SMAI
Assignment 9
Prakash Nath Jha
2018201013

Question 1:

Part 1:

3-Layer Autoencoder

Layers and thers respective size = N, 17 , N

Where N = number or dimensions of the training data and size of hidden layer is 17 which is derived from PCA variance loss calculation earlier.

Activation functions used = “linear”, “linear”

bottleneck_layer_neurons = 17
bottleneck_layer_index = 1
learning_rate = 0.01
epochs = 1000

R2 Score of reconstructed training data: 0.890843
R2 Score of reconstructed validation data: 0.707980

Deep Autoencoder

Layers and thers respective size = N, 25, 17, 25 , N

Where N = number or dimensions of the training data and size of hidden layer is 17 which is derived from PCA variance loss calculation earlier.

Activation functions used = “relu” , “relu” , ”relu” ,“relu”

bottleneck_layer_neurons = 17
bottleneck_layer_index = 2
learning_rate = 0.001
epochs = 1000

R2 Score of reconstructed training data: 0.7931389
R2 Score of reconstructed validation data: 0.5135365

We can observe that with non linear activation functions R2 score decreases drastically because we have used mean square error as the loss function keeping in mind that the dataset contains more number of non catergoical features and thus the use of non linear function is not an ideal choice

Part 2:

Kmeans using 3-Layer Autoencoder

Layers and thers respective size = N, 17 , N

Where N = number or dimensions of the training data and size of hidden layer is 17 which is derived from PCA variance loss calculation earlier.

Activation functions used = “linear”, “linear”

bottleneck_layer_neurons = 17
bottleneck_layer_index = 1
learning_rate = 0.01
epochs = 1000

R2 Score of reconstructed training data: 0.895049
R2 Score of reconstructed validation data: 0.746920

Custom Implementation of K Means Clustering

Purity of cluster 1 is: 0.915273132664437
Purity of cluster 0 is: 0.9932815452445938
Purity of cluster 4 is: 0.43927408096789206
Purity of cluster 2 is: 0.6919682259488085
Purity of cluster 3 is: 0.6363636363636364

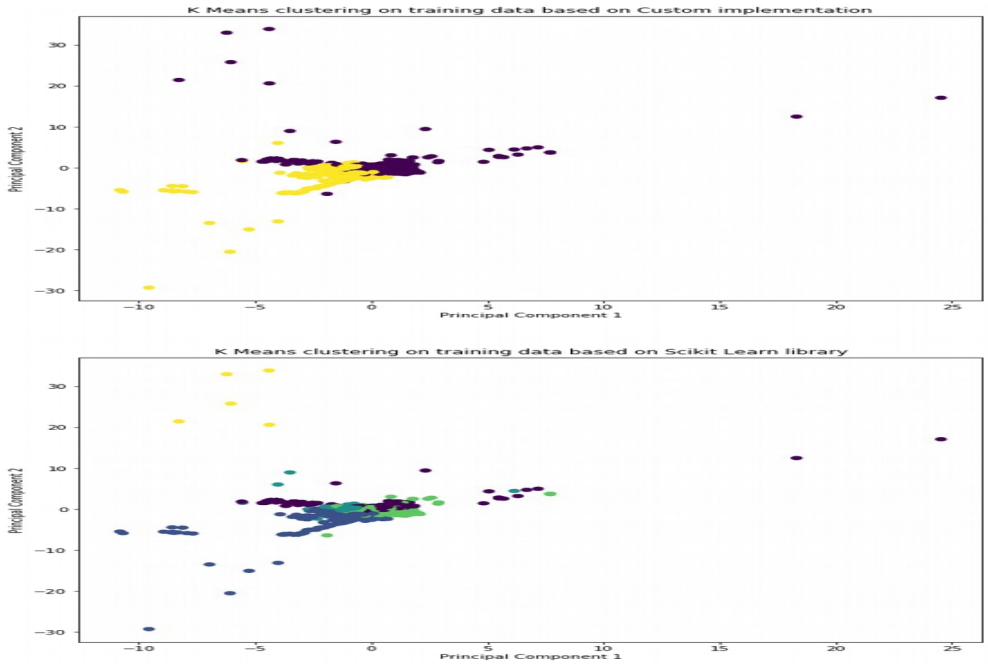
Performance on train data
Accuracy: 0.8344953708995314

