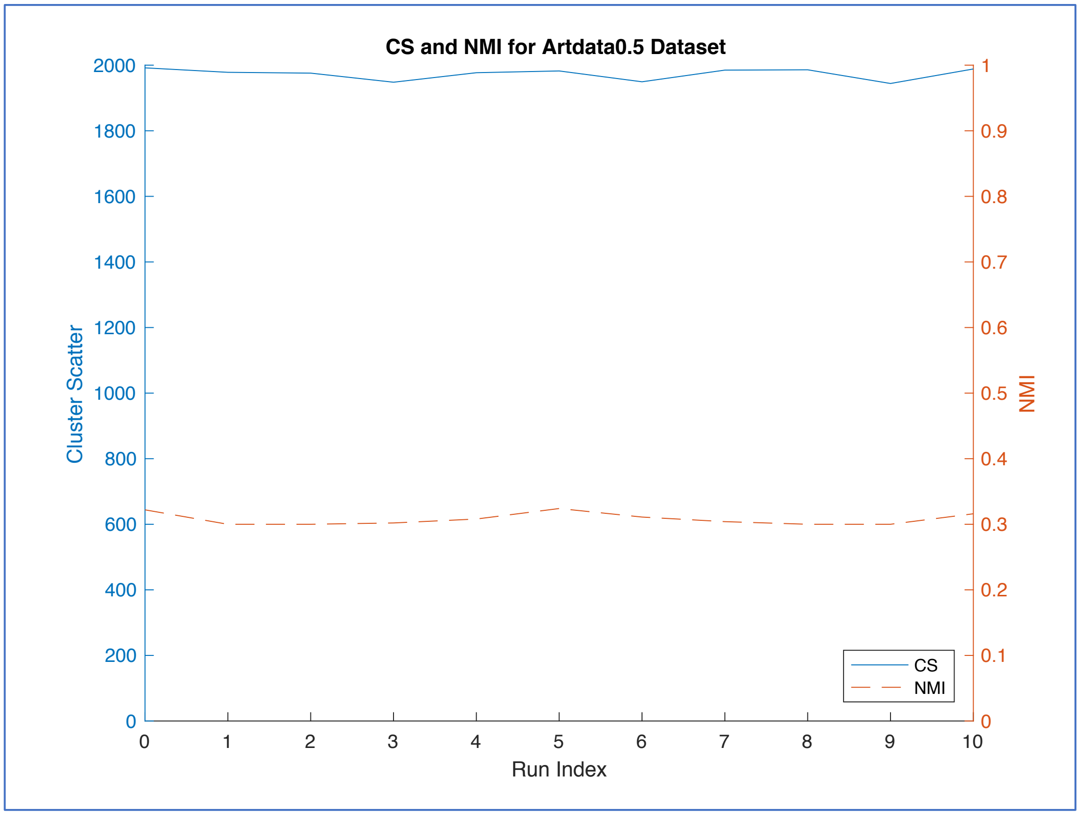
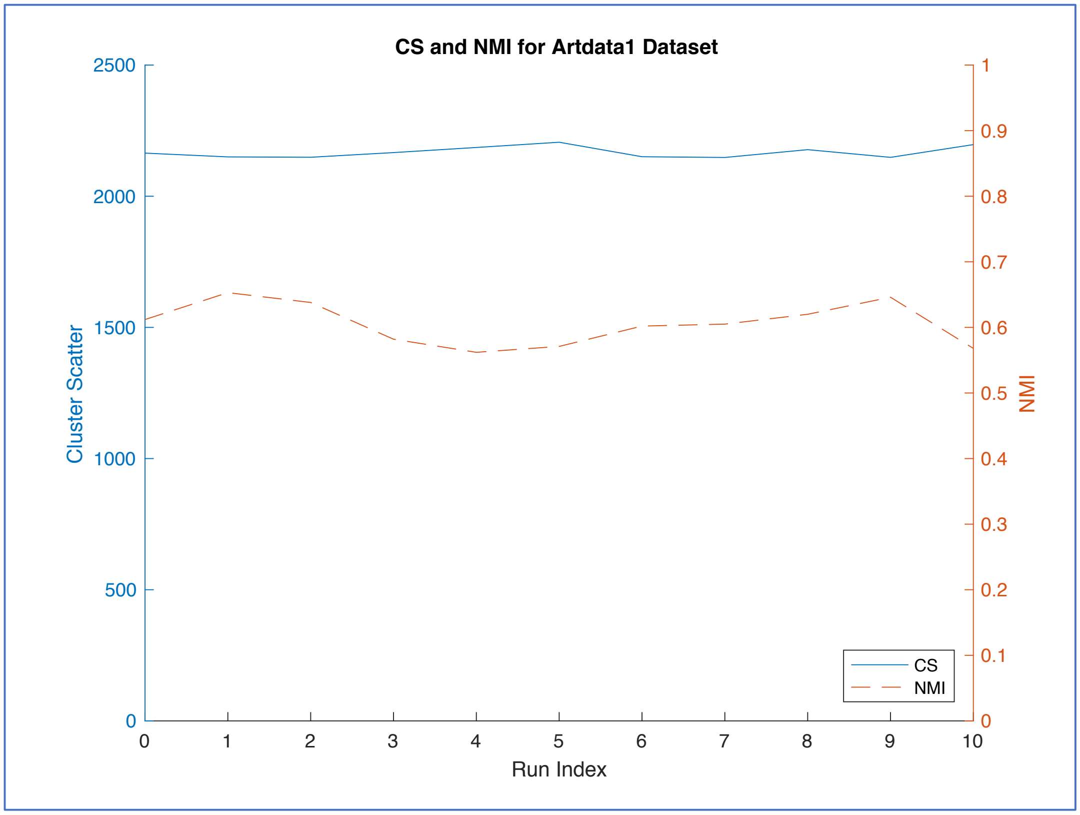
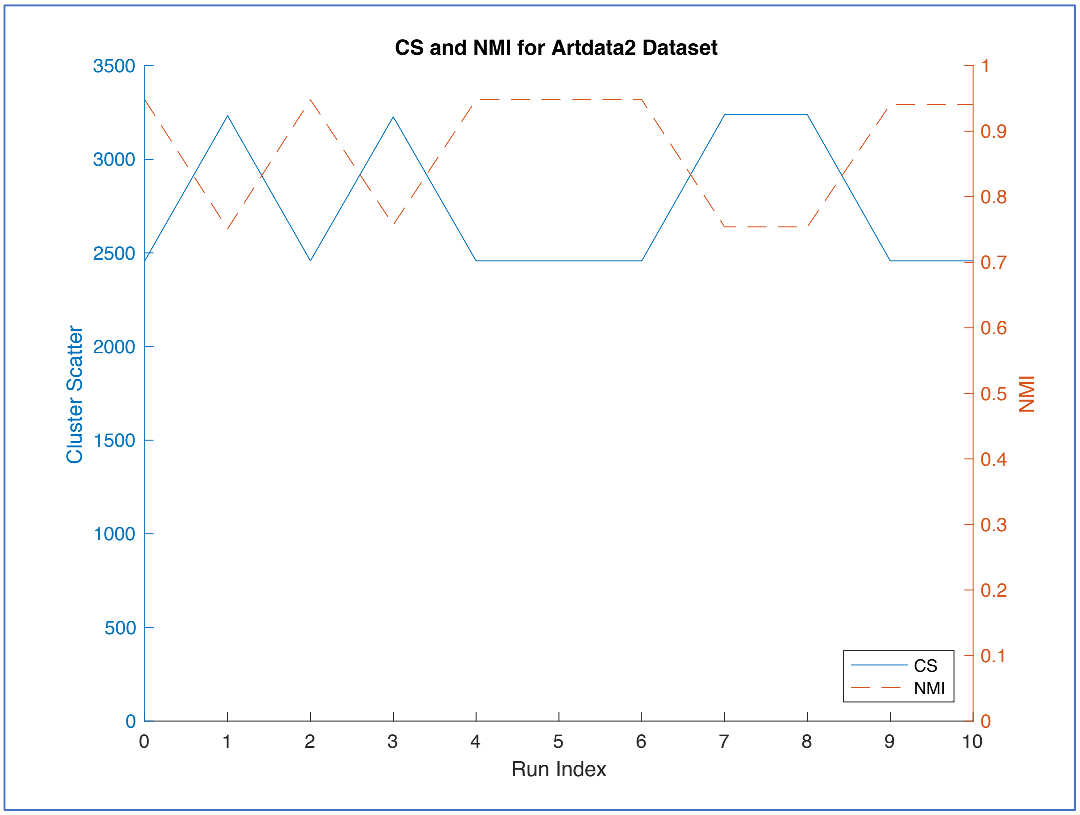
Kenny Alperin
COMP 135
Programming Assignment 3

