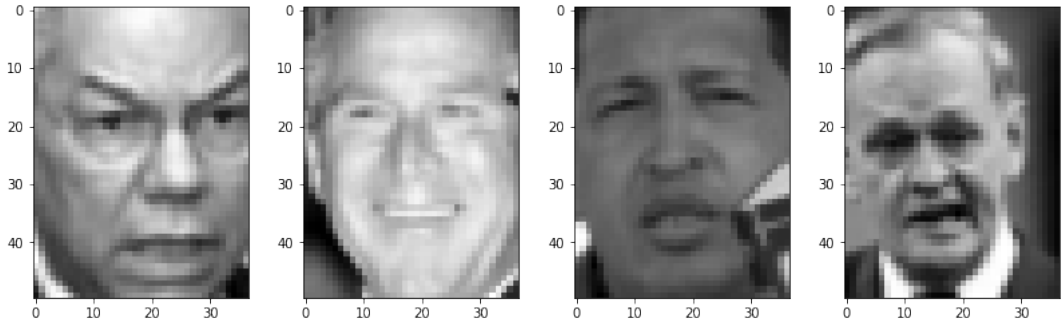
CM146, Winter 2020
Problem Set 4: Clustering and PCA
Due March 15, 2020
Submitted by Guy Ohayon

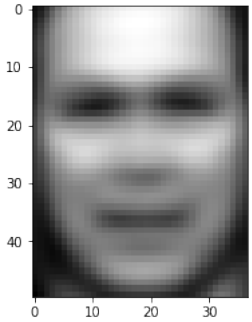
1 PCA and Image Reconstruction

Solution:

(a) These are plots of a couple of the input images in the dataset:



And this is the mean image:



We can see that the average face (mean image) looks like an actual face. In the average face, we can see a mouth, nose, the frames of the eyes, and a facial frame.

(b) These are the top twelve PCA of the dataset:

These twelve eigenfaces are the twelve principal components with the highest variance (sorted). This means that these are the most informative eigenfaces out of all principal components. We can see that the plotted faces truly are of high variance, and that they capture important face features (such as nose, eyes, face frame, and mouth).

(c) $l = 1$:

$l = 10$:

$l = 50$:

$l = 100$:

$l = 500$:

$l = 1288$:

We can see that $l >= 500$ is highly effective for facial recognition. Also, $l <= 100$ is not sufficient for recognition, but we can see good

facial reconstruction. Thus, there might be a middle ground between $l = 500$ and $l = 100$ that's small and sufficient for recognition.

2 K-Means and K-Medoids

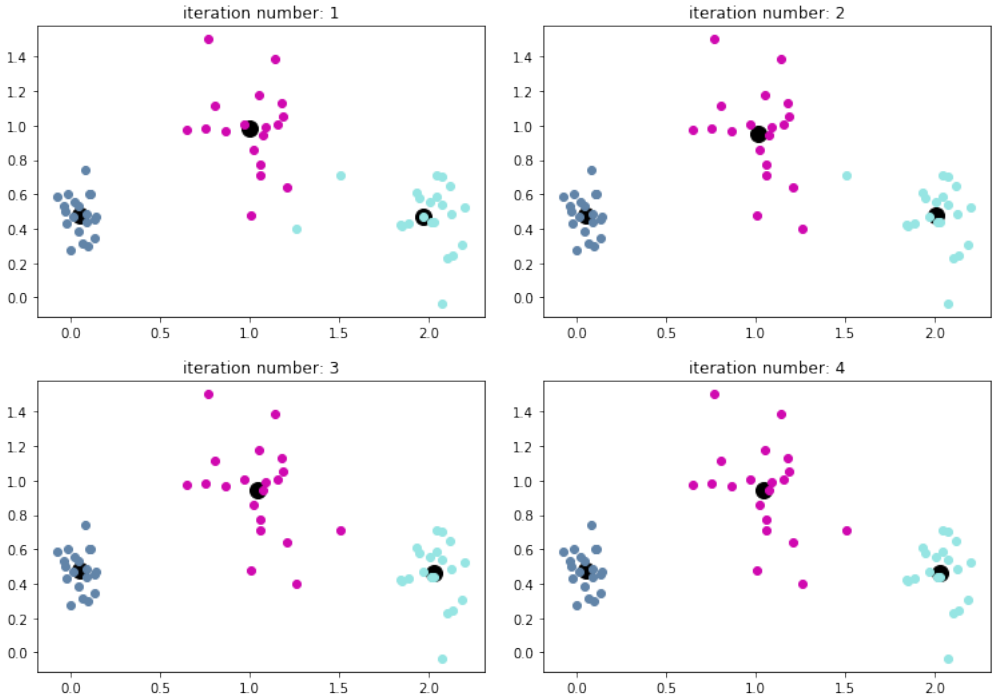
Solution:

