# CS573_assignment5_XiaofengOu

Xiaofeng Ou

April 2019

## 1 Exploration


Figure 1: Raw digits view

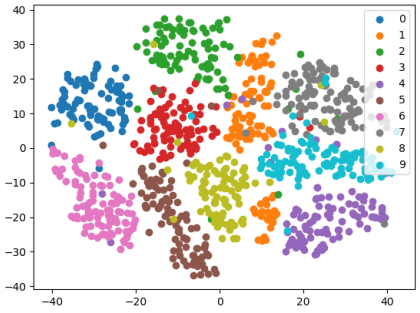


Figure 2: Embedding digits view

# 2 K-means Clustering

## 2.1 Code

**Output** :

```
WC-SSD: 1433531.47
SC: 0.71
NMI: 0.36
```

## 2.2 Analysis

### 2.2.1

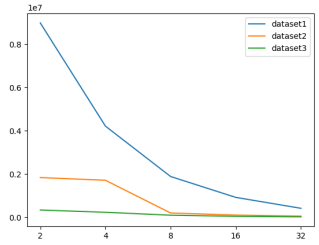


Figure 3: Within-cluster sum of squared distances for datasets 1 2 3

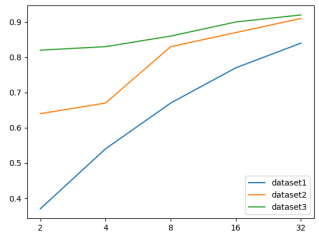


Figure 4: Silhouette coefficient for datasets 1 2 3

### 2.2.2

Choose $K = 8, 8, 2$ for datasets $1, 2, 3$ respectively.

1. For dataset 1, 8 is the elbow point of WC-SSD.
2. For dataset 2, 8 is also the elbow point of WC-SSD while 8 has relatively high SC score.

3. For dataset 3, there is no apparent elbow point and the graph is relatively flat. So choose $K = 2$ for preventing overfitting.

### 2.2.3

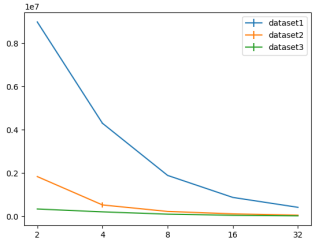


Figure 5: WC-SSD with different random seeds

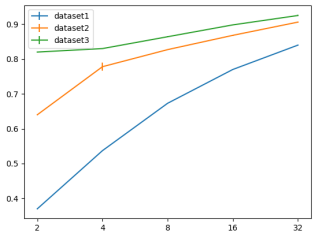


Figure 6: SC with different random seeds

K-means is not sensitive to initial starting position.

### 2.2.4

**NMI output** :

```
NMI of dataset1 with K=8: 0.35
NMI of dataset2 with K=8: 0.33
NMI of dataset3 with K=2: 0.49
```

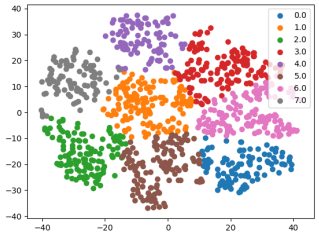


Figure 7: Visualize for dataset1 with K=8

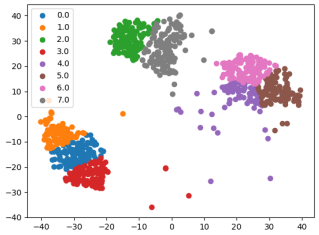


Figure 8: Visualize for dataset2 with K=8

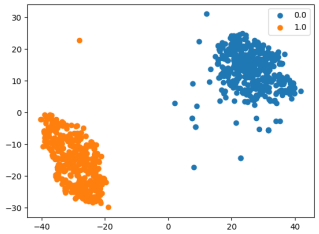


Figure 9: Visualize for dataset3 with K=2

# 3 Hierarchical Clustering

## 3.4

1. Single linkage: $K = 8$
2. Complete linkage: $K = 8$
3. Average linkage: $K = 8$

Reason: 8 is the elbow point for each linkage. The result is the same as that in kmeans.

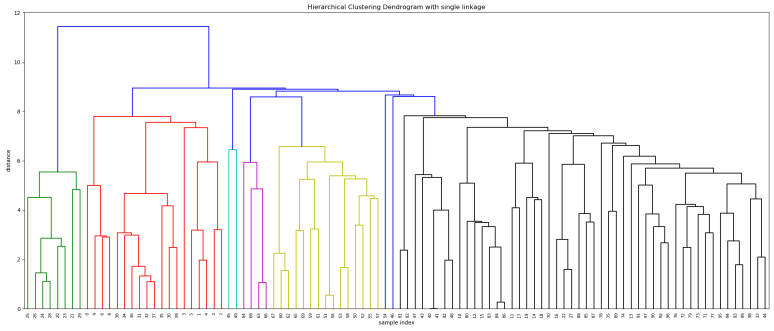


Figure 10: Dendrogram with single linkage

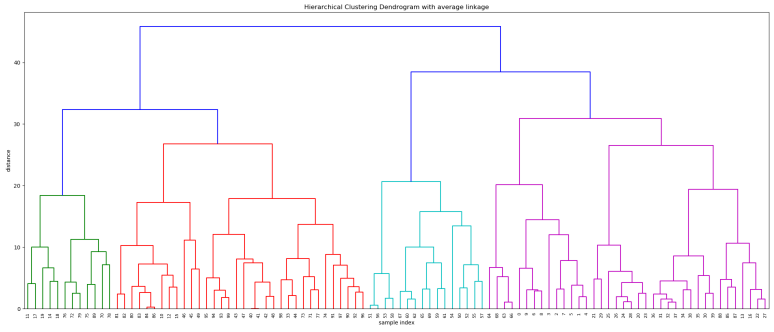


Figure 11: Dendrogram with average linkage

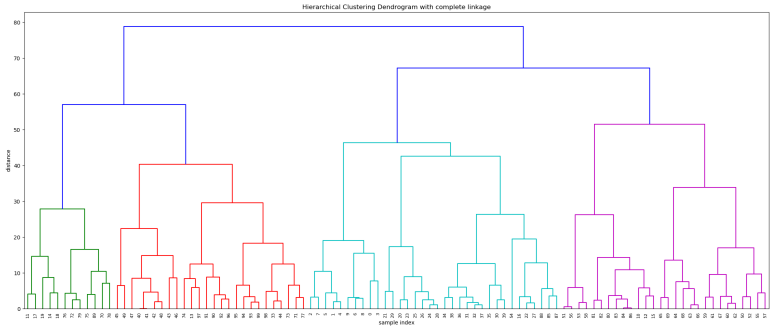


Figure 12: Dendrogram with complete linkage

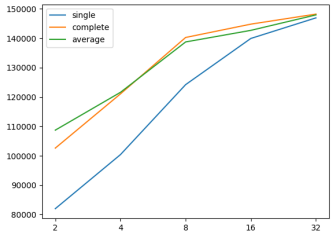


Figure 13: WC-SSD

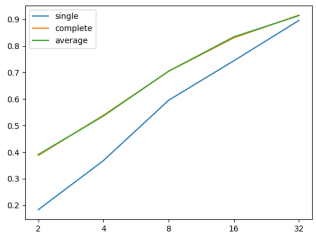


Figure 14: SC score

## 3.5

**Outputs** :

```
NMI of single linkage with K=8: 0.32
NMI of complete linkage with K=8: 0.36
NMI of average linkage with K=8: 0.34
```

They are approximately the same as the value in kmeans.