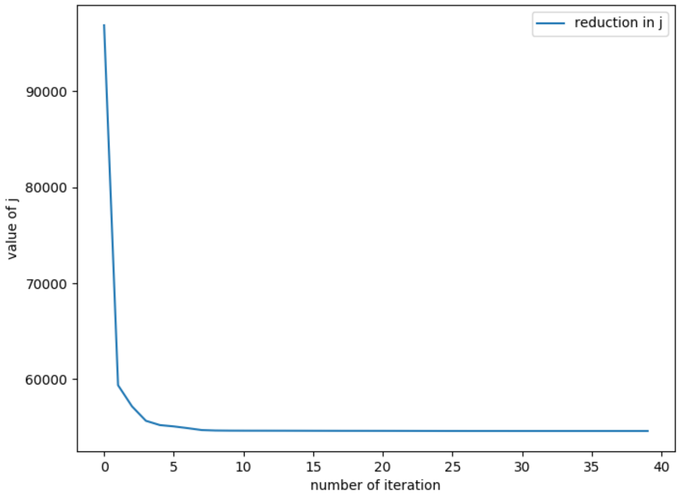


Figure 1: Reduction in objective function

1.3 Conclusion

I find that the program have converged within 40 iterations since the objective of the 40^{th} iteration is the same as that of the 30^{th} iteration. Besides, the derivative of reduction in objective is always negative and monotonically increasing.

2 Problem 2. K-means for sparse data

2.1 Algorithm

Almost the same with Problem 1, except that dataset X is a scipy.sparse matrix and that centers are initialized with the first 10 rows in the dataset to avoid extra computation.

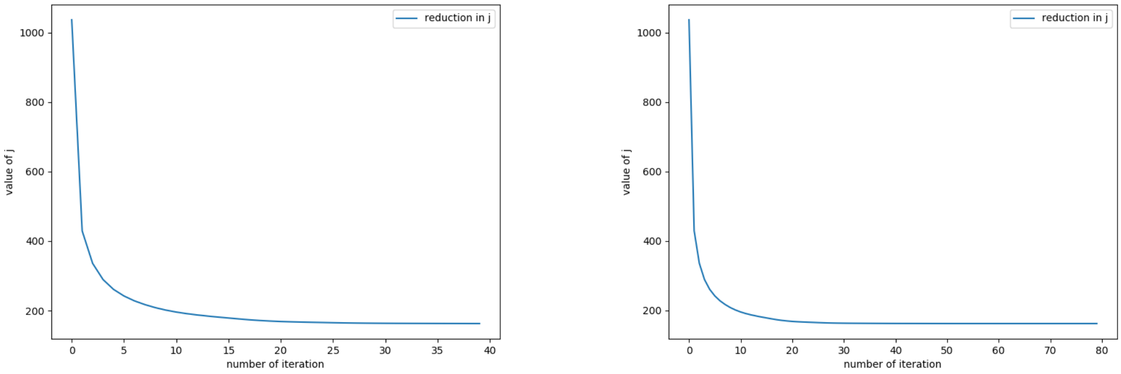
Table 4: Default parameters

name	value
# of centers	10
# of iteration	40

2.2 Result

Table 5: Result of sparse dataset

iteration	objective function
10	201.334622
20	169.798514
30	163.539468
40	162.579801
approximate time/iter(sec)	240



(a) 40 iterations

(b) 80 iterations

Figure 2: Reduction in objective function

2.3 Conclusion

I find that the program have not converged yet at the 40^{th} iteration since the objective of the 40^{th} iteration does not equal that of the 30^{th} iteration, though quite close. Besides, the derivative of reduction in objective is always negative and monotonically increasing.