ASSIGNMENT 2

FIT5201 MACHINE LEARNING: SEMESTER 1 2020

NAME: PARUL
STUDENT ID: 29507960

Question 1 [EM for Document Clustering, 40 Marks]

I. Derive Expectation and Maximization steps of the hard-EM algorithm for Document Clustering.

The Q Function. Let's look at the Q function, which will be the basis of our EM Algorithm:

$$Q(\theta, \theta old) := \sum_{n=1}^{N} \sum_{k=1}^{K} p(z_{n,k} = 1 \mid d_n, \theta^{old}) \ln p(z_{n,k} = 1, d_n \mid \theta)$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} p(z_{n,k} = 1 \mid d_n, \theta^{old}) (\ln \phi_k + \sum_{\varphi \in A} c(w, dn) \ln \mu_{k,w})$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{n,k}) (\ln \phi_k + \sum_{w \in A} c(w, d_n) \ln \mu_{k,w})$$

Where $\theta := (\phi, \mu_1, \ldots\ldots, \mu_k)$ is model parameters

and $\gamma(z_{n,k}) := p(z_{n,k} = 1, d_n \mid \theta old)$.

To maximise the Q function, we can form the LaGrange and set the derivatives to zero:

The mixing components: $\phi_k = N_k/N$ where $N_k := \sum_{n=1}^{N} \gamma(z_{n,k})$

The mean of cluster is calculated as follows:

$$\mu_{k,w} = \frac{\sum_{n=1}^{N} \gamma(z_{n,k})\, c(w, d_n)}{\sum_{w' \in A} \sum_{n=1}^{N} \gamma(z_{n,k})\, c(w', d_n)}$$

EM Algorithm:

Initial setting for the parameters $\theta_{old} = (\phi^{old}, \mu_1^{old}, \ldots, \mu_k^{old})$

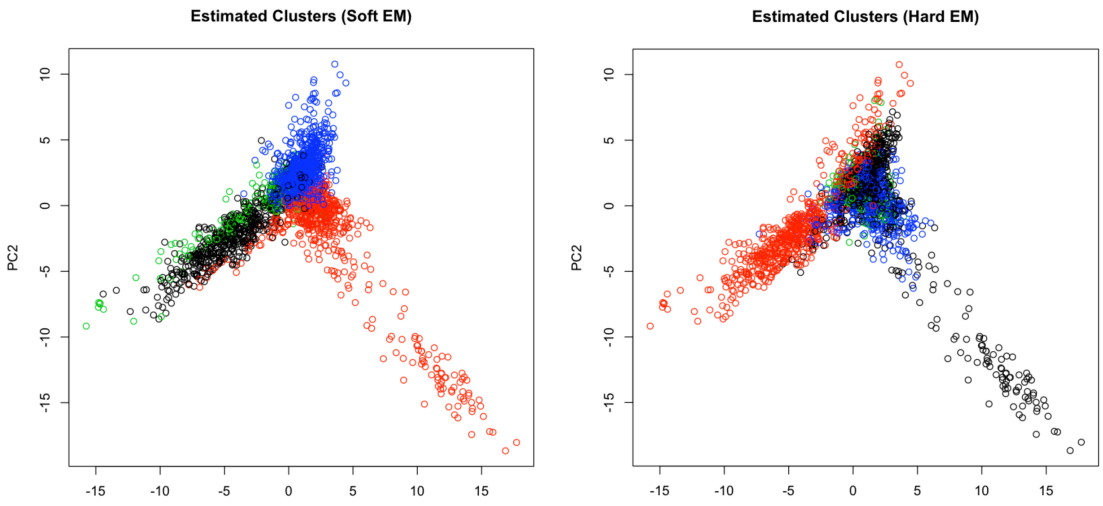
While the convergence is not met:

E step: Set $Z_* \leftarrow \text{argmax}_z\, p(Z \mid X, \theta^{old})$

M Step: Set $\theta_{new} \leftarrow \text{argmax}_\theta \ln p(X, Z_* \mid \theta)$

$\theta_{old} \leftarrow \theta_{new}$

III. Perform a PCA on the clustering that you get based on the hard-EM and soft-EM. Then, visualize the obtained clusters with different colours where x and y axes are the first two principal components. Report how and why the hard and soft-EM are different, based on your plots in the report.



