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.

















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.

















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.