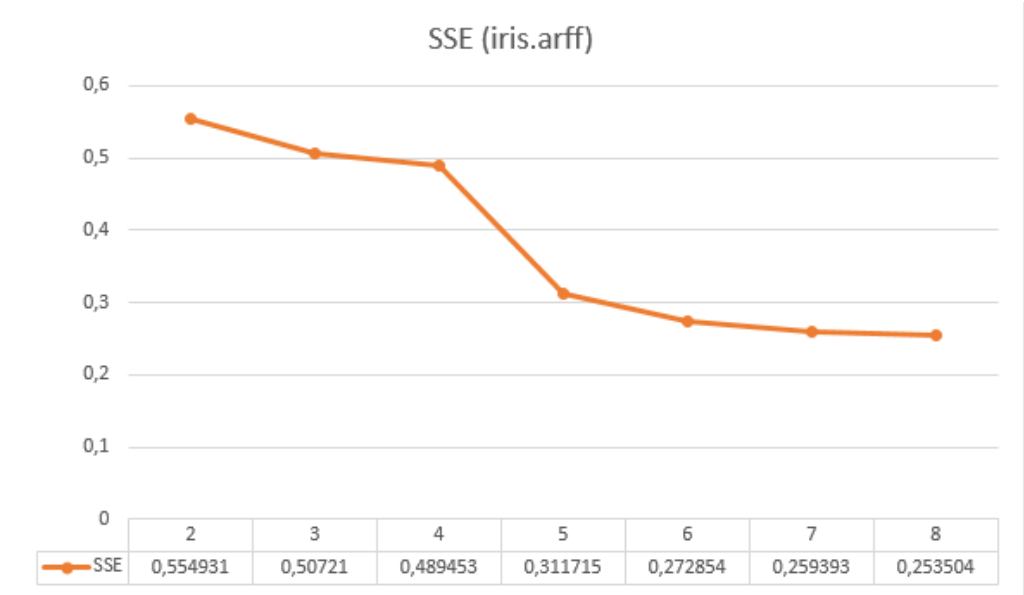
TDT4300 Datavarehus og datagruvedrift

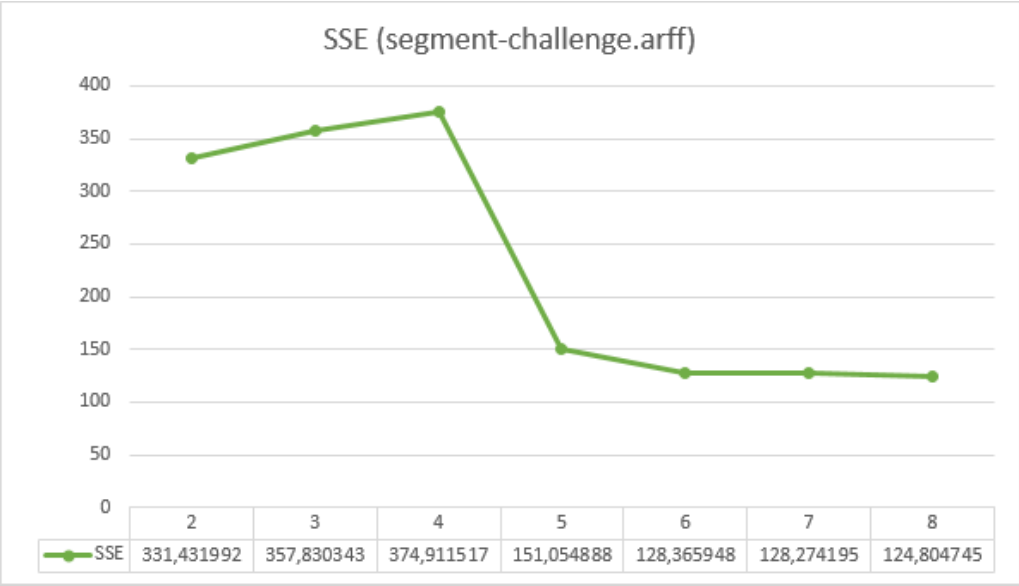
Kristian Snare & Audun Skjervold
11.04.2014

Assignment 5: Cluster Validation

2 Evaluation

We've run the code with a varying k (from k = 2 to k = 8). The graphs below show the result on SSE (y-axis) of variation in the number of clusters, k (x-axis):

