Informatics 2B - Coursework 2

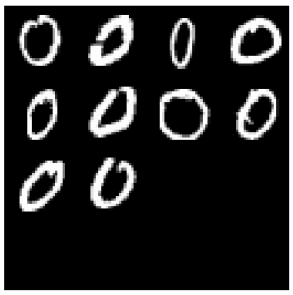
Task 1 - PCA and Clustering

s1765026 - University of Edinburgh

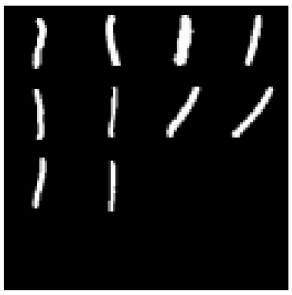
April 2019

Task 1.1

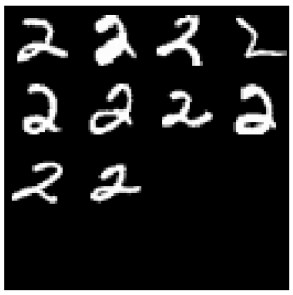
First ten smaples of digit 0



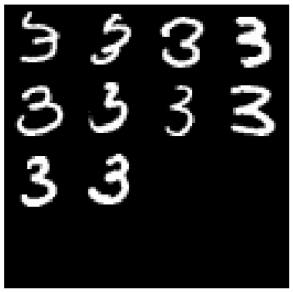
First ten smaples of digit 1



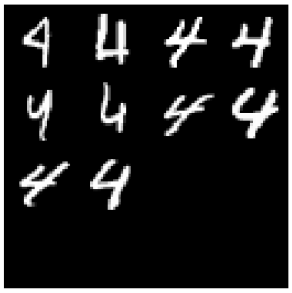
First ten smaples of digit 2



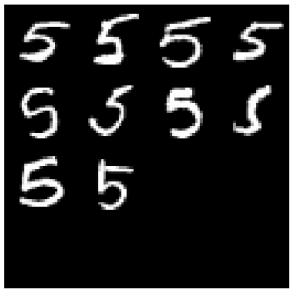
First ten smaples of digit 3



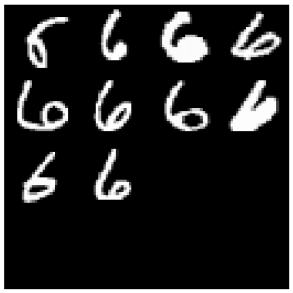
First ten smaples of digit 4



