Part A. Document Clustering

Question 1

Part 1

Ans. As per the problem we would like to partition a collection of documents $\{d_1, d_2, ..., d_N\}$ into K clusters. Since this is an unsupervised learning problem the document clusters are not given to us.
Each document d_n is made up of some text and we assume that the words in the documents come from a dictionary denoted by $\mathcal{A}$.

Generative Model - The following hypothetical generative story is for generating the collection of our documents:
For each document d_n

1. Toss a K-face dice(with parameter $\boldsymbol{\varphi}$) to chose the face (i.e. the cluster) k to which d_n belongs to
2. For eachword placeholder in the document d_n generate the word by tossing the dice (with parameter μ_k) corresponding to the face k

Model Parameters - The parameters of the model are:

1. The cluster proportion $\mathcal{A}$ - A probability vector of size K where, $\sum_{k=1}^{K} \varphi_k = 1$
2. The word proportion μ_k which corresponds to the kth face of the dice where, $\sum_{w \in \mathcal{A}} \mu_{k,w} = 1$; we have K such word proportion vectors each corresponding to a different cluster.

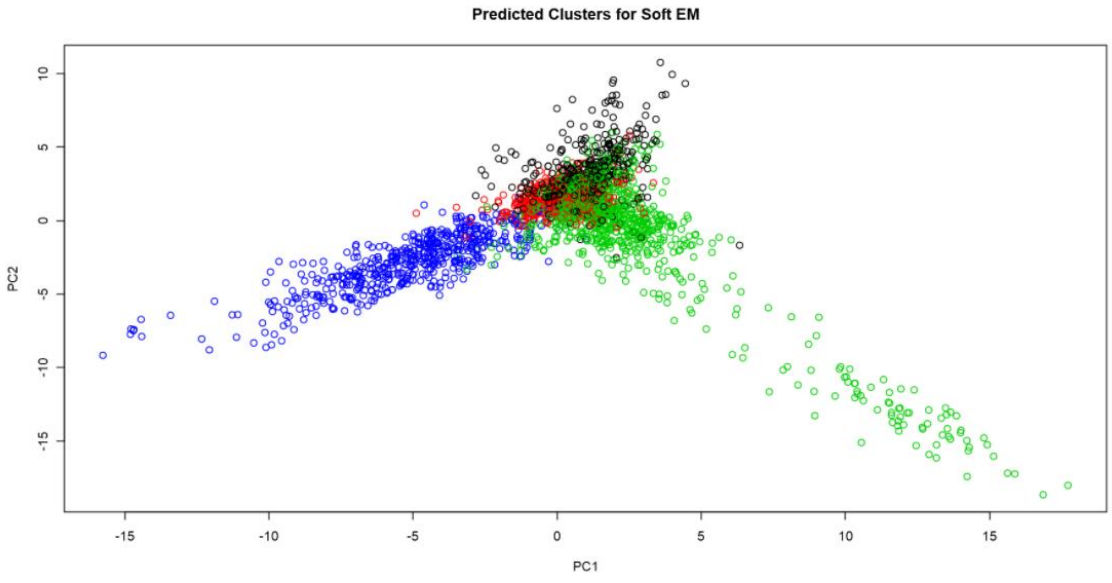
The probability of generating a pair of a document and its cluster (k,d) is:

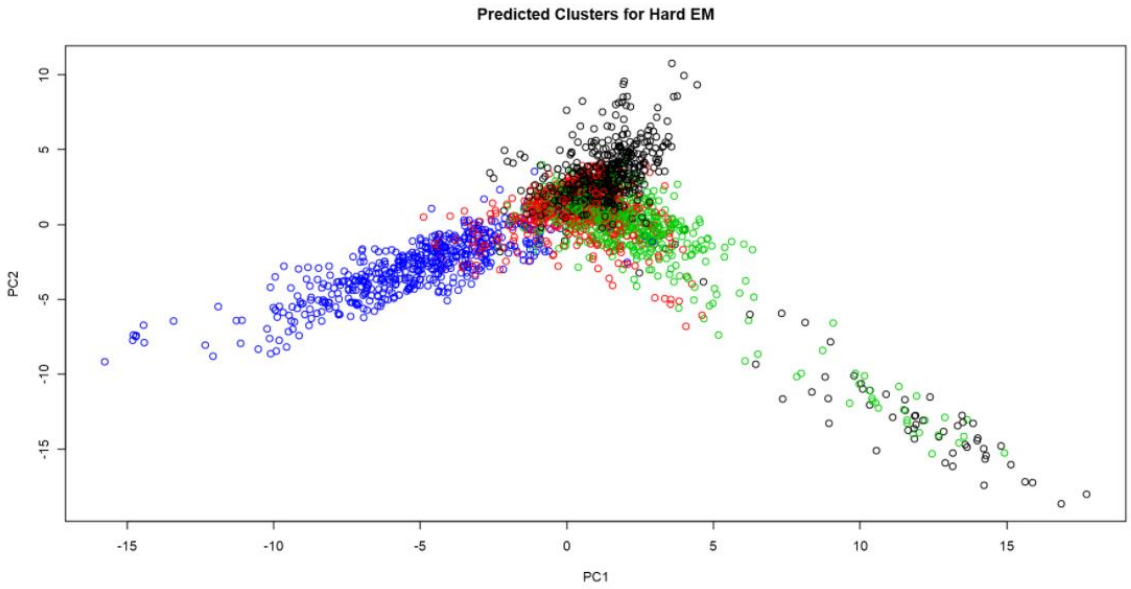
$$p(k,d) = p(k)p(d|k) = \varphi_k \prod_{w \in d} \mu_{k,w}$$
$$= \varphi_k \prod_{w \in \mathcal{A}} \mu_{k,w}^{c(w,d)}$$