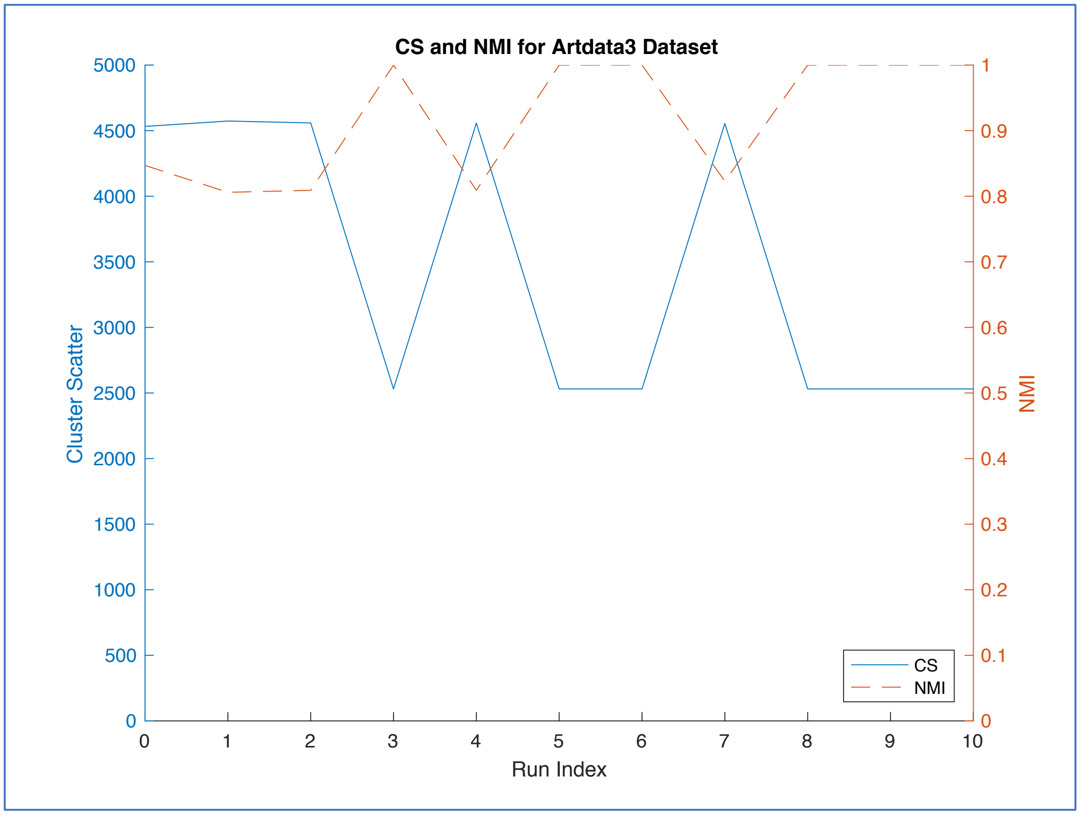
Section 2-2

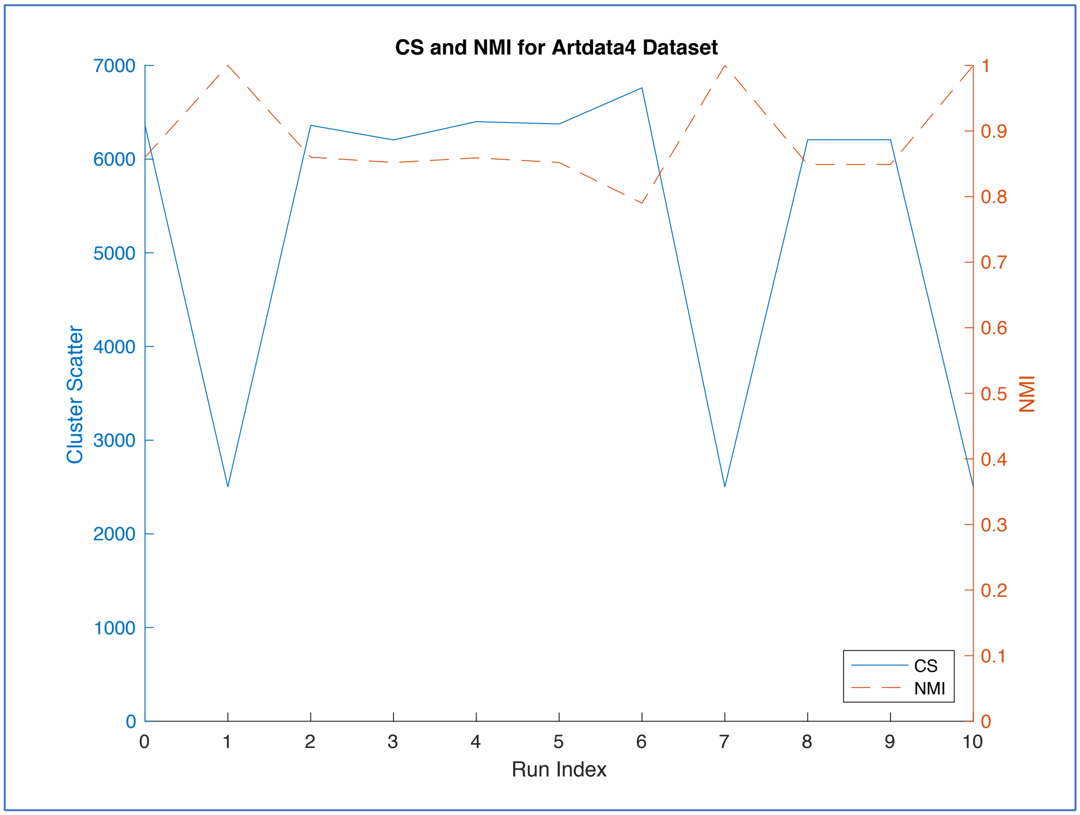
Below are the eight plots representing each dataset for the 10 random restart initializations and 1 smart initialization. Note that Run Index = 10 represents the smart initialization clustering run. The solid line in each plot of this section is the cluster scatter, and the dashed line in each plot of this section is the NMI for the clustering.