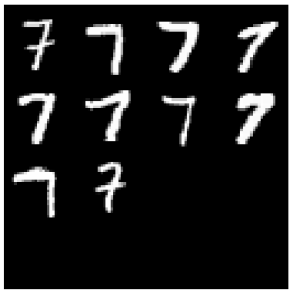
First ten smaples of digit 5



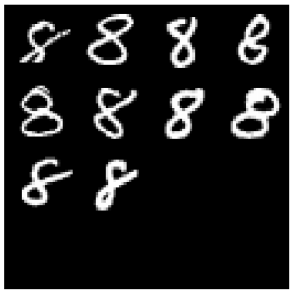
First ten smaples of digit 6



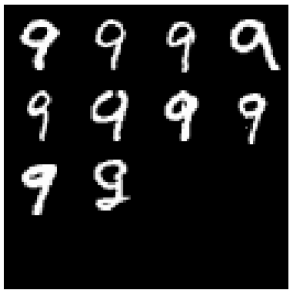
First ten smaples of digit 7



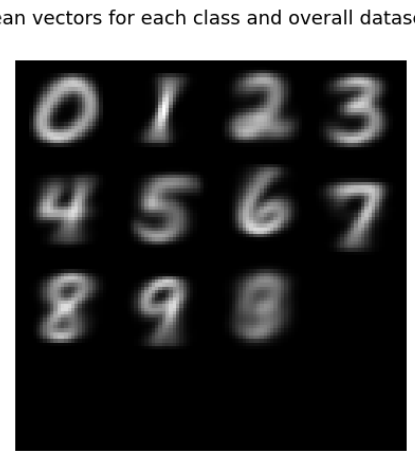
First ten smaples of digit 8



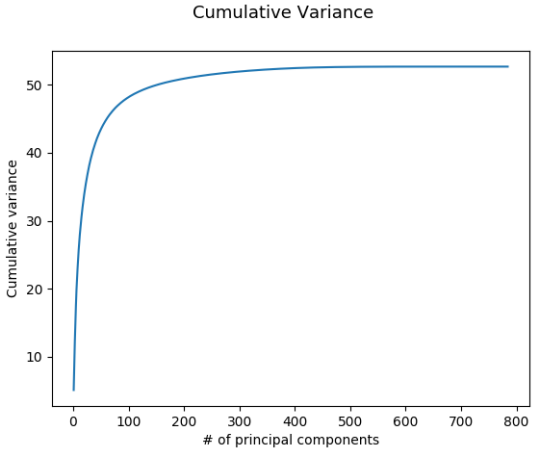
First ten smaples of digit 9



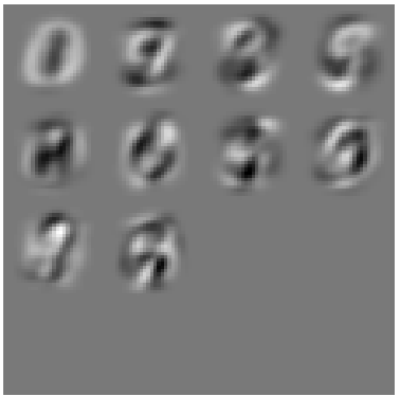
Task 1.2



Task 1.3

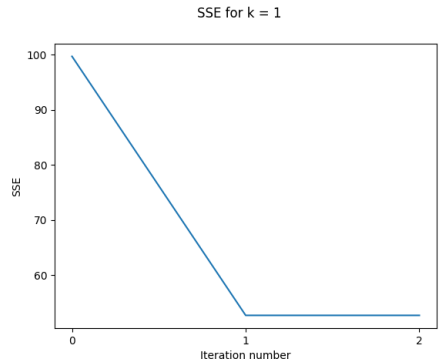


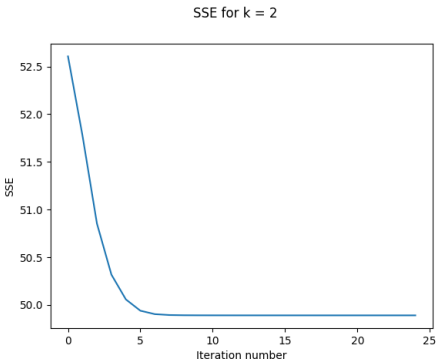
Task 1.4

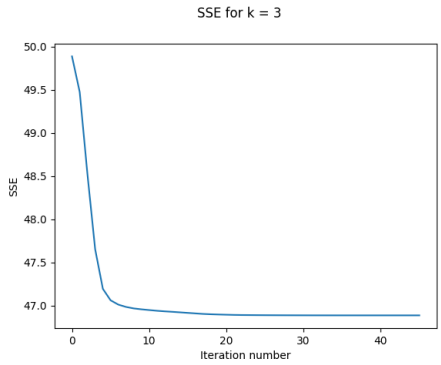