where c(w,d) is simply the number of accurences of the word w in document d.

Part 4

The predicted clusters for hard and soft EM algorithm vary since in hard EM the datapoints belong to one and only one cluster, while in soft EM the datapoints are assigned to clusters based on the probability distribution. In hard EM the cluster with the maximum probability is assigned the probability 1 whereas the other clusters are given a probability of zero.

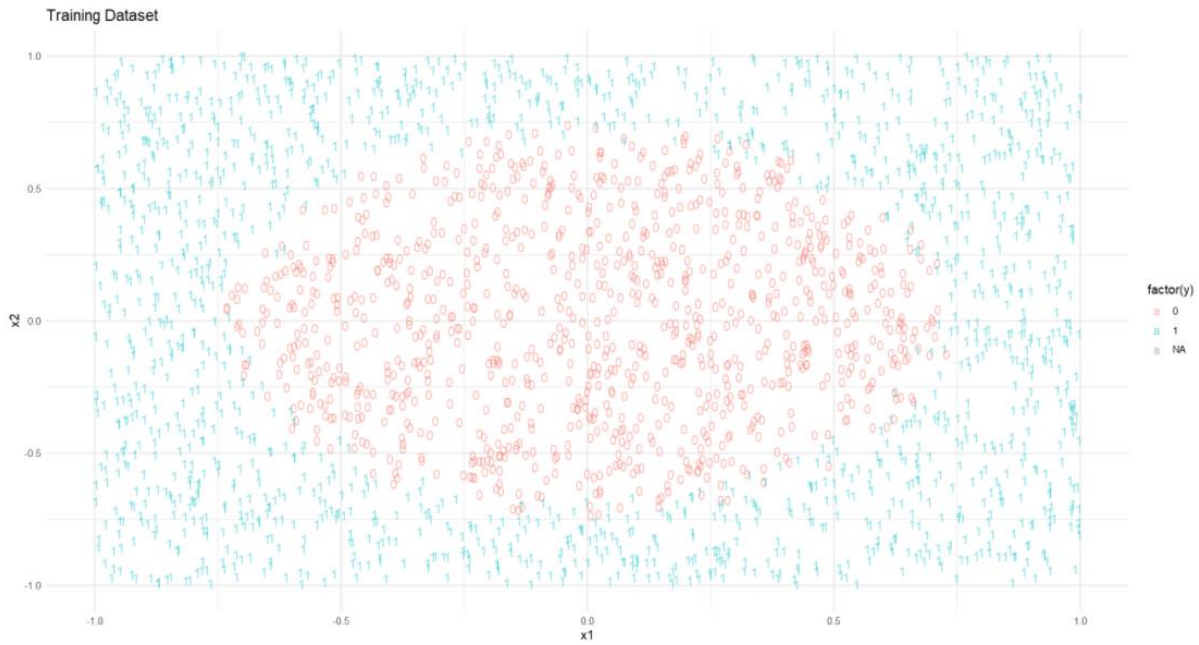




Part B. Neural Network vs. Perceptron

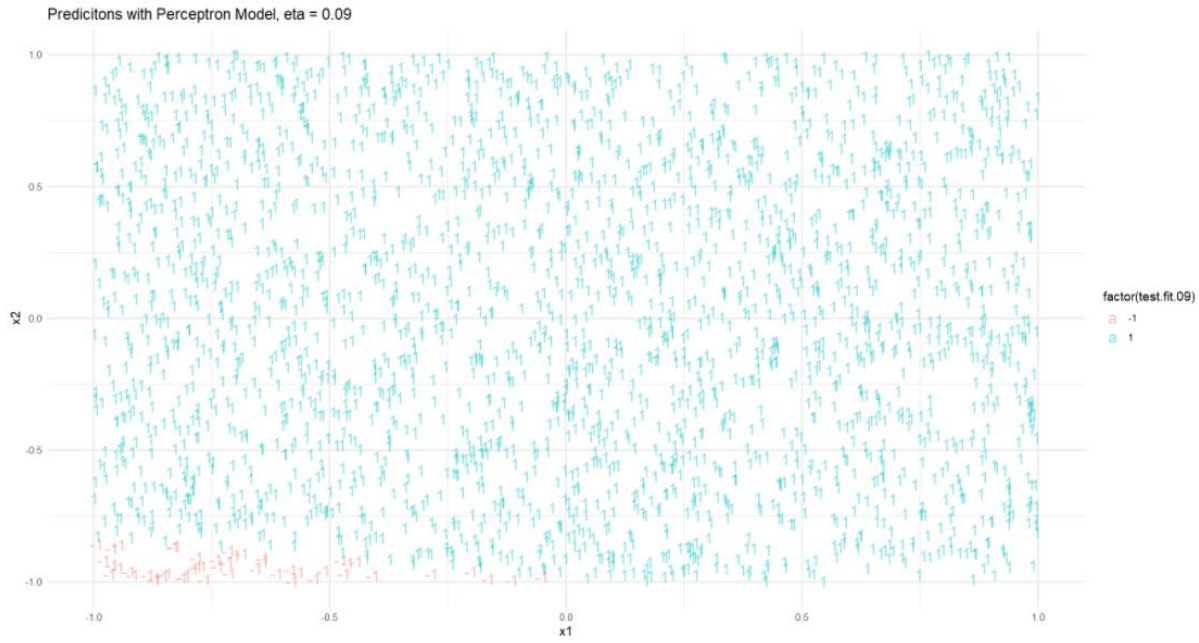
Question 2

Part 1

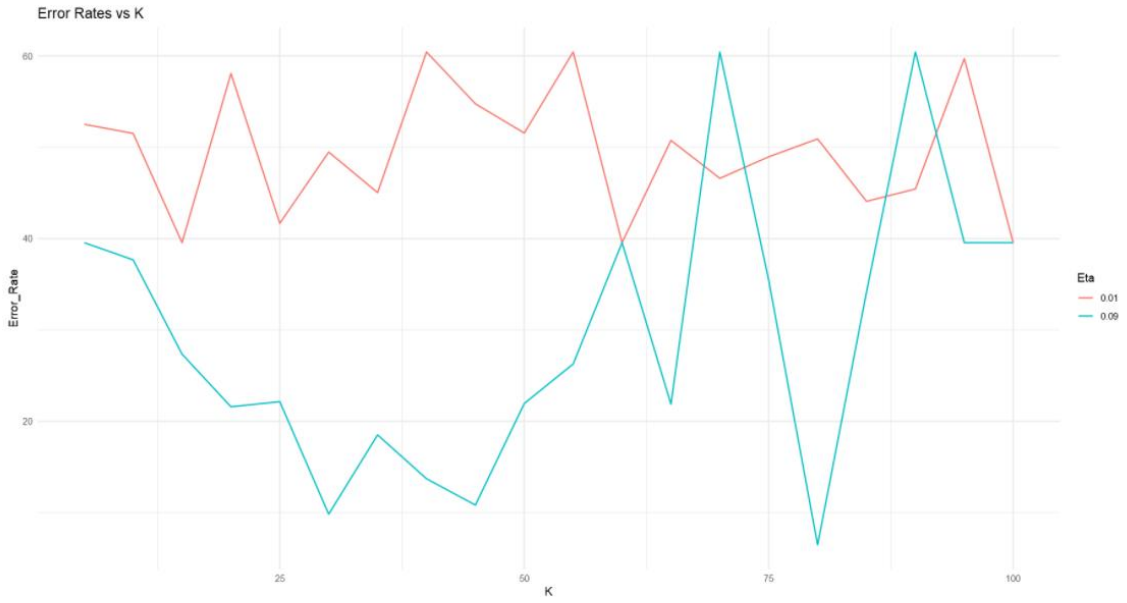


Part 2

The best value of eta was found to be 0.09 since it had a lower test error. The corresponding plot with eta = 0.09 has been attached.

