

CS 573000 : Homework 5

Prashant Ravi | ravi18@purdue.edu

April 29, 2017

1 Question A.1

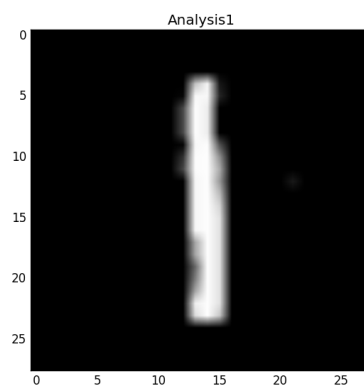


Figure 1: Visualization of digit "1"

2 Question A.2

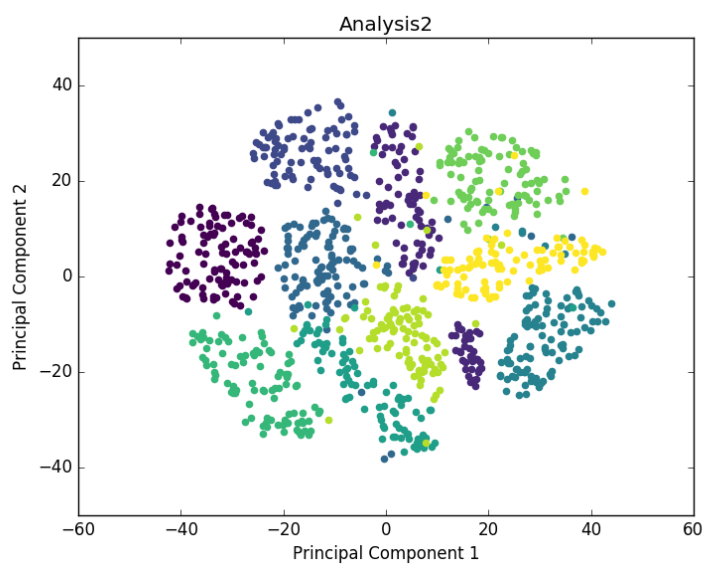


Figure 2: Visualization of random selected sample

3 Question B.1

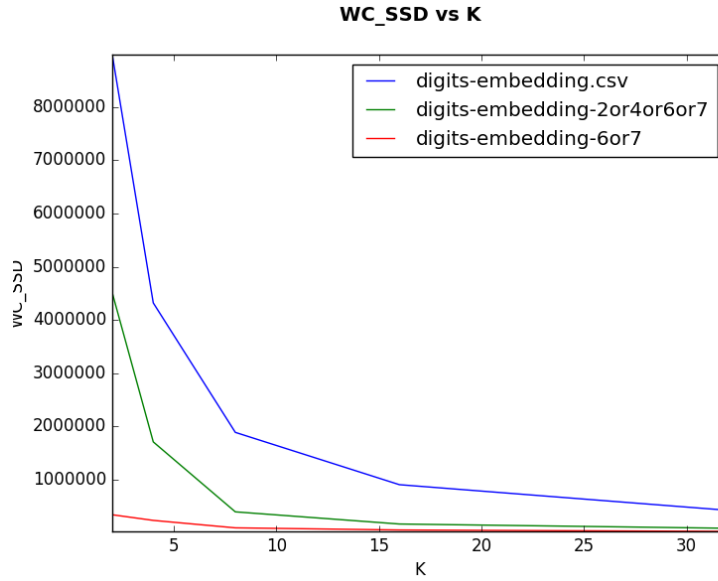


Figure 3: WC SSD vs K

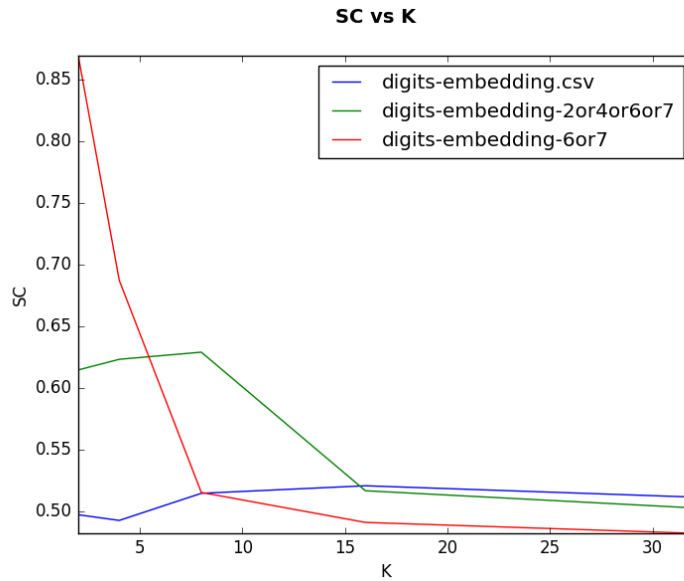


Figure 4: SC vs K

4 Question B.2

The Silhouette coefficient decreases for increasing K however, looking at the SC and WC SSD graphs we find that the selection of 8,4,2 are the best choices for the digits-embedding.csv, digits-embedding-2or4or6or7.csv, digits-embedding-2or4.csv files, respectively. The argument for this is that since each number would represent a cluster ideally, we should set to be atleast greater than or equal to the number of expected