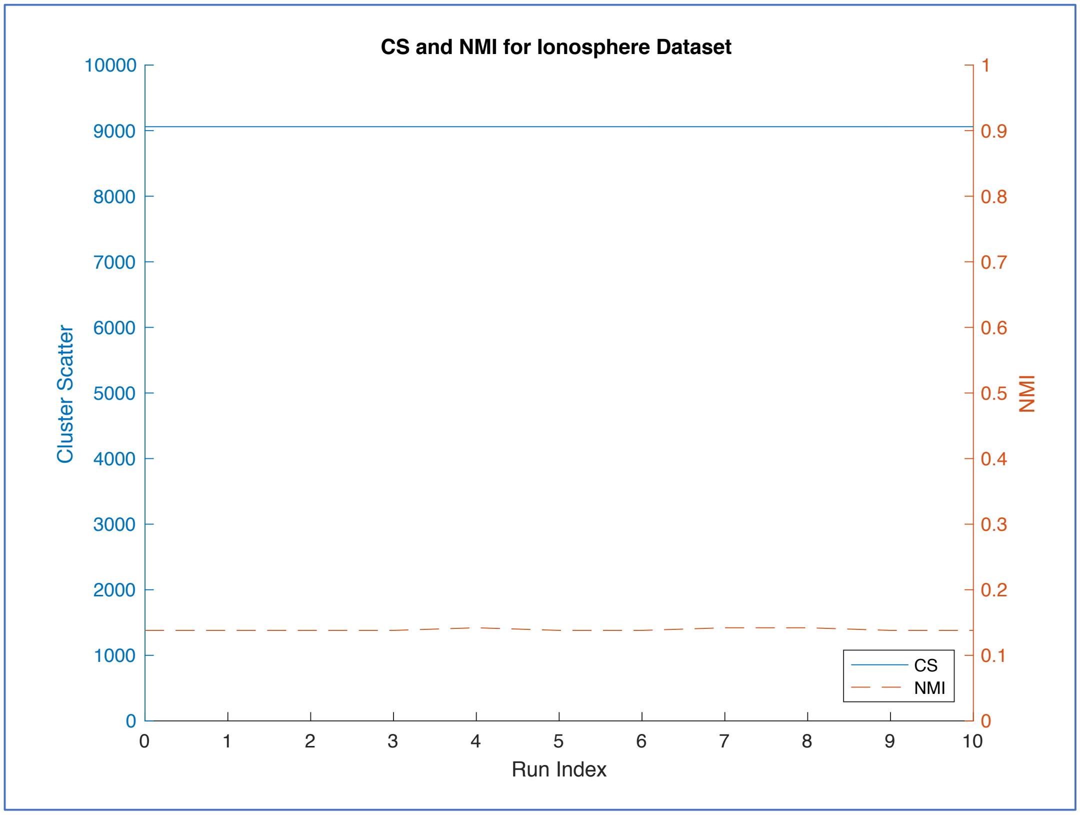


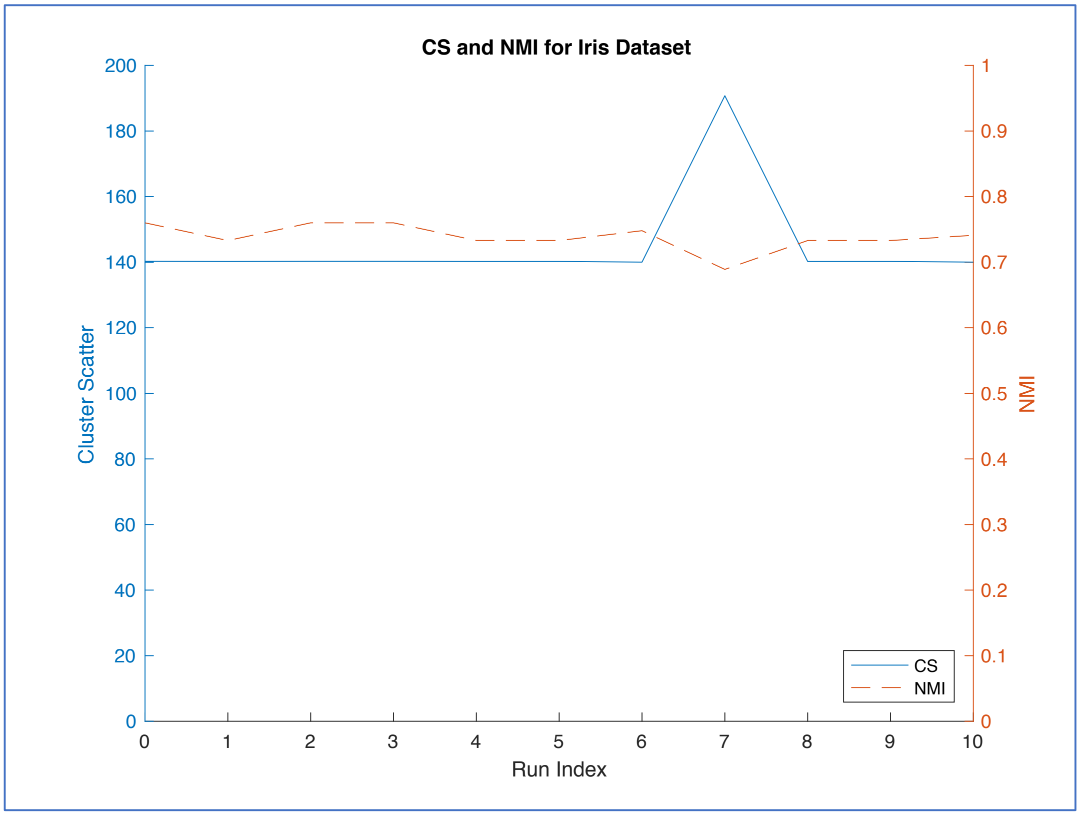

CS and NMI for Artdata1 Dataset
Cluster Scatter
0
500
1000
1500
2000
2500
NMI
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
1
2
3
4
5
6
7
8
9
10
Run Index
CS
NMI

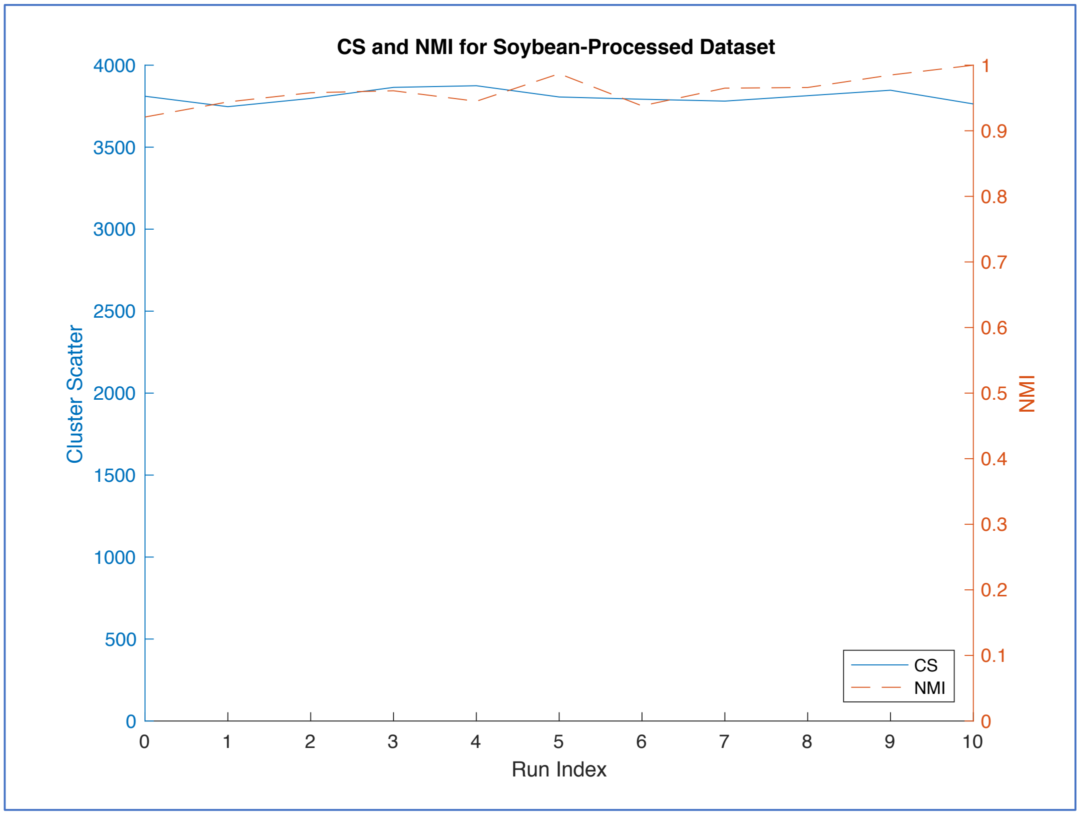

CS and NMI for Artdata2 Dataset
Cluster Scatter
0
500
1000
1500
2000
2500
3000
3500
NMI
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
1
2
3
4
5
6
7
8
9
10
Run Index
CS
NMI


CS and NMI for Artdata3 Dataset
Cluster Scatter
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
NMI
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7
8
9
10
Run Index
CS
NMI


CS and NMI for Artdata4 Dataset
Cluster Scatter
7000
6000
5000
4000
3000
2000
1000
0
NMI
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7
8
9
10
Run Index
CS
NMI


CS and NMI for Ionosphere Dataset
Cluster Scatter
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
NMI
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7
8
9
10
Run Index
CS
NMI

