report.md

Data Mining Lab 3

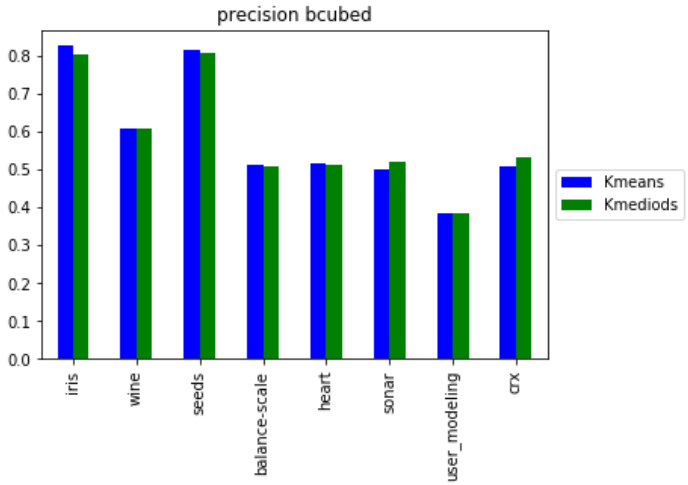
Abdullah Al Thaki

Roll: 12

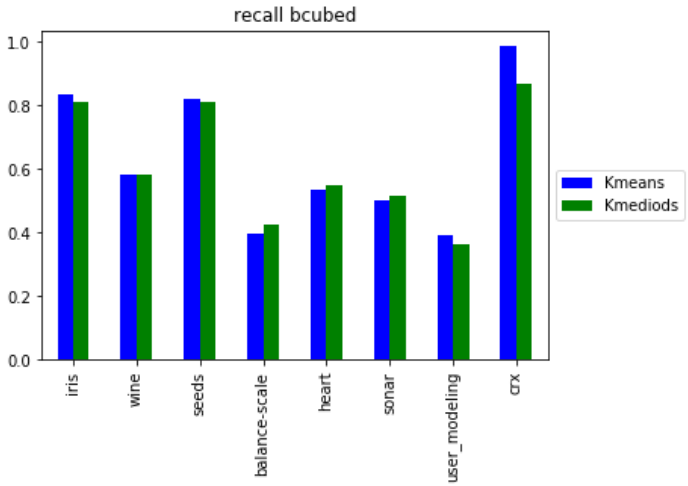
Introduction

K-means and K-medoids are two well known clustering algorithm for unsupervised setting. K-means clustering uses mean as the centroid while K-medoid clustering uses medoid as the centroid. Even though K-means clustering is faster than K-medoids, K-medoids doesn't have outlier problems. Here I have experimented these clustering methods on several real life datasets.

