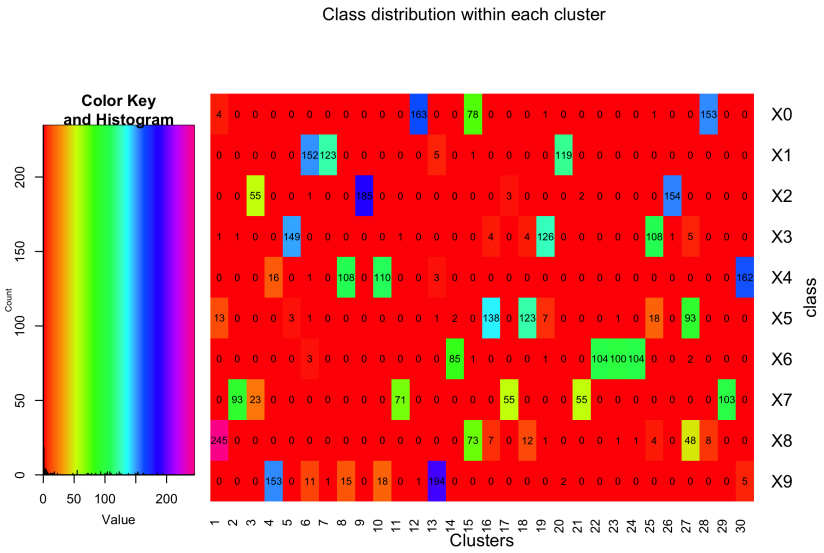
1 K - means clustering

Clustering is the task of grouping a set of data points in such a way that data points in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). The clustering is done by minimizing the sum of square distance between each data point and their correnspondingcluster centroid.

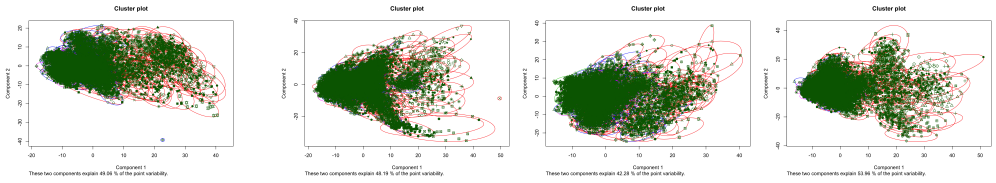
The number of clusters needed is determined by the user. It usually depends on how the data points is distributed. Choosing a large number of cluster will result in a reduced squared distance, and indeally zero as each data point gets it own cluster. The optimal choice of k will consist of having a balance between having the data compressed as possible, and still retain a decent amount of accuracy. More fomally is the goal to partition the data into k clusters in order to minimze the total within cluster sum of squares. For which the elbow method might become handy.

Elbow

confusion table



(a) Class distribution within each cluster

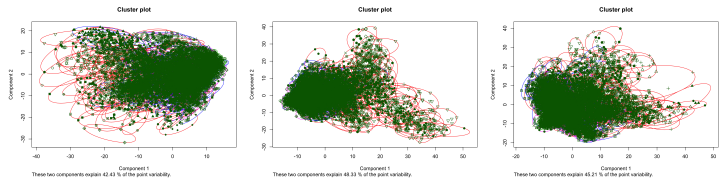


(b) Clustering data consiting of digit 0

(c) Clustering data consiting of digit 1

(d) Clustering data consiting of digit 2

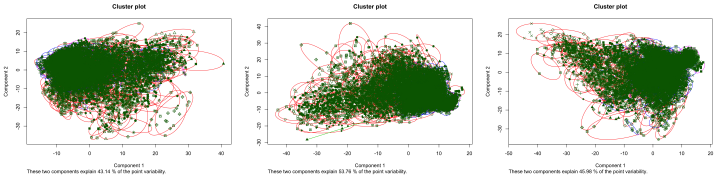
(e) Clustering data consiting of digit 3



(f) Clustering data consiting of digit 4

(g) Clustering data consiting of digit 5

(h) Clustering data consiting of digit 6



(i) Clustering data consiting of digit 7

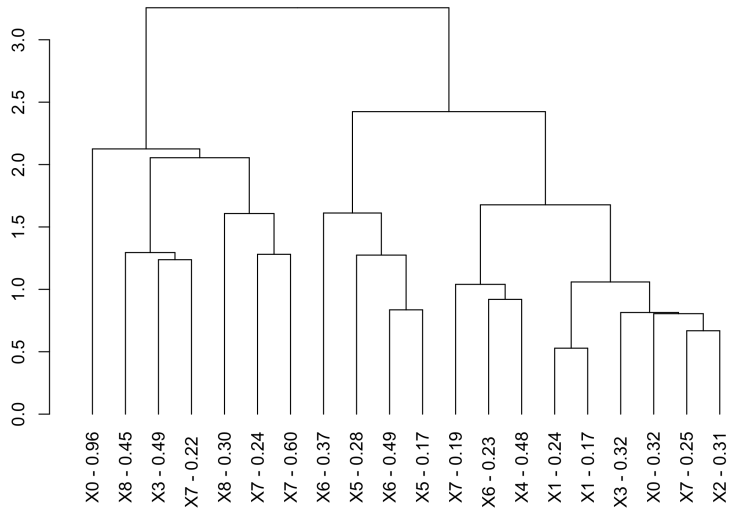
(j) Clustering data consiting of digit 8

(k) Clustering data consiting of digit 9

Figur 1: This and that

some text

some text



(a) Dendrogram

Insert some text