PRML Assignment-1 Report:

ED20B068
Tharun Anand M

1. i) For finding the principal components i first centred my data around the mean of the dataset(dimension:60000,784) then found the covariance matrix(784*784) .From that i was able to get the eigen vectors and eigenvalues and thus the projections on each principal Component

   For demonstration i plotted the first three principal components of a sample image

   The Variance of each component is the EigenValue Associated with that Corresponding EigenVector

ii) I displayed the reconstructed Data for an image as a weighted linear combination sum of all the principal components and plotted the original image beside it for comparison.Though the result from PCA is identifiable to its original image there were some white patches concentrated around the centre(Owing to the fact most of the data's white pixels are in the middle)

Also a suitable D for reconstruction would be 550.Because a large number of the eigen values are negative so my strategy of selecting the most important components was setting a threshold for **abs(eigen_value)>150**. By setting this condition i found near to 550 eigen values satisfying it.

iii)I took first 1000images For kernel PCA .
I proceeded with centered the K matrix (not the individual elements because for 1000 images the code took a long time to calculate 1000*1000 matrix).
Then calculated the eigen vectors and values and the corresponding Alphas.
Then I reconstructed the point as a linear combination of alphas with the corresponding kernel element.
In the graphs I plotted the projections of thesee points on the first two principal components.

iv)From the graphs of projection of the points on the principal components we can see that Polynomial kernels do better at representing the data on the first two principal components compared to the rbf kernels and among the polynomial kernels we see the best representation for d=2.
So the graph of projections on w1 vs w2 tells us was well and how much of the data those principal components were able to represent and it seemed to be highest for polynomial kernel for d=2

2.i) To implement K-means Clustering on the dataset the following steps were involved:

a)Initializing K random means

b)Initially randomly allocate points to clusters

c)Calculating The cost function.

d)Updating the points to the nearest mean and then updating the new means of every cluster

For each value of K the error decrease progressively after every iteration.Also the error depend on the random initializations taken.

iii)Spectral Clustering: I initially tried out Ploynomial Kernels for a few values of d but in the final results of the clustering towards the corner and edges of the two clusters there were a large number mismatches. When tried RBF kernels for sigma=1 to 10 and found sigma=8 giving the best cluster assignment.

So my final choice of kernel was RBF with sigma=8

iv) I observed a poorer cluster assignment based on the mappings they gave compared to the case where we used out RBF kernels.Selecting the maximum argument as the cluster no may not be an optimized way of clustering. Because when we find the maximum argument for each eigen value in the H matrix it is ordered columnwise so when we do the lloyd’s algorithm on the rows that may not always end up with the same results