Performance on validation data
Accuracy: 0.5298666666666667

Scikit Learn Library Implementation of K Means Clustering

Performance on train data
Accuracy: 0.8343810721225283

Performance on validation data
Accuracy: 0.5472



Kmeans using Deep Autoencoder

Layers and thers respective size = N , 25 , 17 , 25 , N

Activation functions used = “relu”, “relu”,“relu”, “relu”

bottleneck_layer_neurons = 17
bottleneck_layer_index = 2
learning_rate = 0.001
epochs = 1000

R2 Score of reconstructed training data: 0.808496
R2 Score of reconstructed validation data: 0.554741

Custom Implementation of K Means Clustering

Purity of cluster 2 is: 0.9423584076353133
Purity of cluster 1 is: 0.9829663481512256
Purity of cluster 3 is: 0.5230427540255413
Purity of cluster 4 is: 0.6976360637713029
Purity of cluster 0 is: 0.5449591280653951

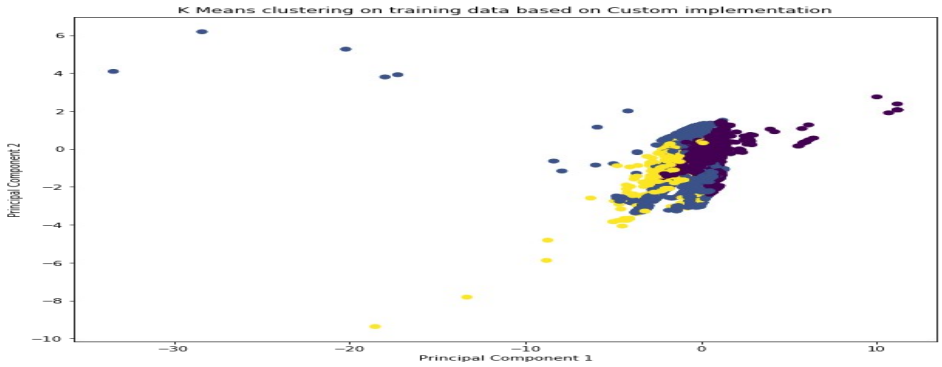
Performance on train data
Accuracy: 0.8599268487827181

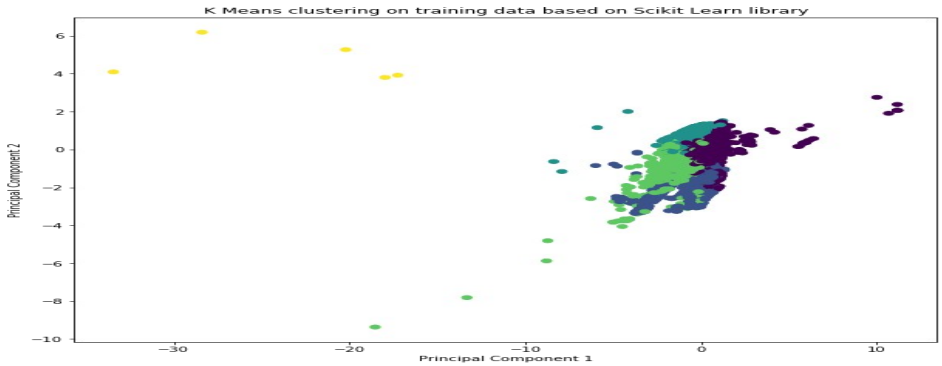
Performance on validation data
Accuracy: 0.8017333333333333

Scikit Learn Library Implementation of K Means Clustering

Performance on train data
Accuracy: 0.8515830380614927

Performance on validation data
Accuracy: 0.8457333333333333





Part 3:

Gaussian Mixture Model using 3 layer Autoencoder

bottleneck_layer_neurons = 17
no_of_clusters = 5
bottleneck_layer_index = 1
learning_rate = 0.01
epochs = 1000
act_func = ["linear","linear"]

