CS 434: Due May 27th 2018

Scott Russell

Jacob Volkman

Justin Sherburne

**2 Non-hierarchical clustering - K-Means algorithm**

1. (25 pts) Implement the K-means algorithm. Run your K-means algorithm

with k = 2. To verify that your algorithm converges, please plot

the objective of the K-means algorithm (i.e., the SSE) as a function of the

iterations. From one run to another run, this curve may look different.

Just present the results of a typical run.

**Above is a typical example of convergence. From our testing we were able to converge anywhere from 10-20 iterations. You can see a rapid improvement over the first few iterations, and the final half are converging very slowly to the optimal SSE objective. Note that the decrease in SSE is extremely rapid for the first few iterations, these tappers off fairly quickly and only after several more iterations do we reach our convergence.**

2. (25 pts) Now apply your K-means implementation to this data with different values of k (consider values 2-10). For each value of k, please

run your algorithm 10 times, each time with a different random initialization, record the lowest SSE value achieved in these 10 repetitions for each

value of k. Plot the recorded SSE values against the changing k value.

What do you think would be a proper k value based on this curve? Please

provide justification for your choice.

**From the data we recorded it seems like there is a linear correlation with increasing K and decreasing the SSE objective. It does seem to be tapering off to a point, which seems to show that as we continue to increase the K value we gain less and less optimization. There are also a couple of “knee” cases but they seem very minor and overall the improvement with a higher K does seem to be worth the investment to the dataset.**