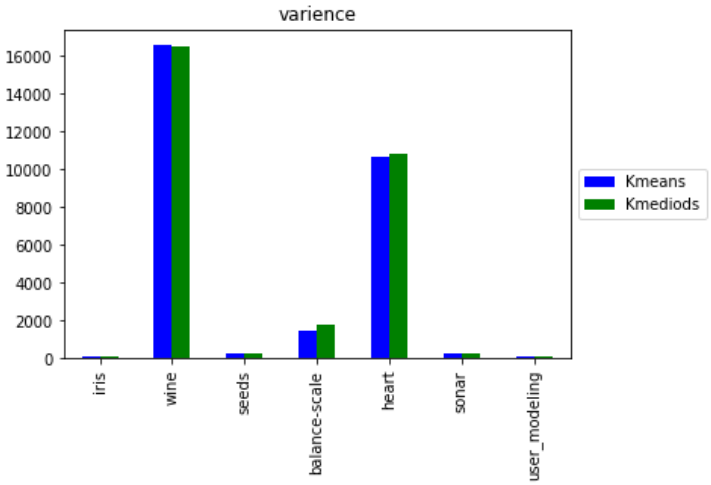
Precision bcubed



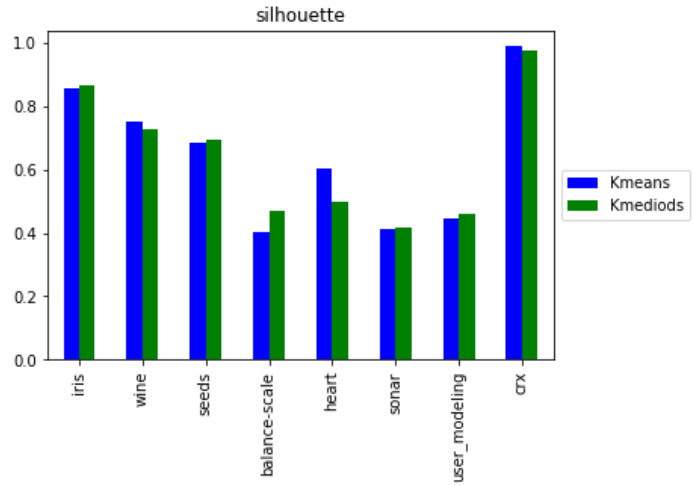
Recall bcubed



Varience

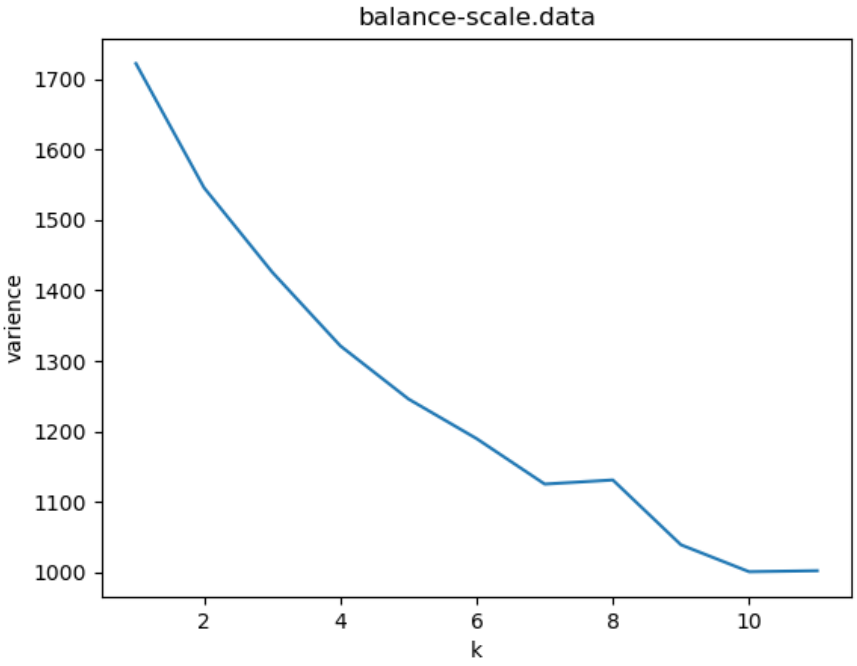