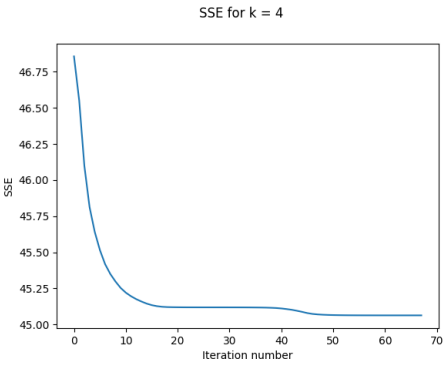
Task 1.5

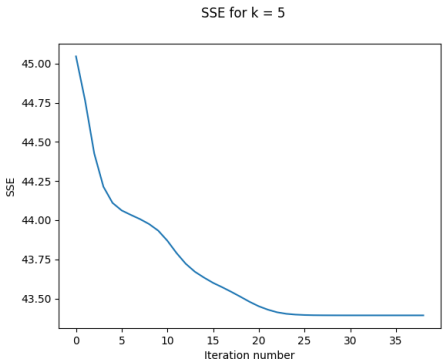
SSE over iterations for each K

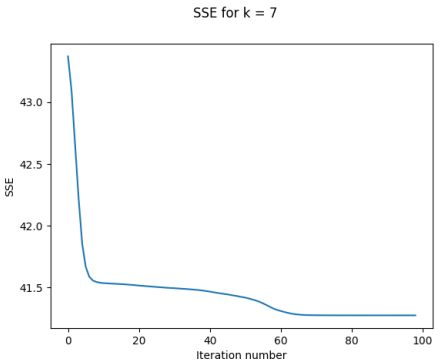


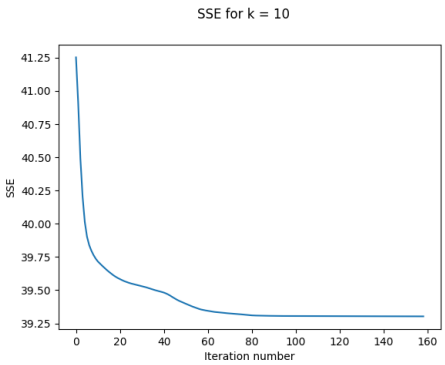


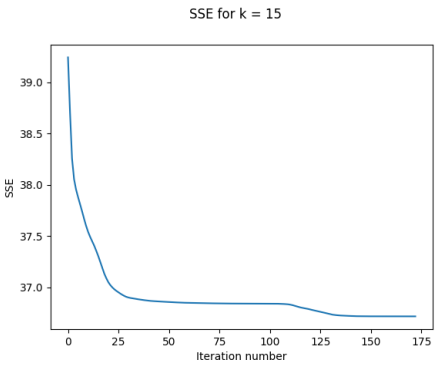


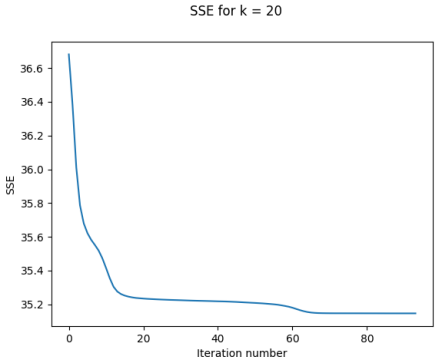












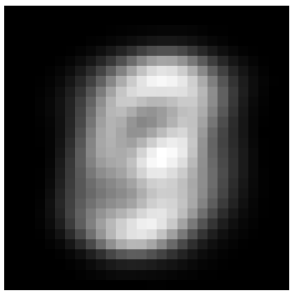
Runtimes for each K

K	1	2	3	4	5
Runtime (s)	0.42485	6.31980	15.18739	17.51115	9.24558

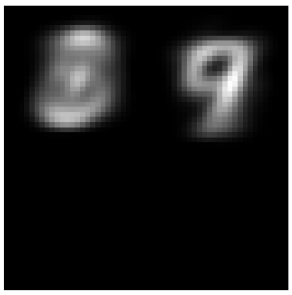
K	7	10	15	20
Runtime (s)	25.36083	36.59445	42.42998	27.17146s