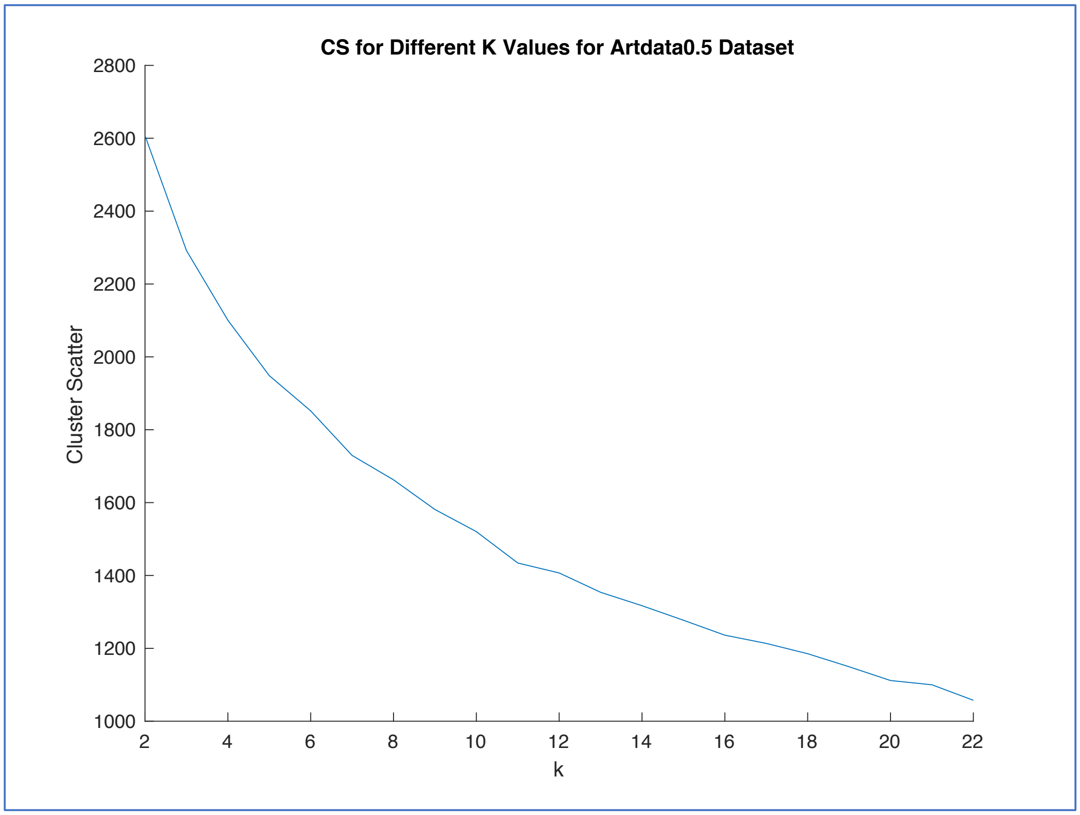

CS and NMI for Iris Dataset
Cluster Scatter
200
180
160
140
120
100
80
60
40
20
0
NMI
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
7
8
9
10
Run Index
CS
NMI

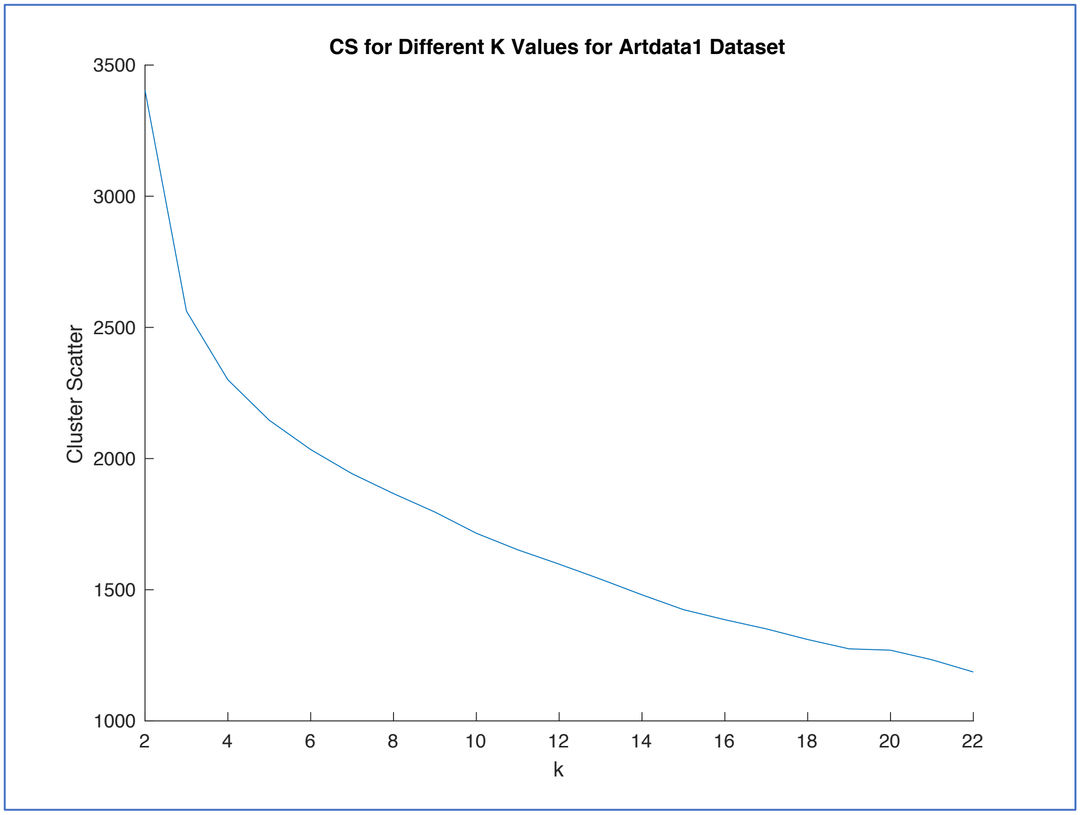


Some of the datasets have consistent CS and NMI across each run, including the smart initialization, but some have a lot of variety in the values. For example, the Ionosphere dataset has steady CS and NMI values. To the contrary, the Artdata2 dataset has the CS and NMI values jump for each run. When there was a decrease in CS from one run to the next, there was a corresponding increase in NMI, which shows a decrease in CS leads to an increase in NMI, since the data is better clustered. For most of the datasets, the smart initialization had the lowest CS and highest NMI scores. In all of the datasets, using only the CS as criterion over a bunch of runs would have also led to picking the right clustering. Another observation is that just because two datasets have similar cluster scatters, that doesn't mean they have similar NMIs. This is likely because different datasets have different numbers of attributes to account for, so the scores are only relative to a particular dataset.