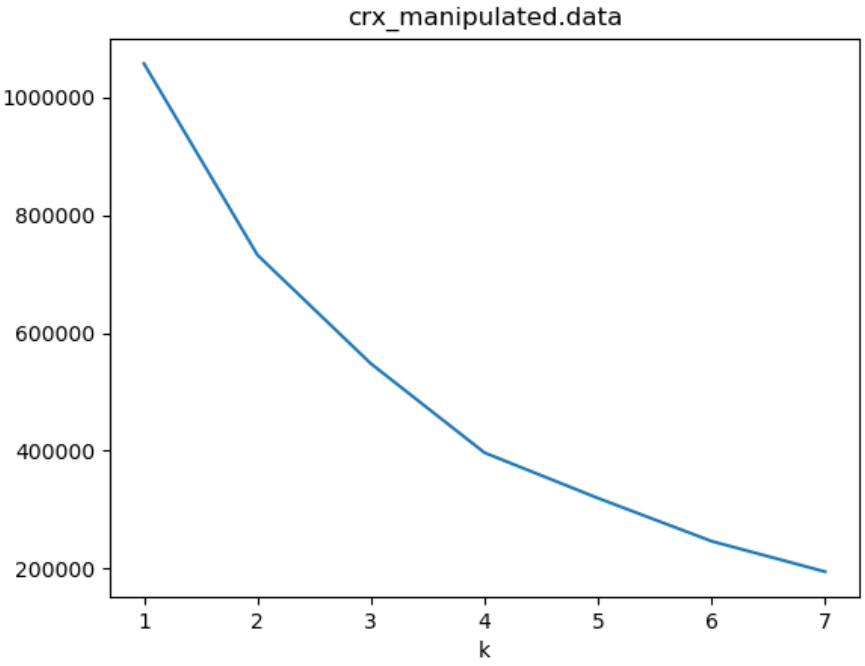
Silhouette

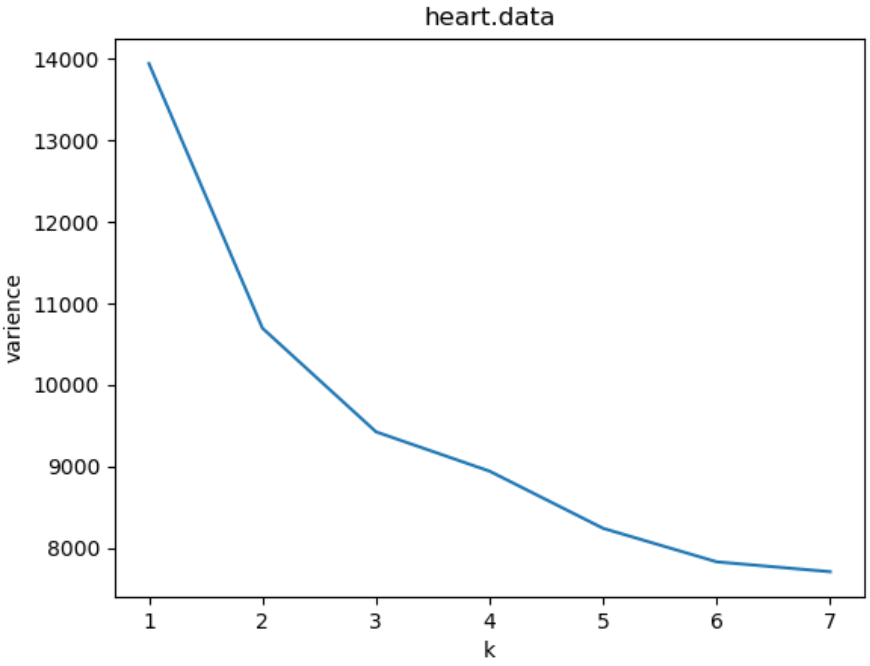


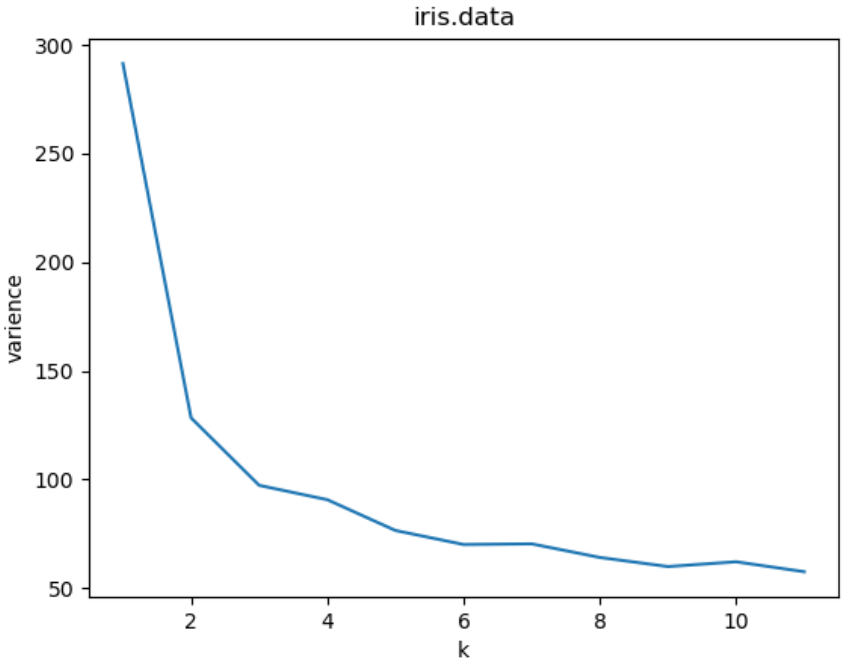
Elbow method