R2 train data: 0.8856887628614438
R2 validation data: 0.6836509718871968

Purity of cluster 3 is: 0.9643570440927945
Purity of cluster 2 is: 0.9878483134297088
Purity of cluster 4 is: 0.44210032817627753
Purity of cluster 1 is: 0.7244973938942666
Purity of cluster 0 is: 0.6233576642335766

Accuracy on train dataset: 0.7969482226540177
Accuracy on validation dataset: 0.7121333333333333

Part 4:

Hierarical Clustering using 3 layer Autoencoder

bottleneck_layer_neurons = 5 (Because of memory constraint)
no_of_clusters = 5
bottleneck_layer_index = 1
learning_rate = 0.01
epochs = 1000
act_func = ["linear","linear"]

R2 train data: 0.6038968088063424
R2 validation data: 0.08403993528611232

Purity of cluster 0 is: 0.901029213015906
Purity of cluster 1 is: 0.9860369609856263
Purity of cluster 2 is: 0.44990458015267176
Purity of cluster 3 is: 0.6321585903083701
Purity of cluster 4 is: 1.0

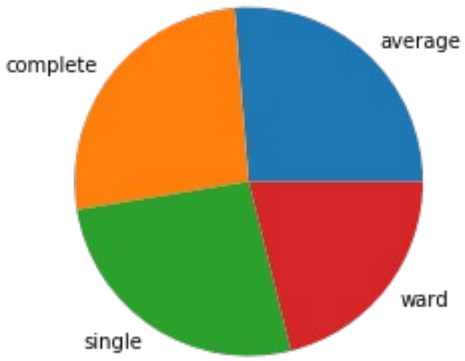
Linkage_Type Accuracy (On train data)

ward = 0.801749
single = 0.999543
average = 0.999314
complete = 0.999086

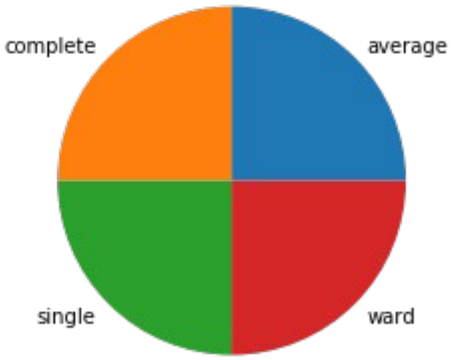
Linkage_Type Accuracy (On validation data)

ward = 0.9972
single = 0.9972
average = 0.9972
complete = 0.9972

Cluster purity pie chart on training data



Cluster purity pie chart on validation data



Hierarical Clustering using Deep Autoencoder

bottleneck_layer_neurons = 5 (Because of memory constraint)
no_of_clusters = 5
bottleneck_layer_index = 2
learning_rate = 0.01
epochs = 1000
act_func = ["relu","relu","relu","relu"]

R2 train data: 0.6062825457010889
R2 validation data: 0.41724706749606222

Linkage_Type Accuracy (on train data)

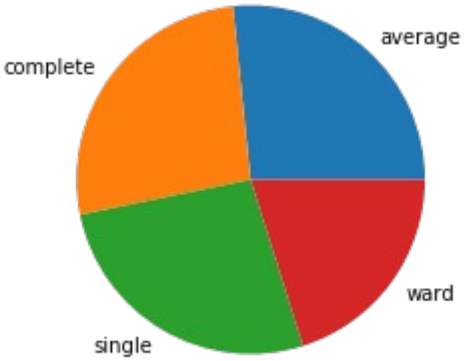
ward = 0.756144
single = 0.999429
average = 0.999314
complete = 0.999086

Linkage_Type Accuracy (on validation data)