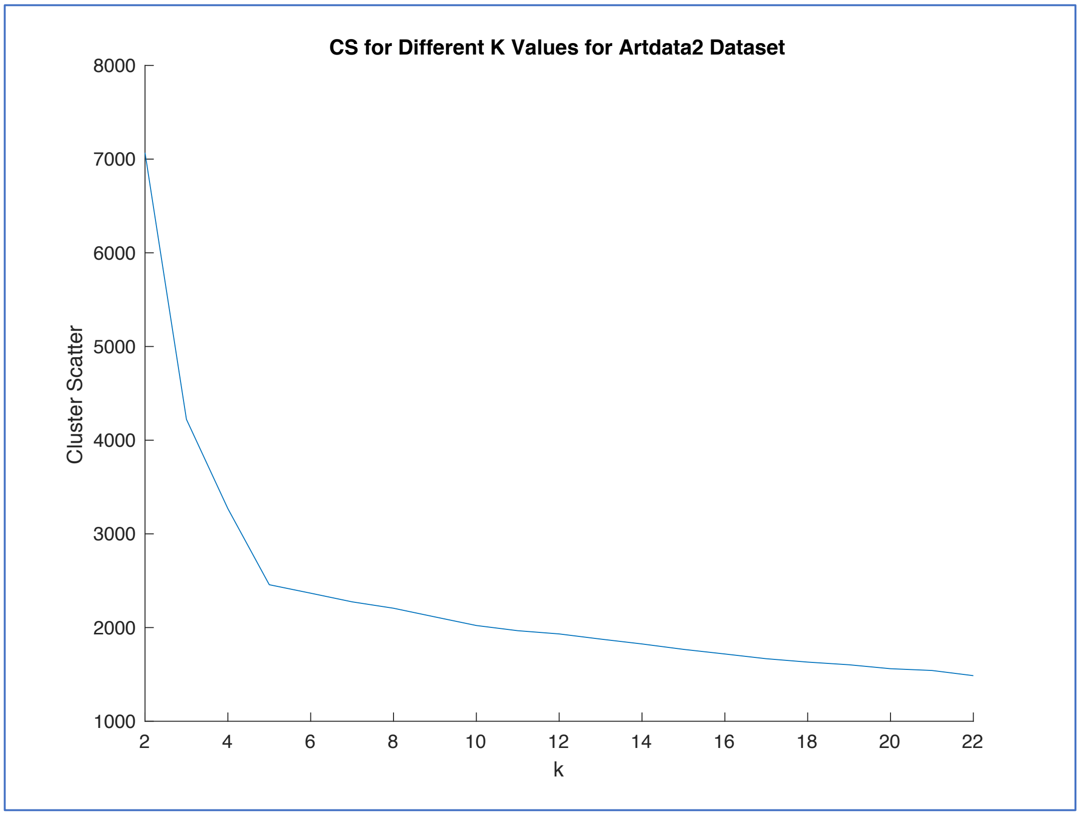
Section 2-3

Below are the eight plots representing each dataset for running clustering for k values from 2 to 22.