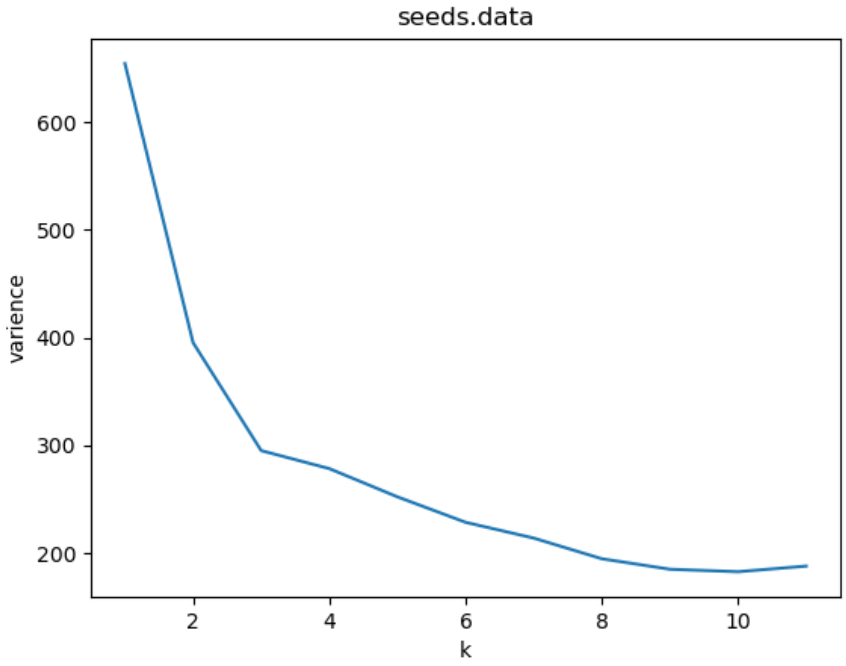
Elbow method is used to find out best k. I have manipulated Kmeans clustering algorithm to show elbow method.