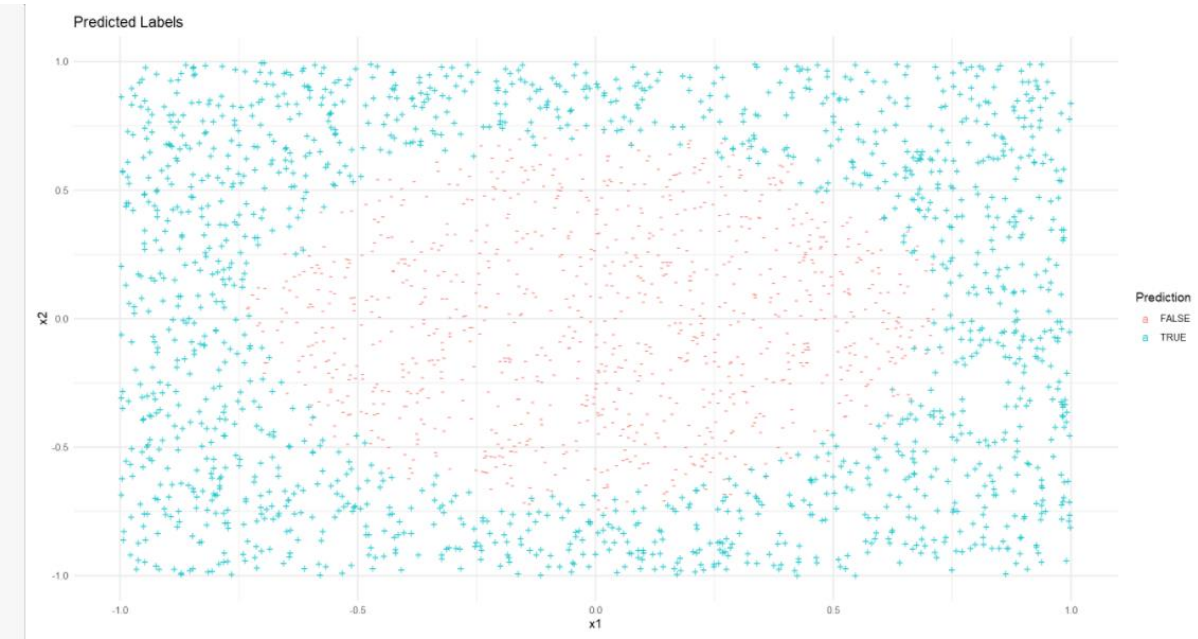


Part 3



Best value of eta = 0.09 and corresponding K = 80 with Error-Rate = 6.48

The predicted labels with eta = 0.09 and K = 80 are plotted below.



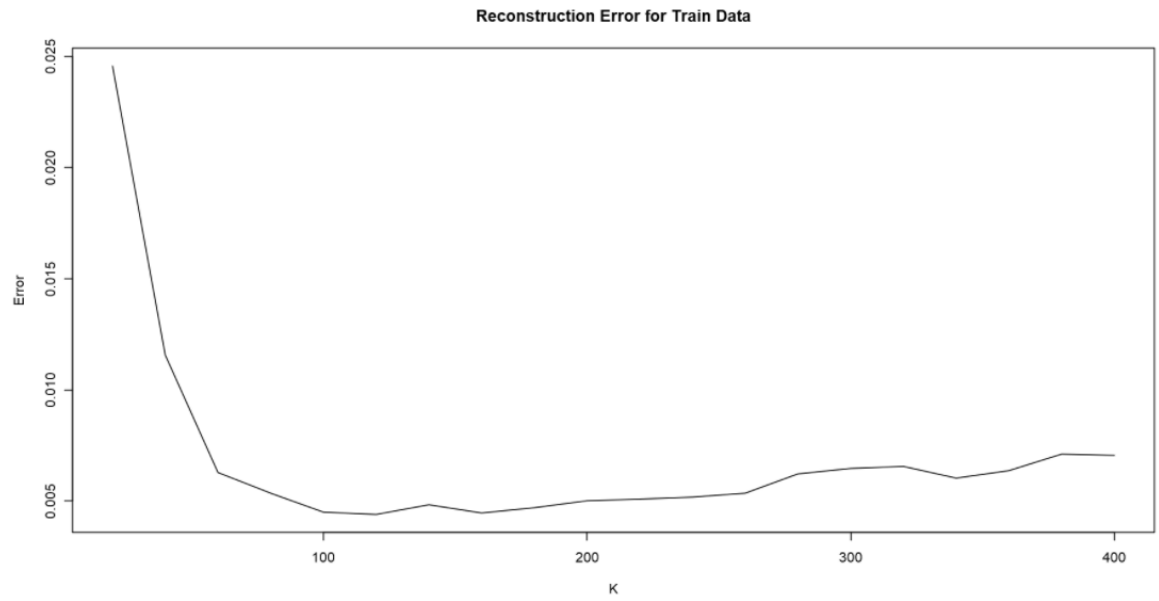
Part 4

As we can visualize in the above plots the predicted labels are much better for the neural network than the perceptron model. This is because a perceptron model fits a linear model to the given data whereas a neural network provides a more accurate classification. Also the perceptron model is sensitive to the initial values given whereas such is not the case with neural network.

Part C. Self-Taught Learning

Question 3

Part 3



We can see that the reconstruction error decreases significantly as the number of units in the middle layer(K) increases and then gradually begins to rise after a certain point.