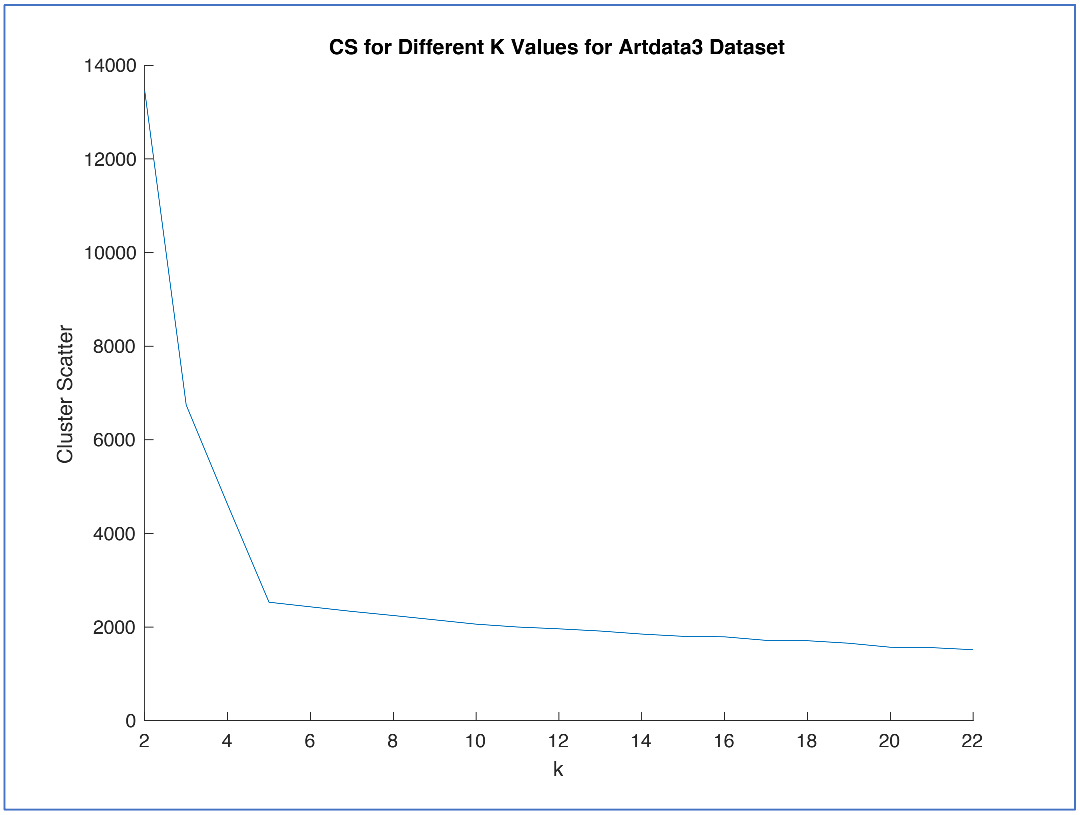


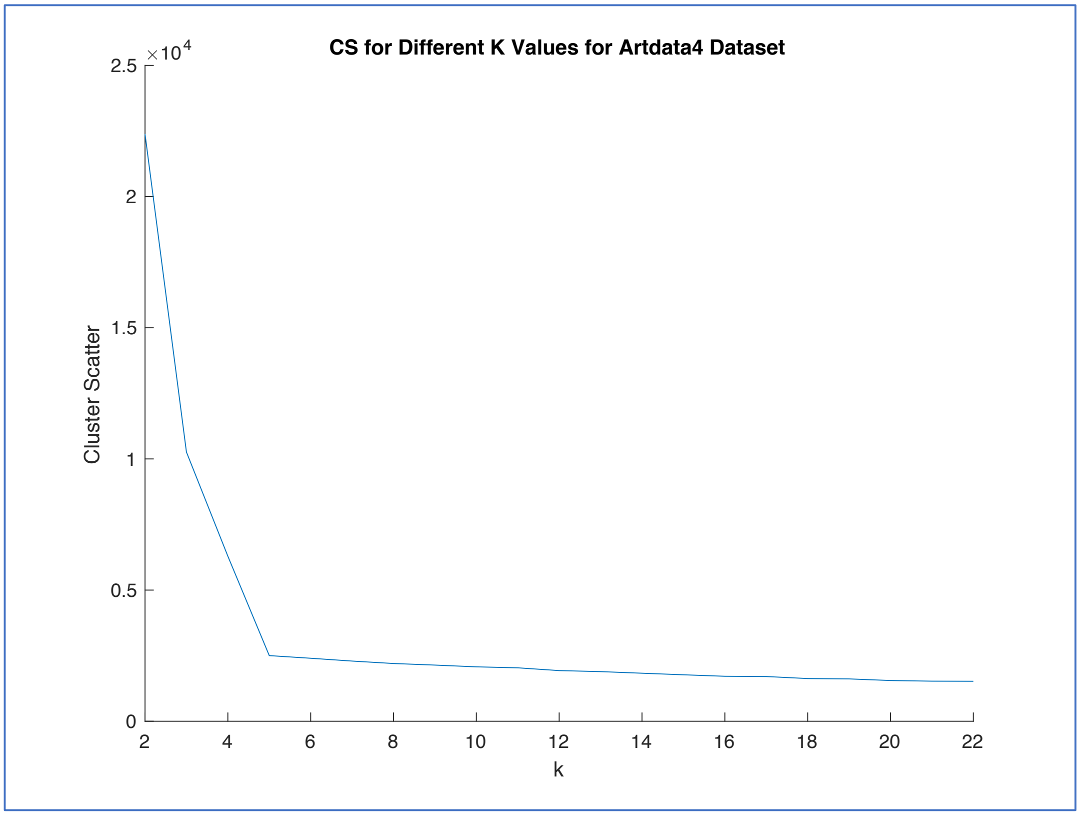


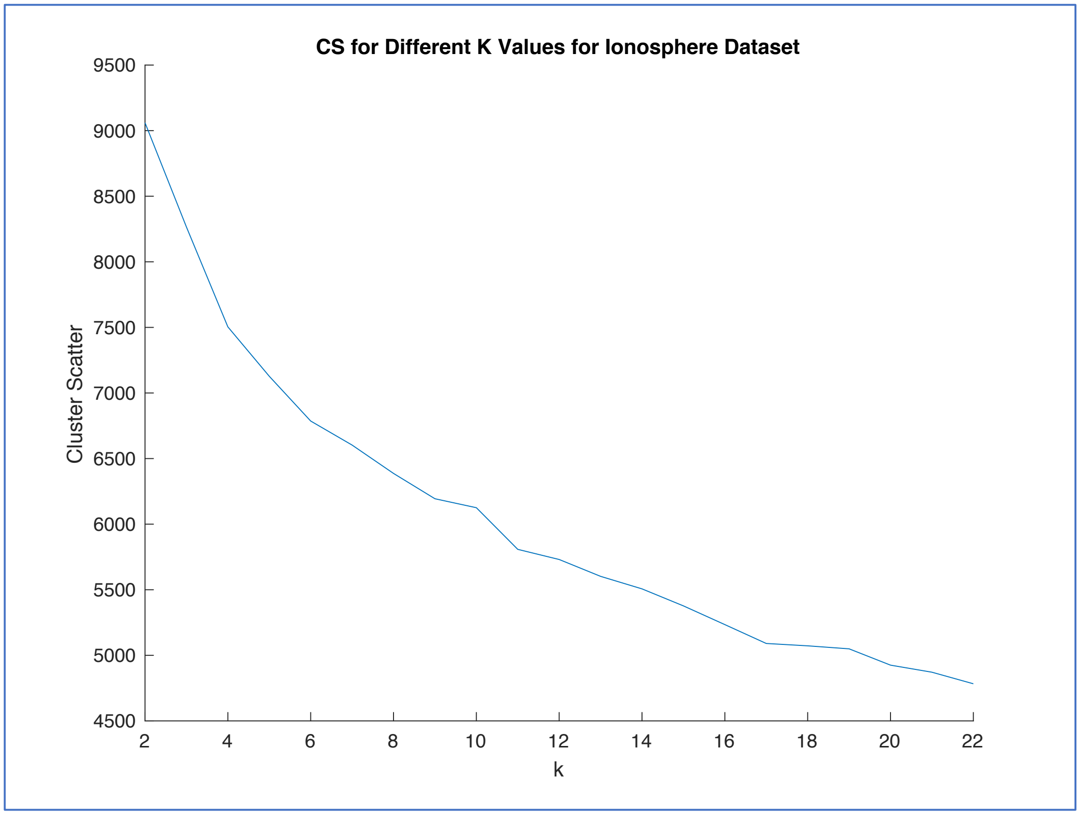
CS for Different K Values for Artdata1 Dataset

