Assignment 4 Report - Kmeans and Clustering

Helena Bales and Natalie Suderman

CS434 - Spring 2017

May 23, 2017

Contents

1 Non-Hierachical Clustering Implementing K-means algorithm

1.1 Implementing K-means with K of 2

The results of typical run of the kmeans algorithm when k is 2. I plotted multiple runs to show the trend.

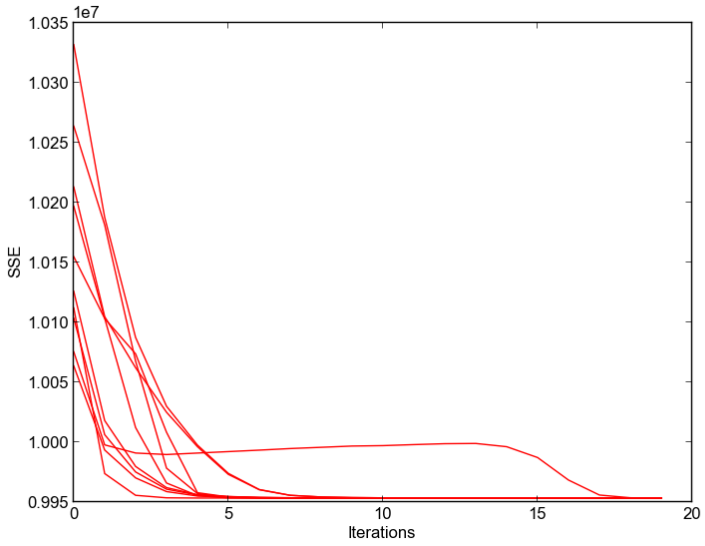

Figure 1: A typical run. SSE of Kmeans with k of 2 over number of iterations.

1.2 Apply K means to different values of K

The algorithm was tested by using different values of K (3, 4, 10, etc) and the minimum SSE from 10 runs for each K was plotted to demonstrate a trend in the size of the SSE vs number of K.

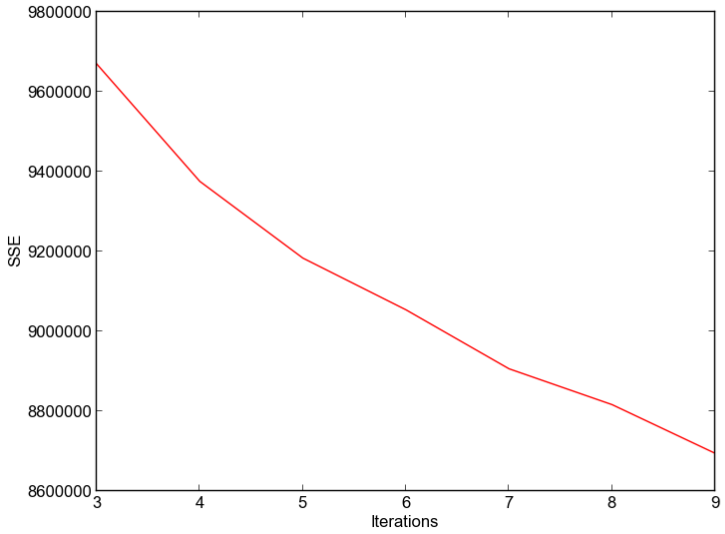

Figure 2: Min SSE found over 10 runs of the algorithm for each k tested.

2 Hierarchical Agglomerative Clustering

2.1 Compute HAC Using Single Link

similarities reported: (1,2) 1499919.0 (3,4) 1208826.0 (3,4,5) 1207976.0 (6,7) 1127700.0 (8,9) 1065419.0 (1,2,3,4,5) 1057888.0 (1,2,3,4,5,6,7) 632494.0 (1,2,3,4,5,6,7,8,9) 280324.0 (10,11) 1489698.0

2.2 Compute HAC Using Complete Link

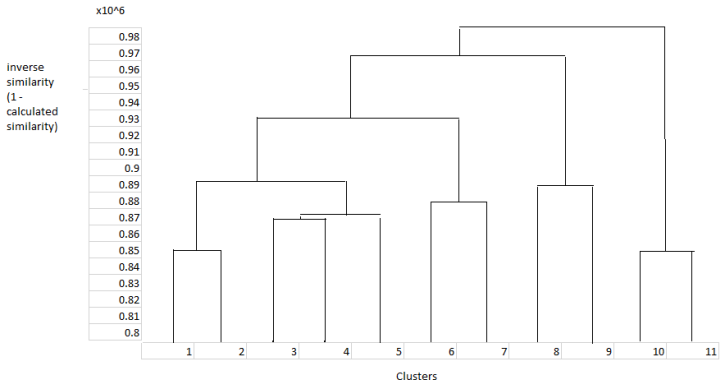

Figure 3: Dendrogram starting with 10 clusters. Difference/height is reported as an inverse of similarity. ie. Higher similarity calculated equals lower difference/height of dendrogram

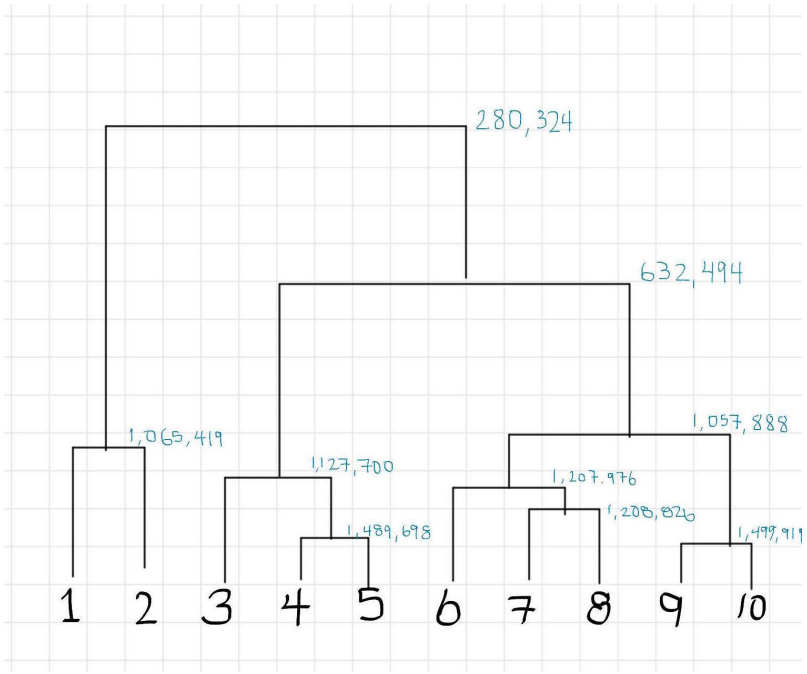

Figure 4: Dendrogram starting with 10 clusters. Reported similarities are indicated in light blue