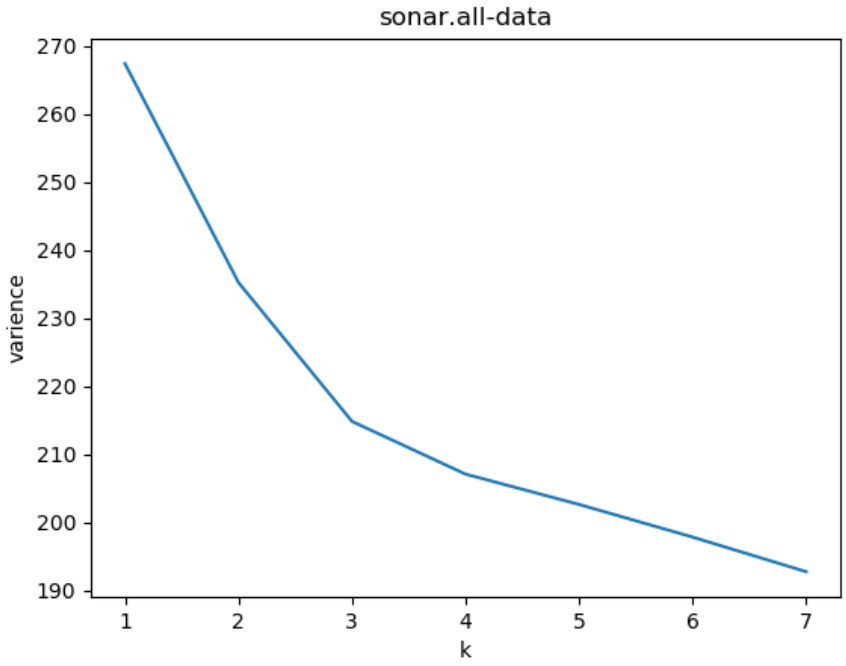

balance-scale.data
varience
1700
1600
1500
1400
1300
1200
1100
1000
2
4
6
8
10
k

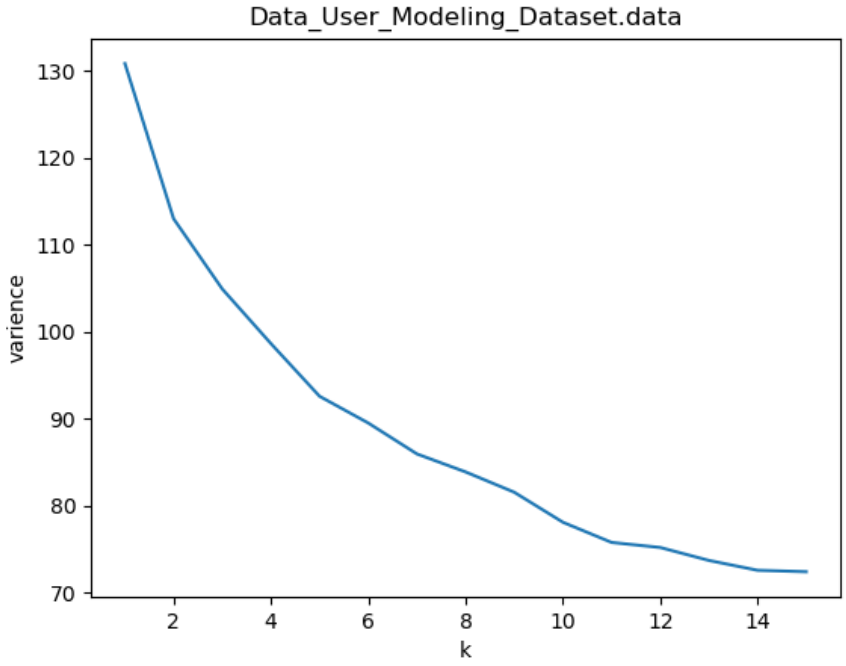

crx_manipulated.data
1000000
800000
600000
400000
200000
1
2
3
4
5
6
7
k

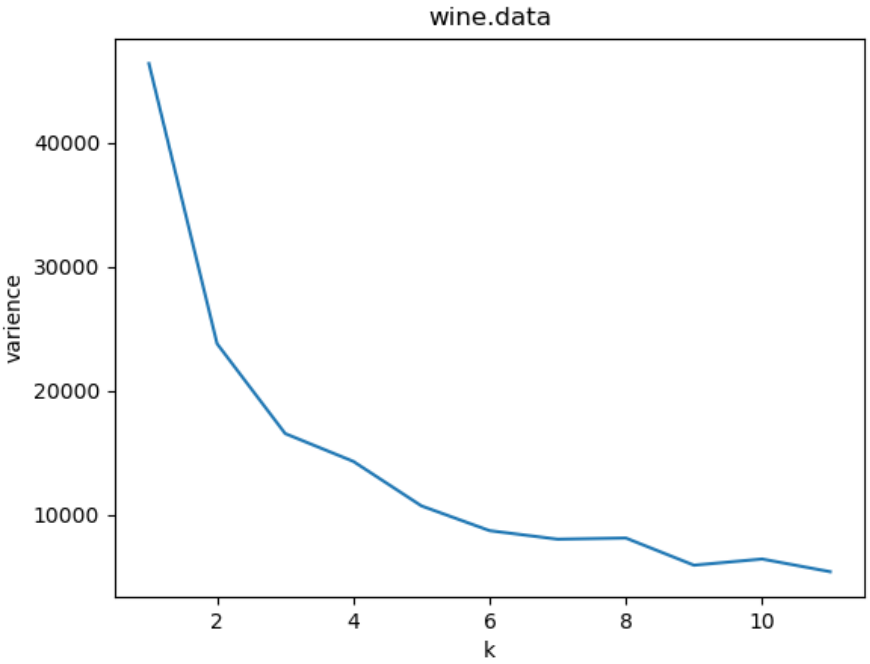

heart.data
varience
k
14000
13000
12000
11000
10000
9000
8000
1
2
3
4
5
6
7


iris.data
varience
k
300
250
200
150
100
50
2
4
6
8
10


seeds.data
varience
k
600
500
400
300
200
2
4
6
8
10


sonar.all-data
varience
k
270
260
250
240
230
220
210
200
190
1
2
3
4
5
6
7





Findings

From elbow method, the real k is not very intuitive all the time. However elbow method can give us some hint. Kmeans takes very less time compare to Kmedoids. Kmeans has very low implementation complexity. But Kmedoids can give better result compared to Kmeans.

In the four evaluation metrics, there is no exclusive winner. Kmeans have done well in some cases, Kmedoids have done well in others. However the result could be different in some other real life huge/complex dataset.