(a) $J(c, \mu, k) = \sum_{i=1}^{n} ||x^{(i)} - \mu_{c^{(i)}}||^2$ where $c^{(i)} \in 1, \ldots, k$. To minimize this function, we can choose $k = n$, $c^{(i)} = i$, and $\mu_i = x^{(i)}$ (each example in a cluster alone, and each example is also the centroid of its cluster). Therefore, in this case, the minimum possible value of $J(c, \mu, k)$ will be 0, because $\forall i, ||x^{(i)} - \mu_{c^{(i)}}||^2 = 0$. This is clearly not a good idea since it results in overfitting.

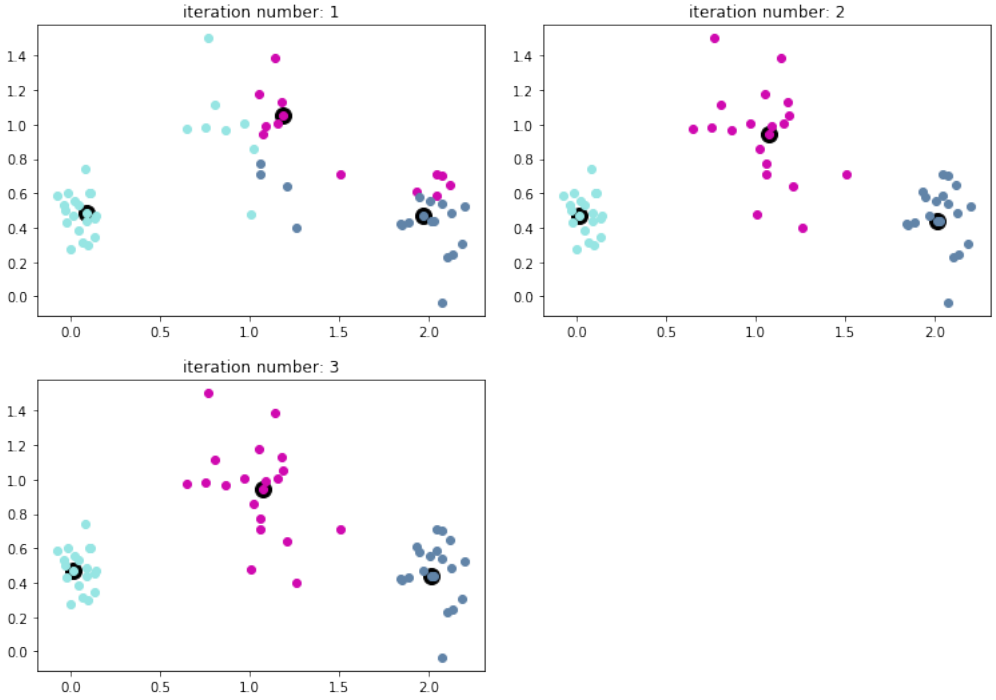
(b) Code section.

(c) Code section.

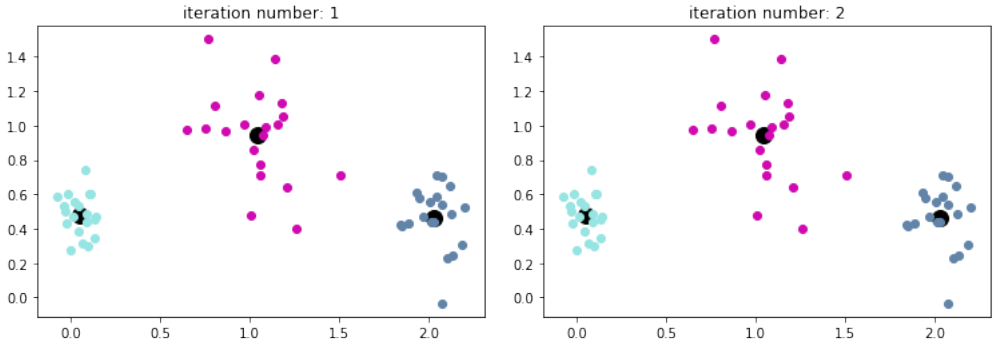
(d) The following are plots for k-means cluster assignments and corresponding cluster centers for each iteration (when using random initialization):