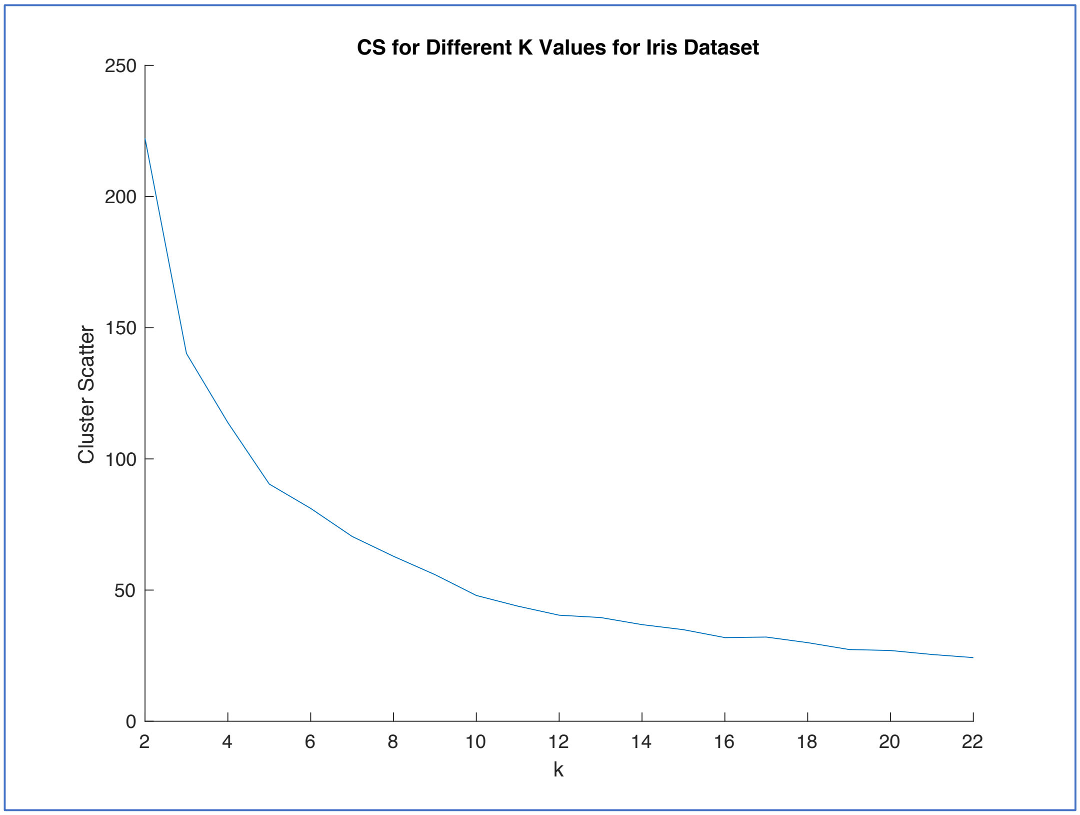


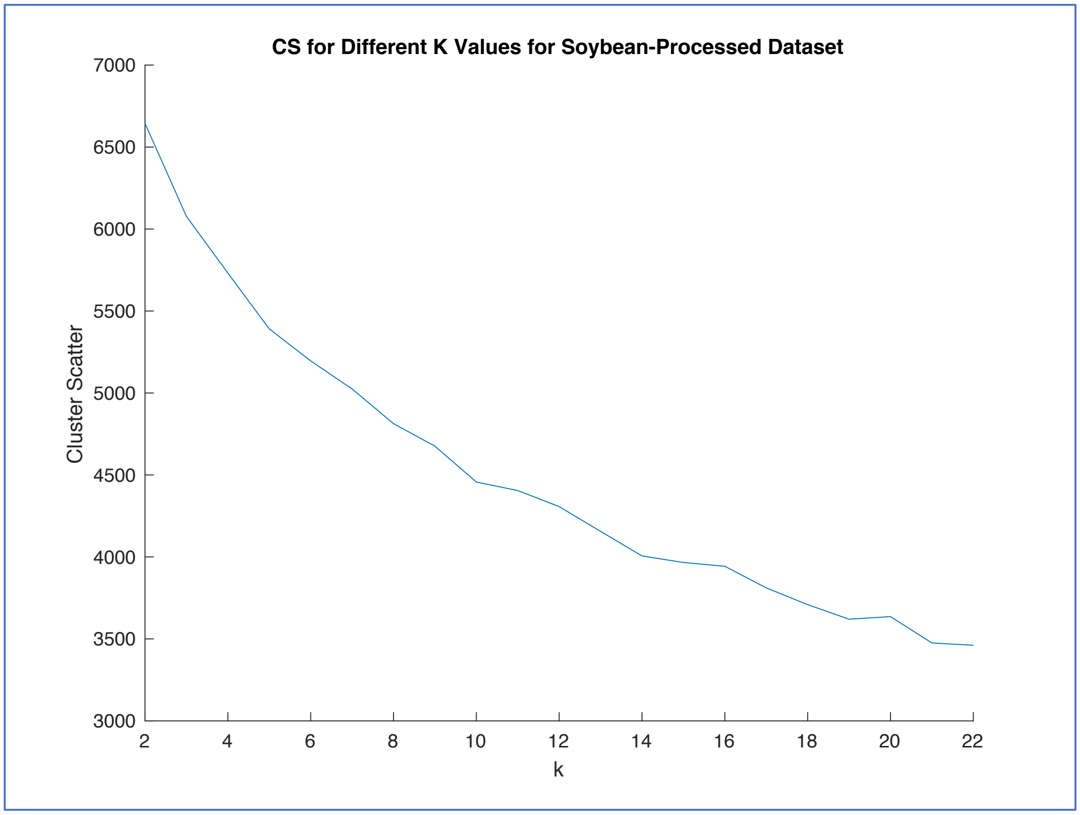
CS for Different K Values for Artdata2 Dataset


CS for Different K Values for Artdata3 Dataset
Cluster Scatter
14000
12000
10000
8000
6000
4000
2000
0
2
4
6
8
10
12
14
16
18
20
22
k


CS for Different K Values for Artdata4 Dataset
Cluster Scatter
×10^4
2.5
2
1.5
1
0.5
0
2
4
6
8
10
12
14
16
18
20
22
k


CS for Different K Values for Ionosphere Dataset
Cluster Scatter
9500
9000
8500
8000
7500
7000
6500
6000
5500
5000
4500
2
4
6
8
10
12
14
16
18
20
22
k


CS for Different K Values for Iris Dataset
Cluster Scatter
250
200
150
100
50
0
2
4
6
8
10
12
14
16
18
20
22
k



It makes sense that the cluster scatter decreases as the k value increases since there are more clusters, allowing for better grouping of the data. However, I would look at values of k where the rate of CS decreasing decreases, as in look for the points where there appears to be a corner and the CS decreases at a smaller rate. For example, in the Iris dataset, there is one of these corners at k = 3, which suggests a good k value. This is supported by the fact that the Iris dataset does have 3 classes. From the artificial datasets, it is easier to see the corners for picking k in the datasets that have values farther apart (artdata4), and it is difficult to pick a k from artdata0.5. As the artificial datasets have values that are farther apart, the CS increases, but it is also easier to see where the best k value is. For the Ionosphere dataset, there are two classes, so you can't see the corner in the plot.