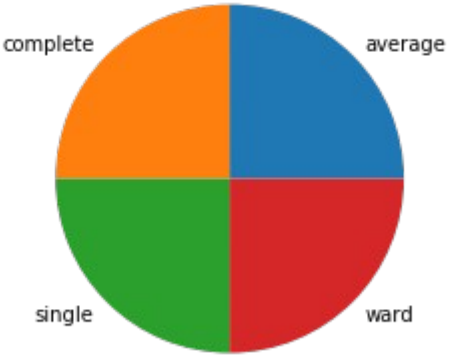
ward = 0.996667
single = 0.997200

average = 0.997200
complete = 0.996667

Cluster purity pie chart on training data



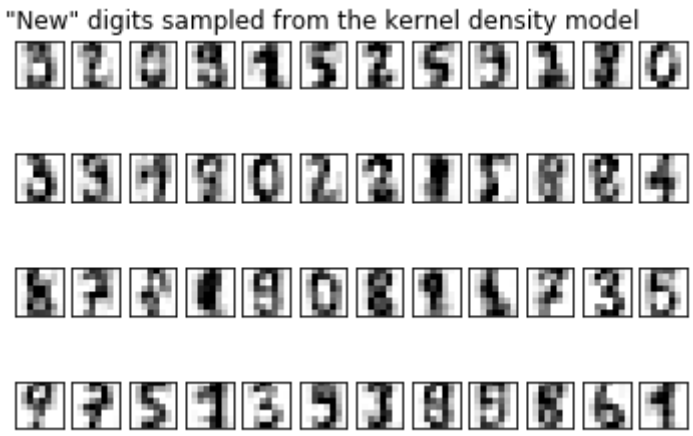
Cluster purity pie chart on validation data



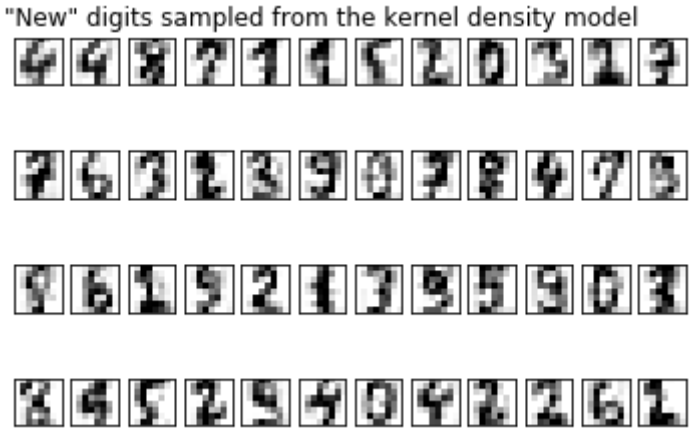
Question 2:

Kernel Density Estimation:

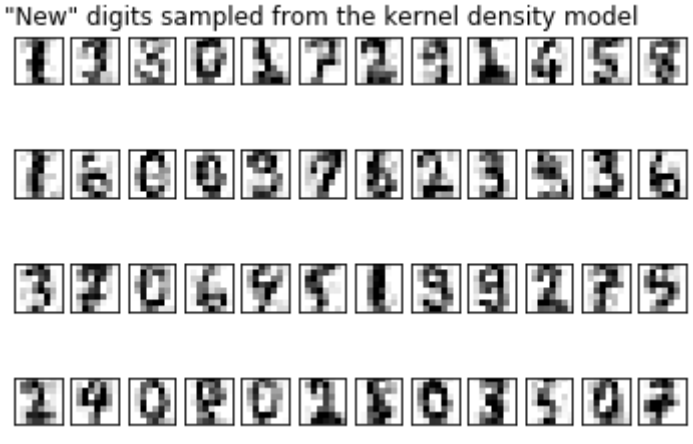
Best bandwidth: 3.3598182862837818
Output with number of principal components as 15



Best bandwidth: 3.3598182862837818
Output with number of principal components as 25

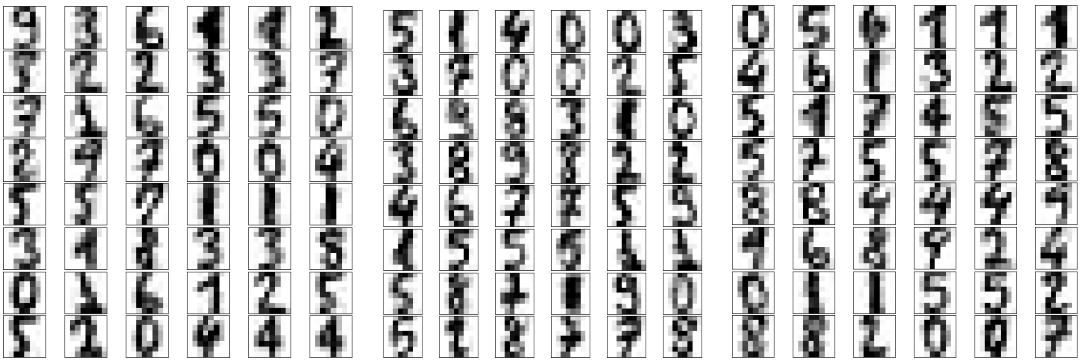


Best bandwidth: 2.976351441631318
Output with number of principal components as 35