Task 1.6

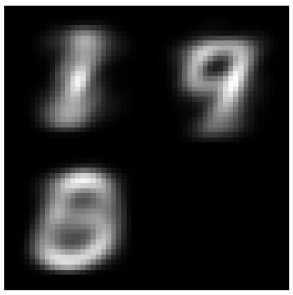
Cluster centres for k = 1



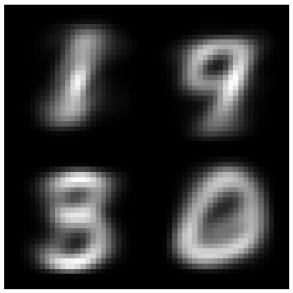
Cluster centres for k = 2



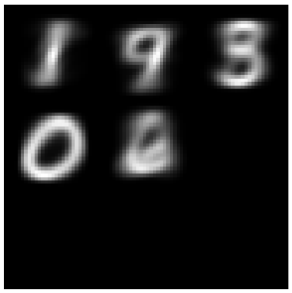
Cluster centres for k = 3



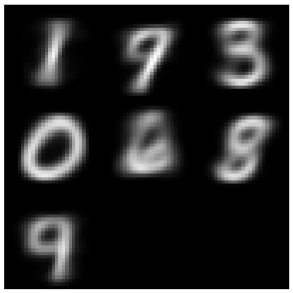
Cluster centres for k = 4



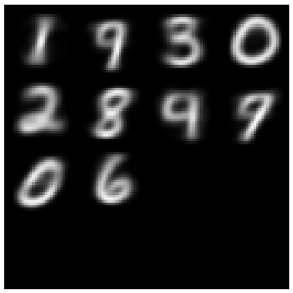
Cluster centres for k = 5



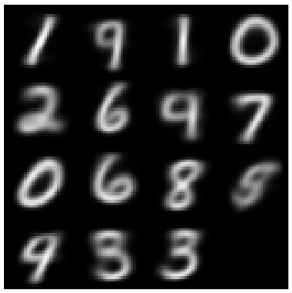
Cluster centres for k = 7



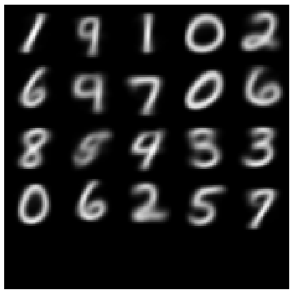
Cluster centres for k = 10



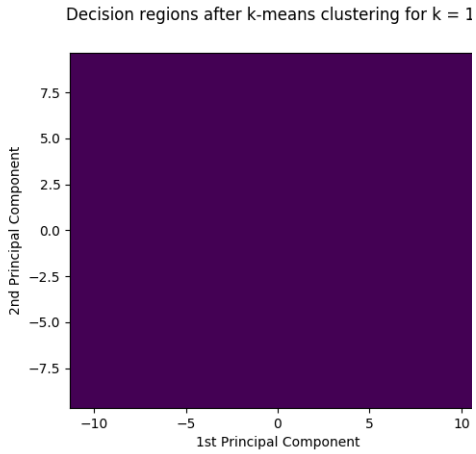
Cluster centres for k = 15

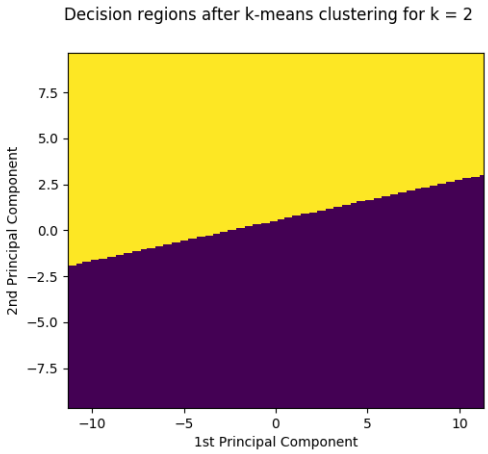


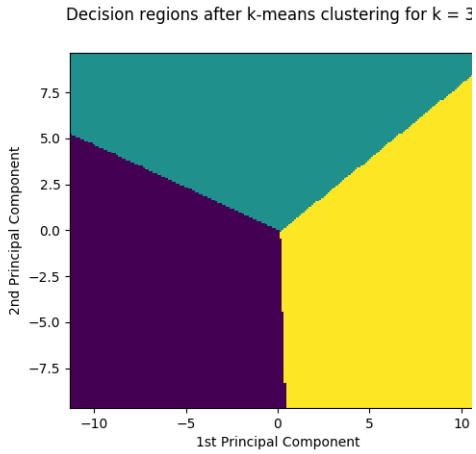
Cluster centres for k = 20

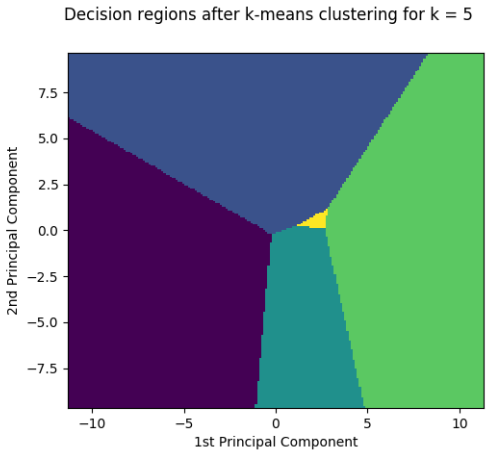


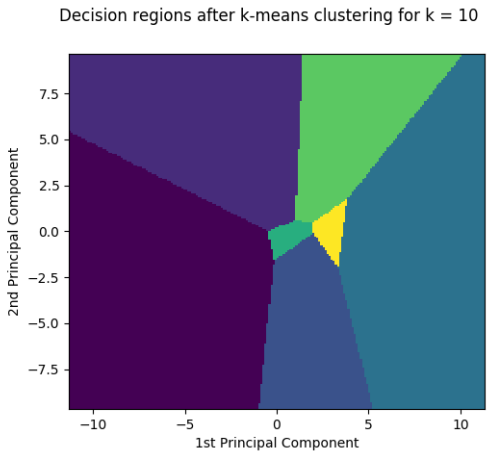
Task 1.7


Decision regions after k-means clustering for k = 1
2nd Principal Component
1st Principal Component
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10
−5
0
5
10