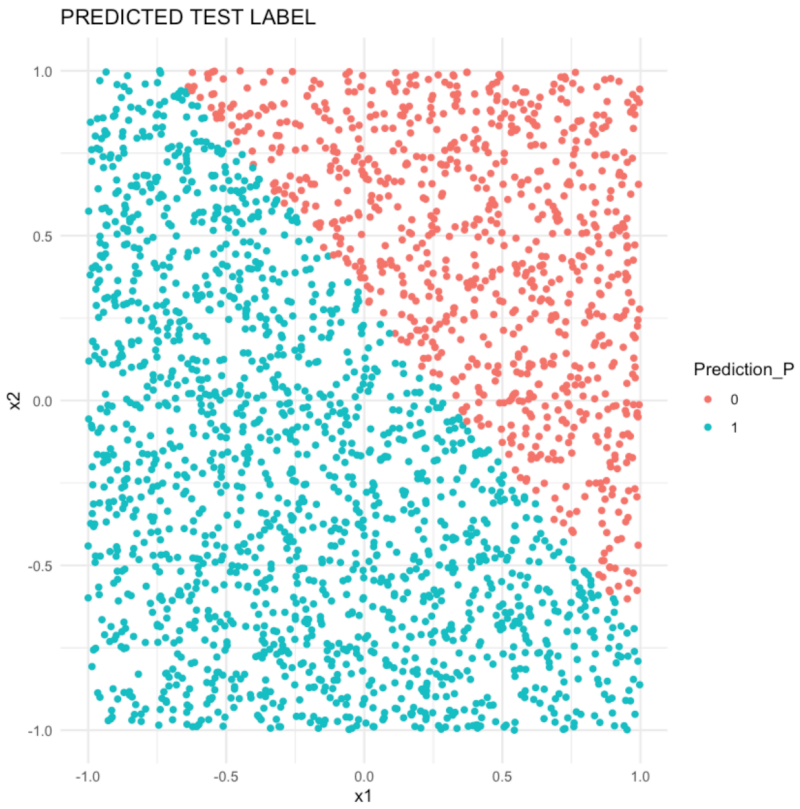
In soft clustering, one data point is assigned to more than one cluster having diff probabilities for each cluster whereas in Hard clustering data points strictly belong to only one class of cluster. From the graphs above you can see, in graph one there are some points belonging to multiple classes and the clusters are no separated, whereas the graph for hard clustering has data points belonging to only one cluster.

Part B. Neural Network vs. Perceptron

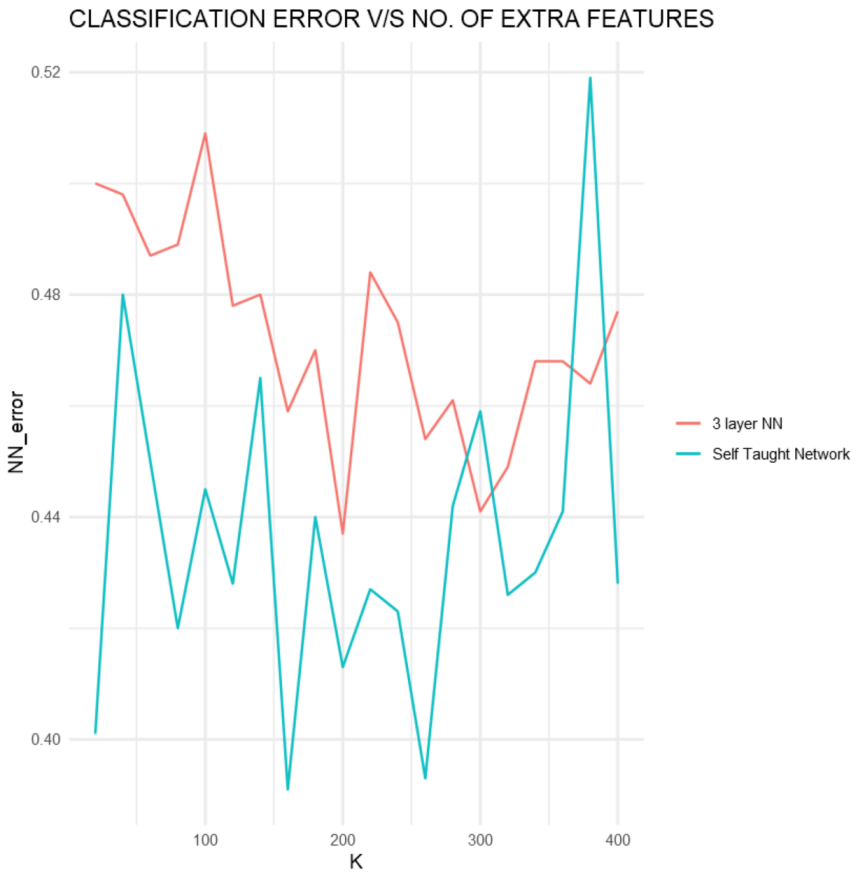
Question 2 [Neural Network's Decision Boundary, 30 Marks]

I. Plot of training data with classes are marked with different colours.

II. Plot of test data while the points are colored with their estimated class labels using the best Perceptrone model.



III. Plot of error for μ 0.01 and 0.09 vs K (one line for μ 0.01 and another line for μ 0.09 in a plot) and attach it to your PDF report. Based on this plot, find the best combination of K and μ and the corresponding model, then plot the test data while the points are colored with their estimated class labels using the best model that you have selected.