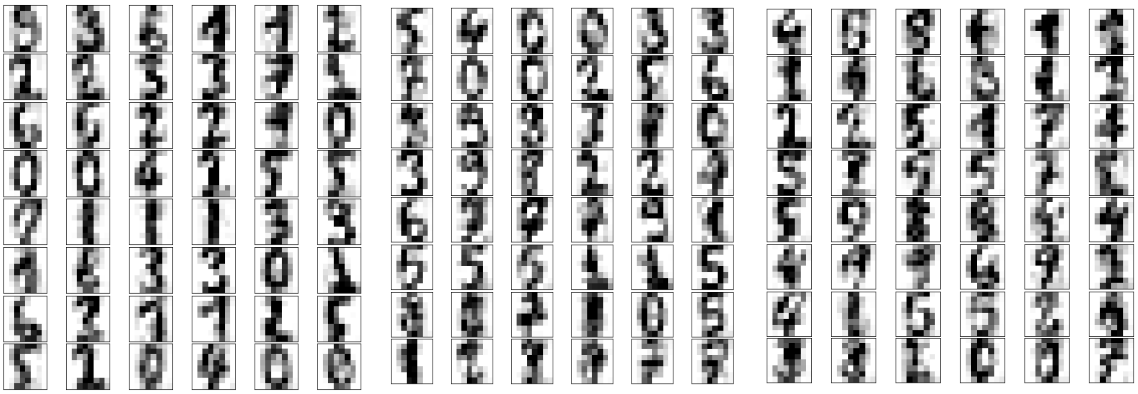


Guassian Mixture Model Density Estimate:

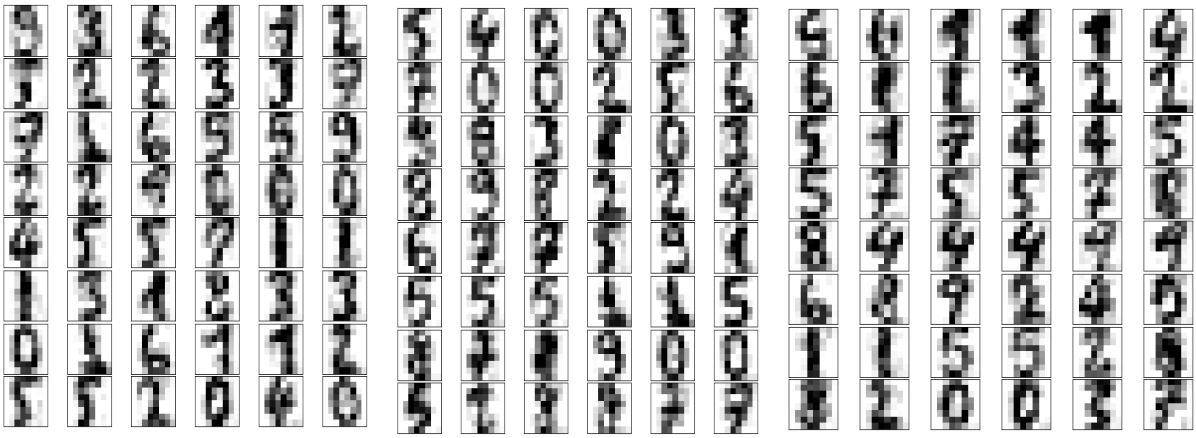
Output with number of principal components 29, 34 and 41 with covariance type full



Output with number of principal components 29, 34 and 41 with covariance type spherical



Output with number of principal components 29, 34 and 41 with covariance type diagonal



Output with number of principal components 29, 34 and 41 with covariance type tied