COL774 : Assignment 4

Aditi Singla
2014CS50277

21st April 2017

1. K-means

   (a) The K-Means clustering algorithm has been implemented here. The dataset has 10411 examples with each example being represented by 561 attributes.

   - Number of clusters, K : 6
   - Convergence Criteria : Number of iterations = 40
     (Converges after around 30 iterations)
   - Original Distribution : [1722, 1544, 1407, 1801, 1979, 1958]
   - Distribution obtained : [2329, 1139, 1382, 1877, 2817, 867]
   - Error, J : 174376.1658

   (b) Minimising the Error Functions..

   - Error values obtained on 10 random initialisations :
     (174373.469, 173772.019, 173823.334945, 179620.2380, 180128.045, 174371.1746, 180129.693, 173747.746, 174373.216, 174740.769)
   - Best value of J obtained : 173747.746
   - For the best value of J obtained, the variation of accuracy with number of iterations is as follows :

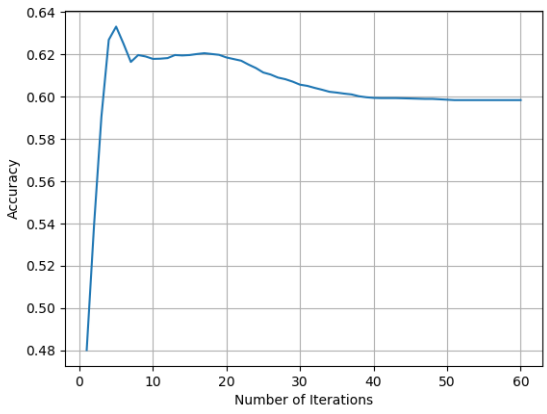

Figure 1: Accuracy v/s Number of Iterations

Here, we observe that as the number of iterations increases, the accuracy shoots to around 63% and then after a small dip, it flattens ar around 61%. Eventually, on convergence, the flattening of the curve takes place. A local maxima might be due to inaccurate labelling of given data.

- For the best value of J obtained, the variation of error, J with number of iterations is as follows :

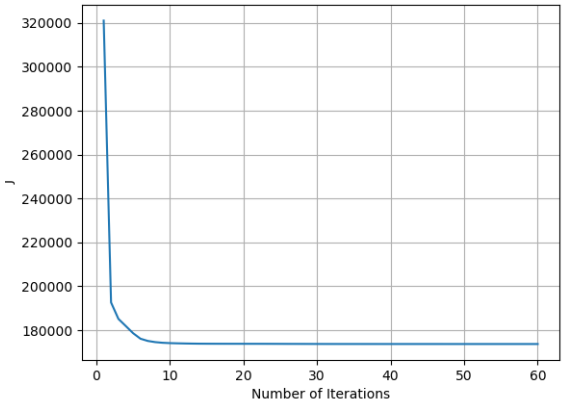

Figure 2: Error(J) v/s Number of Iterations

The error term J is very high when the number of iterations are very less. As the number of iterations increases, the error value drops sharply. The initial high value is because of the random initialisation of the K centres, chosen from the complete data. As the number of iterations increases, we tend towards minimising the error value.
The curve of J almost flattens after a few iterations, since the clusters are more or less same, with a very few data points moving around in every iteration.
Convergence is observed at around 30 iterations, beyond which the cluster centroids donot move.

(c) Accuracies obtained :

- Average Accuracy obtained = 61.0316 %
- Highest Accuracy obtained = 63.54 %
- Error obtained for average, J = 174373.469

(d) Sklearn - Linear SVM for Classification :

- 10-fold Cross-Validation Accuracy = 93.517%
- Time taken = 165.2206 s

The accuracies obtained using cross-validation are way too higher (about 22%) than the values obtained by using k-means.
But a point to be noted here is that SVM classifier needs labels to train, it being a supervised learning model, while K-means doesn't need any labels, being an unsupervised learning model. Also, SVM needs large data to train on, to learn all the parameters, while K-Means doesn't since it just needs to find the means for the data.

2. Principal Component Analysis

(a) Average Faces :

Figure 3: Dataset 1 & Dataset 2

(b) PCA using SVD libraries :

- The data is modified to zero mean and unit variance.
- PCA is performed on the modified data, X, to obtain the eigen values and corresponding eigen vectors.
- Top 50 eigenvectors are chosen as Principal Components, which are then modified to unit variance (already zero mean). These components correspond to the eigen faces.
- These principal components are stored in a file and used to find the projections of each of the data image into reduced dimensions, and stored in a different file.

(c) Eigen Faces for top 5 components: These components (vectors) are then scaled up to get them in the range of 0-255, to get a clearer picture.

Figure 4: Dataset 1

Figure 5: Dataset 2

(d) Sklearn - Linear SVM for Classification :

Data Set 1

- Accuracy on original data before PCA : 41.156% (76.166s)
- Accuracy on normalised data before PCA : 74.527% (54.731s)
- Accuracy on projected data after PCA : 74.907% (2.2843s)

Data Set 2

- Accuracy on original data before PCA : 63.75% (38.725s)
- Accuracy on normalised data before PCA : 98.50% (38.470s)
- Accuracy on projected data after PCA : 97.50% (0.4845s)

Observations :

- We observe huge increase in efficiency after normalising the data.
- The total time to train the normalised data reduces substantially, due to reduction in dimensionality.
- Accuracy slightly decreases in case of Data set 2, this might happen due to over-reduction of the attributes, causing loss of important data as well.

(e) Comparision between original and the projected face :

- For reconstruction, the input file is taken, its projected image is multiplied by the eigen vectors, where the eigen vectors were also unit variance.
- Since the data has been normalised, the image is a bit blur, but still is able to capture the basic features. A few details lack, due to which image is less sharp.
- On increasing the number of principal components, the image becomes clearer and clearer, capturing the whole image when all the eigenvectors are taken as principal components.

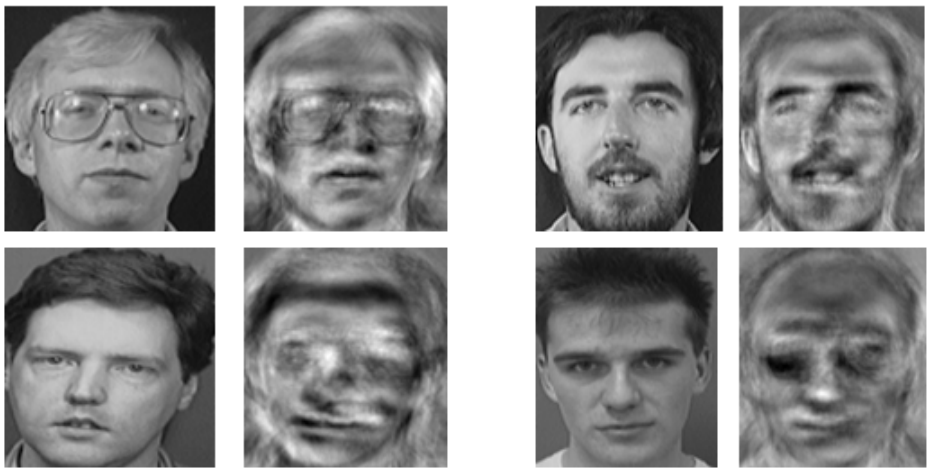

Figure 6: Dataset 1

Figure 7: Dataset 2

.