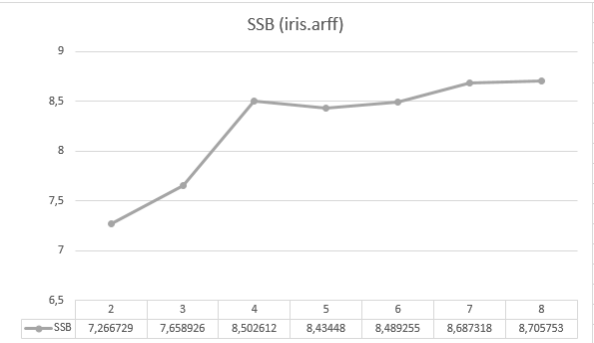


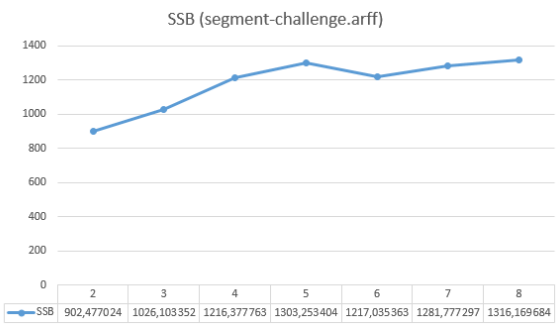


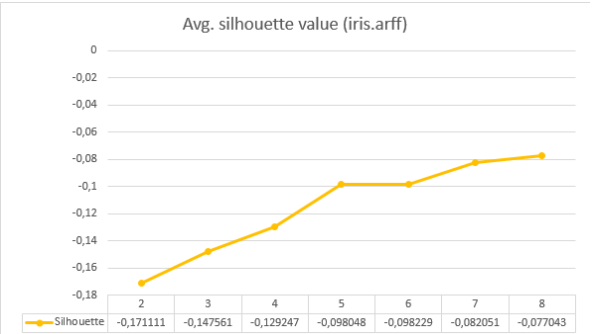
The plot of k versus SSE can show an appropriate amount of clusters via the elbow method. The method makes use of the idea that “the elbow”, an angle in the graph, shows where the adding of more clusters gives little difference to the SSE.

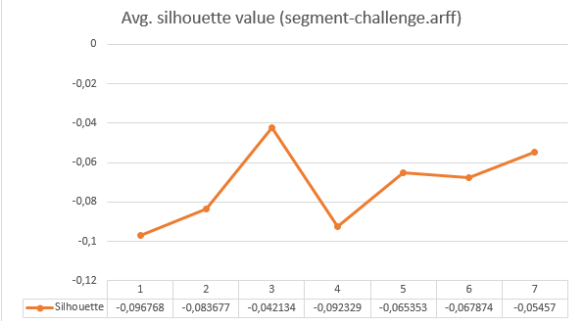
As we can see from the graphs, we get a considerable drop in SSE at 5 clusters for both data sets, after which the graph flats out somewhat.

The following four graphs show the development of the SSB (top row) and average silhouette value (bottom row) for the iris and segment-challenge datasets (left and right respectively).









We exclusively had negative silhouette values, which leads us to believe that we might have made a mistake in the implementation, but we are not sure what that would be.

All of the graphs exhibit “elbows” at 4 or 5 clusters, leading us to believe one of these are the optimal amount of clusters. Considering the considerable drop in SSE at 5 clusters, we will assume this is the best choice.

The relation between SSB, SSE and the silhouette value

- SSE measures cluster cohesion: how closely related objects are in a cluster
- SSB measures cluster separation: how distinct the clusters are from one another
- The silhouette coefficient is a measure of both cohesion and separation

SSB, sum of squares between clusters, is a prototype-based separation measure. SSB is most applicable for well separated clusters.

SSE, sum of squares error, is a prototype-based cohesion measure. SSE is most applicable for comparing of two clusterings and can be used to find an optimal number of clusters.

Silhouette coeffiecient is a measure which combines both cohesion and separation but for individual points in addition to clusters and clustering. s=b(i)-a(i)/max{a(i),b(i)} where a is cohesion and b separation. We believe that the silhouette value is an applicable measure for data sets with clusters with varying density.