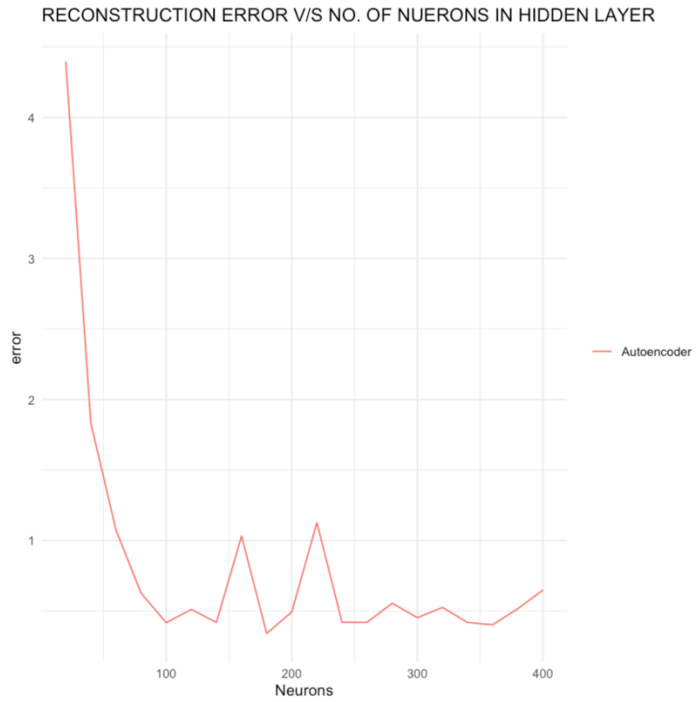
IV. Explain the reason(s) responsible for such difference between perceptron and a 3-layer NN by comparing the plots.

Perceptron is a single layer Neural Network that works as a linear binary classifier. Perceptron trained on the labelled data to predict clusters separated data linearly and gives huge error in predicting lables, and are different from the actual labels in the cluster. Whereas has 3-layers neural networks trained on labelled and unlabelled data produces much better results by separating data non-linearly in comparison to the perceptron.

Part C. Self-Taught Learning

Question 3 [Self Taught Neural Network Learning, 30 Marks]

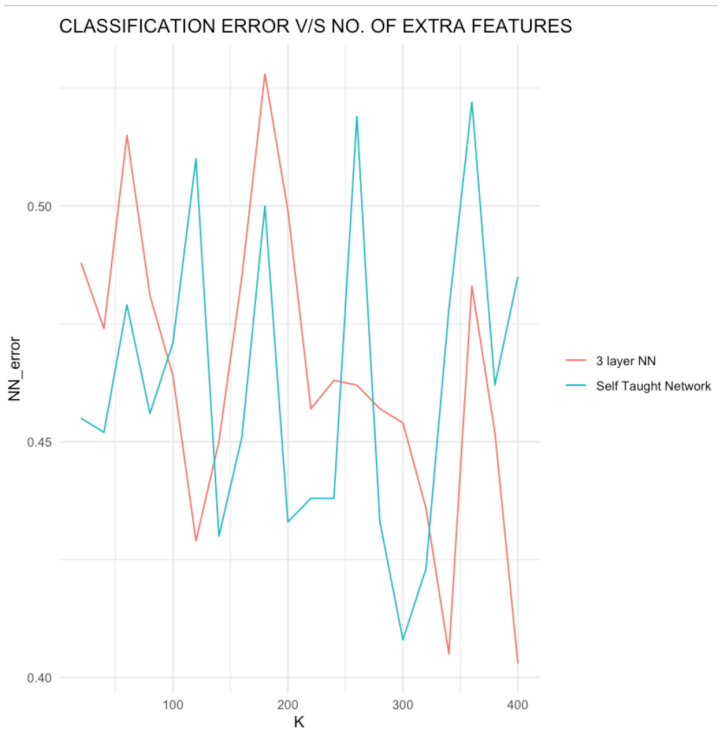
III. Plot these values where the x-axis is the number of units in the middle layer and the y-axis is the reconstruction error. Explain your findings based on the plot.



Autoencoder is an unsupervised neural network trained trained on inputs to produce equal number of targets as inputs. These targets are reconstructed inputs. Autoencoder first reduce the dimensions of input and then reconstruct it. In contrast, anomalies are not reconstructed well and have a high amount of reconstruction error, so in the process of encoding and decoding the instances, the anomalies are discovered.

From plot above, we can say that the reconstruction error of the autoencoder decreases with increase in the number of neurons (units) until a certain point, after that it fluctuates within a small range of values of error as we increase the number of neurons (units).

VI. Plot of error rates for the 3-layer neural networks from Step IV and the augmented self-taught networks from Step V, while the x- axis is the number of extra features and y-axis is the classification error. Explain how the performance of the 3-layer neural networks and the augmented self-taught networks is different and why they are different or why they are not different, based on the plot.



Self-taught neural network is trained on additional features extracted from autoencoder whereas the normal Neural network is trained on the given set of features without any additional features.

Thus, the Self-Taught neural network converges first in comparison with the neural network.