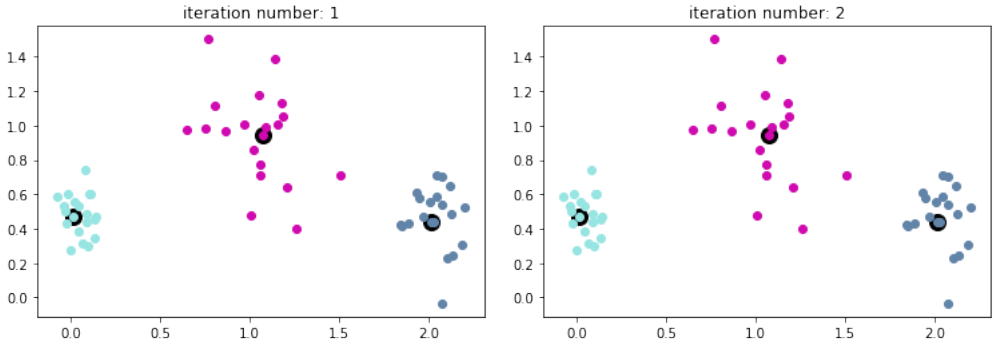
(e) The following are plots for k-medoids cluster assignments and corresponding cluster centers for each iteration (when using random initialization):



(f) By using cheat initialization, we can see that all data points are correctly clustered starting from iteration 1. These are the plots for k-means:



And these are the plots for k-medoids:



3 Clustering Faces

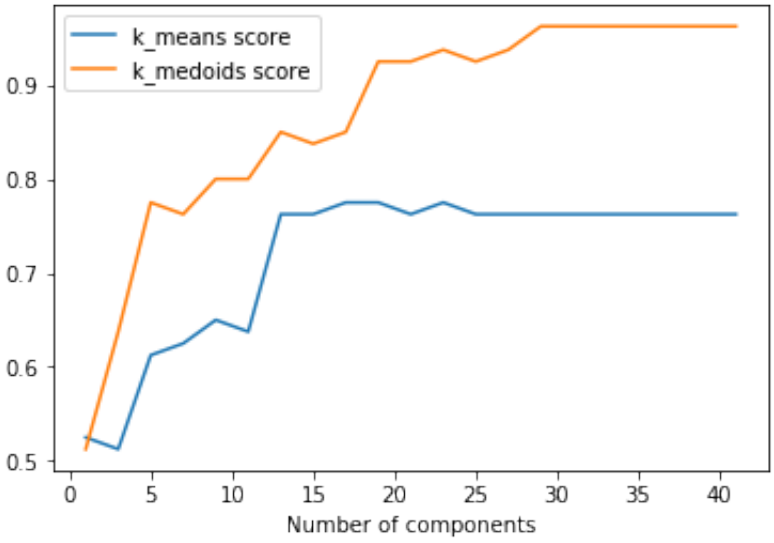
Solution:

(a)

	average	min	max
k-means	0.5875	0.51875	0.7125
k-medoids	0.619375	0.59375	0.71875

In terms of clustering performance, we can see that k-medoids performed better, and both algorithms have approximately the same runtime, which was ≈ 0.25 seconds on average over all 10 iterations.

(b) These are the results obtained for $K = 2$ and $l = 1, 3, 5, \ldots, 41$:



We can see that k-medoids consistently performs better than k-means.

Also, we can see that the performance stalls if we use more than 30 components in k-medoids, or more than 25 components in k-means. Therefore, it's not useful (in terms of clustering with k-medoids or k-means) to use more than 30 components with k-medoids, or to use more than 25 components with k-means.

(c) I chose to use k-medoids to discriminate pairs of individuals. For each pair of labels $i, j = 1, 2, \ldots, 19$ where $i \neq j$, I sampled 40 images for each label. I then performed PCA with $l = 30$ to reduce the dimensionality of each image, executing clustering using kMedoids, and calculated the score of the clustering. The worst discrimination score was for classes 3 and 7, with a score of 0.5. The best discrimination score was for classes 6 and 14, with a score of 0.975. The following are the images of the pair [3,7], which had the worst discrimination:


The following are the images of the pair [6,14], which had the best discrimination:


We can see that the worst discrimination pair of images are a lot more similar than the best discrimination pair of images, which makes sense because better clustering means better discrimination between the images (the 2nd pair received a better clustering score).