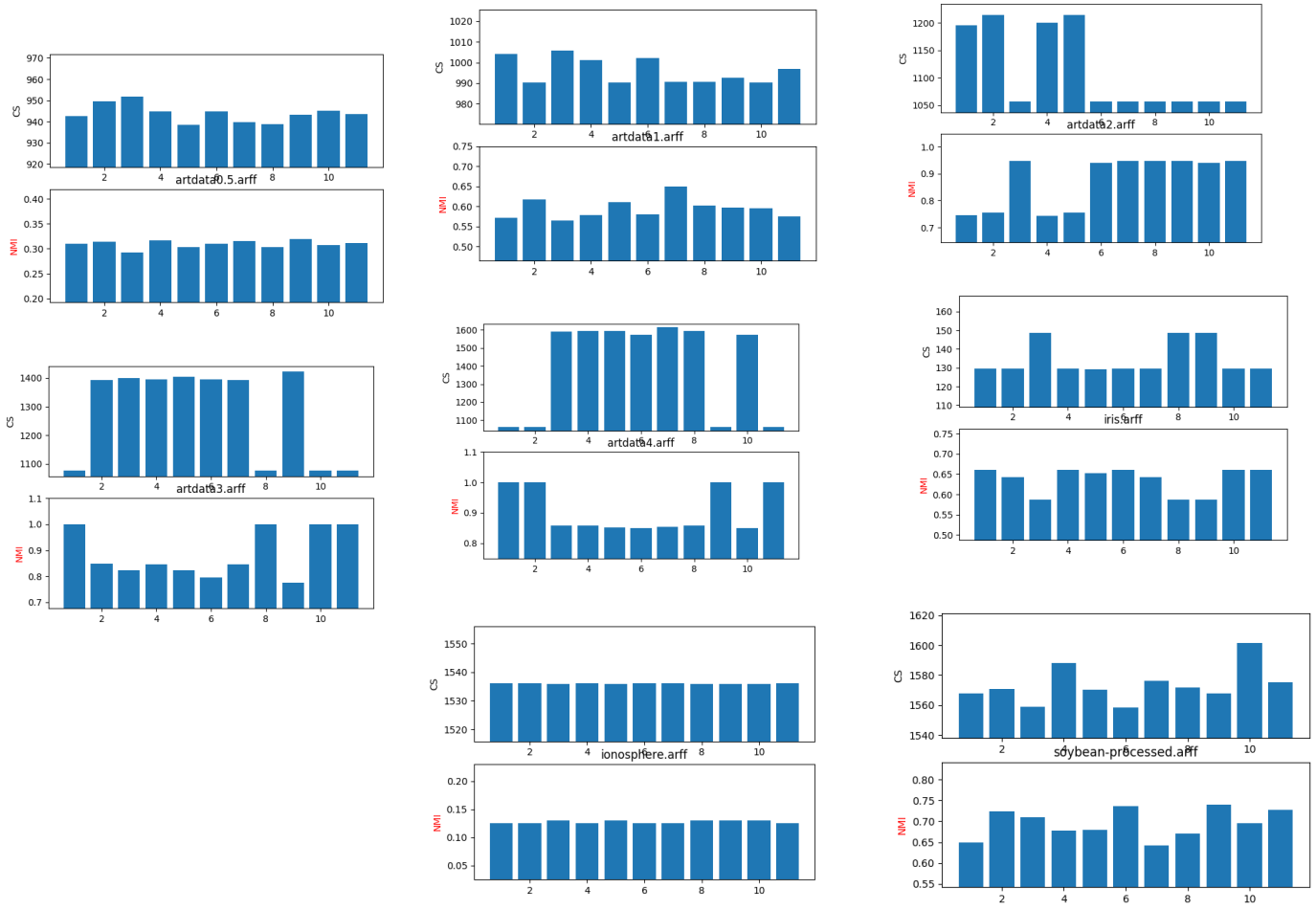
Ethan Hartzell
Machine Learning
Project 3

Below are the results of the first experiment. Each graph is labeled with the dataset it corresponds to. The top portion is Cluster Scatter and the bottom portion is Normalized Mutual Information. The first 10 bars are the random initializations, and the last one is the smart initialization. In the artificial data sets, it's clear that the closer the gaussian spheres are together the harder it is for the data to be clustered. As a result we get low NMI values for all runs, but comparatively lower clusters scatters to other data on the artdata0.5 set since the clusters are all close together. On the data sets where the clusters are further apart, the cluster scatter values have a clear inverse relationship with the NMI values, since correctly clustered data shouldn't have means stuck between clusters (which would yield clusters with high scatters). This does happen on some of the random runs, while others achieve much higher NMI values. These values are less stable across random runs on harder to cluster data. Lowest cluster scatter does not always correspond to highest NMI on some of the harder to cluster data. However, the smart-initializations are always among the highest NMI values, that is, they are more reliable than the random initializations, but not always the best on harder to cluster data. The easy-to-cluster datasets are able to achieve NMI values at or near 1.0. The ionosphere data gets very low NMI values, so this algorithm may not be able cluster data in that shape very well.



Below are the results for the K experiments. Each graph is labeled with its corresponding dataset. As K increases, cluster scatter goes down, which is expected because there are less examples in each cluster to increase the distance. As the distance between spheres increases, the more clearly the knee criterion can be used to judge the appropriate K value. This is most clear on artdata4.arff that the number of clusters should be five. On the ionosphere dataset, this criterion seems to fail. The appropriate K can be determined on some datasets by seeing where the “bend” is, where increasing K past that point only decreases cluster scatter slightly.