clusters. For the 2-4-6-7 cast 4 clusters would be needed, and so follows for the other two csv files. In the graphs we see that the sharpest drop in scores for Silhouette score occurs at these chosen values. In the WC SSD curve, we see a similar trend of drops at these points. Hence, they are considered good values for K.

5 Question B.3

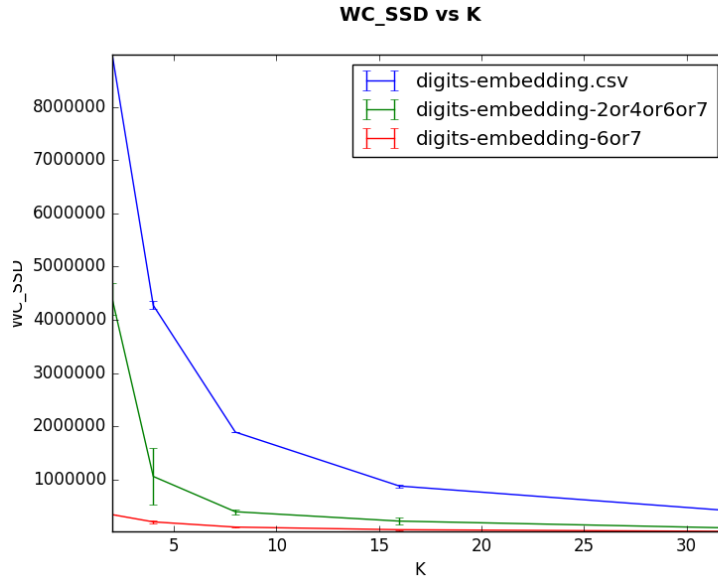


Figure 5: WC SSD vs K

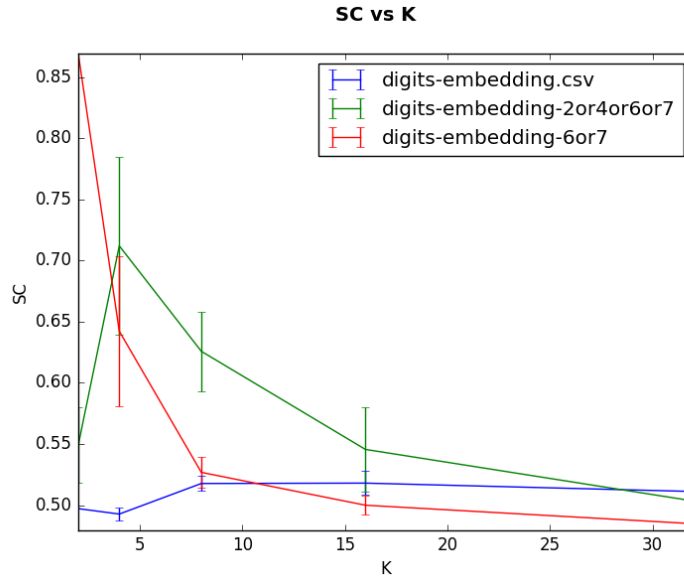


Figure 6: SC vs K

The standard error is the average standard deviation in the plots above across 10 random seed initial start initializations. The $WC - SSD$ plot doesn't show much variance, however, the SC plot shows significant variance across K values, being more sensitive for low values of K but decreasing in effect for larger values. In addition, the digits-embedding.csv is most insensitive to start conditions.