Decision regions after k-means clustering for k = 2
2nd Principal Component
1st Principal Component
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10
−5
0
5
10


Decision regions after k-means clustering for k = 3
2nd Principal Component
1st Principal Component
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10
−5
0
5
10


Decision regions after k-means clustering for k = 5
2nd Principal Component
1st Principal Component
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10
−5
0
5
10


Decision regions after k-means clustering for k = 10
2nd Principal Component
1st Principal Component
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10
−5
0
5
10

Methodology

To plot these graphs we started by loading the previously calculated eigenvalues and eigenvectors of the data. From them, we extracted the standard deviation (σ) for the first two principal components, that is, the square roots of the two largest eigenvalues.
After that, we load the mean vector of the dataset, and project it to the two-dimensional principal subspace, by multiplying it by the matrix containing the eigenvectors as columns.
The next step is to create a grid. However, first we need to determine the range of the plotting. To do that, we use the projected mean vector (μ) to calculate the desired $\mu \pm 5\sigma$ range for each of the axes. In this interval we generate a linear space (`np.linspace()`) with 200 samples for each axis (value of `nbins`). These can then be combined by `np.meshgrid()` to obtain our grid of coordinates.
Now each of the points in that grid needs to be classified, but this grid is in a two-dimensional subspace and we want our classification to be done in the original vector space. To achieve this, we can 'unproject' the grid.
Each point in the two-dimensional subspace (y) is obtained by projecting a corresponding point (x) in the original vector space. This process is described by the formula, where $\mathbf{p}$ is the position vector, $\mathbf{y}$ is the projected data, $\mathbf{x}$ is the original data, and $\mathbf{V}$ is a matrix with the eigenvectors as columns:

$$\mathbf{y} = \mathbf{V^T}(\mathbf{x} - \mathbf{p}) \Leftrightarrow \mathbf{x} = (\mathbf{V^T})^{-1}\mathbf{y} + \mathbf{p}$$

By first padding the grid elements with zeros to match the original dimensions and then applying the right-hand side of the equivalence we 'unproject' our grid. Finally, we can assign each point in the 'unprojected' grid to the closest cluster center obtaining our `Dmap`, which is used to plot the graphs above.

Task 1.8

For this small research project we were asked to investigate the impact of the initial cluster centres chosen in the performance of k-means clustering. The first step was the decision work exclusively with $k = 10$ and use the previously suggested method (using the first ten samples of the dataset as initial centres) as our 'control subject'. Please note that each of these samples belongs to a different class.

From there the research split the experimentation into different branches:

1. Options that would always be available
   (a) Randomly pick 10 samples, from any class, from the dataset
   (b) Calculate the mean of the dataset and choose the 10 samples which are further away from it
2. Options that would only be possible with already tagged data (supervised learning - and so not very useful in a real world context)
   (a) Randomly pick 10 samples, one from each class, from the dataset
   (b) Use the mean of each class as the initial cluster centres

Please find the results obtained in the last page.

Conclusions

These allow us to formulate some hypotheses. The main observation is that choosing initial centres away from the mean seems to be more efficient in the sense that the it results in a significantly lower number of iterations. The final sum squared error is marginally higher but there do not seem to be many fluctuations of that parameter. Complete randomness is the next best approach.

Additionally, we can also observe by the shape of the graphs that the error decreases much quicker when the initial centres are the furthest from the mean. This means that if we set a looser early termination condition we could safely cut the number of iterations down to about one fourth or less (in our example).

Finally, it also becomes apparent that having more information about the data (in this case the actual labels) does not seem to help us do a better job in estimating good initial cluster centres.

Overall, the main hypothesis is that the furthest the initial centres are from the mean of each class, the faster the process seems to converge. However to confidently claim this hypothesis would require further and more advanced testing.

Method	'Control'	1. (a)	1. (b)	2. (a)	2. (b)
Number of iterations	78	74	43	52	112
Final error	39.15347	39.15347	39.35176	39.32280	39.29411