6 Question B.4

NMI 0.347 filename -> digits-embedding.csv K- > [8] NMI 0.455 filename -> digits-embedding-2or4or6or7 K- > [4] NMI 0.491 filename -> digits-embedding-6or7 K- > [2]

7 Question C1

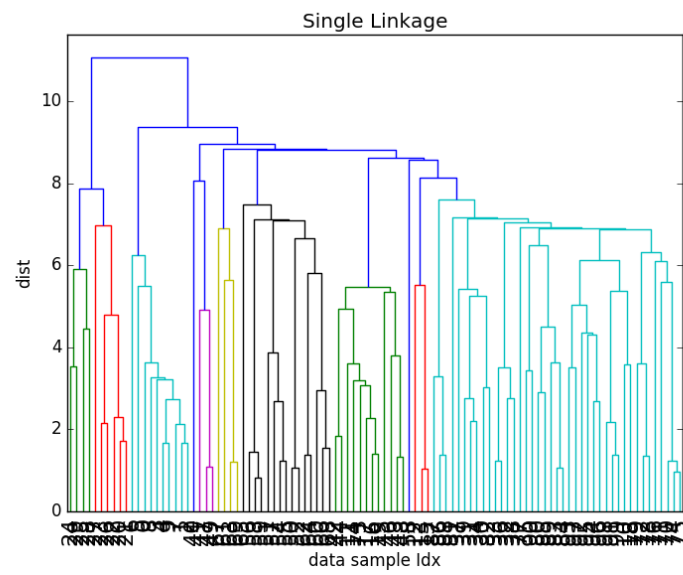


Figure 7: Single Linkage

8 Question C2

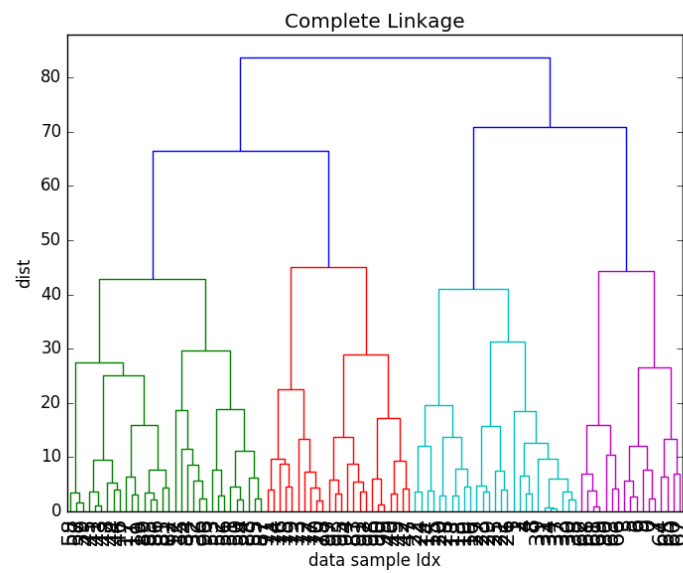


Figure 8: Complete Linkage

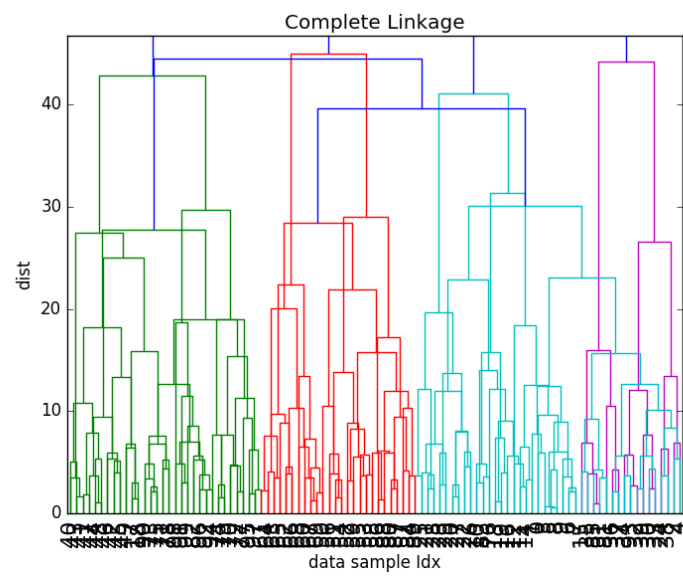


Figure 9: Average Linkage

9 Question C3