Run AT&T Face Image. Total 40 images contains X= (x1 x2 ... x40)

Kmeans results are:  Kindex = kmeans(X,4) =

Columns 1 through 20

1 1 1 1 1 1 1 1 1 1 4 4 4 4 4 4 4 4 4 4

Columns 21 through 40

2 2 2 3 3 3 3 3 2 2 3 2 3 2 2 3 2 3 2 2

This means :

x1 belongs to predicted-cluster 1
x2 belongs to predicted-cluster 1
.....
x19 belongs to predicted-cluster 4
x20 belongs to predicted-cluster 4
x21 belongs to predicted-cluster 2
.....
x24 belongs to predicted-cluster 3
.....
x38 belongs to predicted-cluster 3
x39 belongs to predicted-cluster 2
x40 belongs to predicted-cluster 2

The correct (ground-truth) cluster labels for x1 ... x40 are:

Columns 1 through 20

1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2

Columns 21 through 40

3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4

confusion matrix :
C(i,j) = the number of data points that belong to i-th true-cluster, but are clustered to j-th predicted cluster
i = true-cluster
j = predicted cluster
(You should play around with the k-means clustering to be familiar with these concepts.)

Based on the above gound-truth labels and predicted cluster labels given by K-means,
we compute confusion matrix

The results are:

| | | | |
|---|---|---|---|
| 10 | 0 | 0 | 0 |
| 0 | 0 | 0 | 10 |
| 0 | 5 | 5 | 0 |
| 0 | 6 | 4 | 0 |

Interpretation:
1st column: 10 images are clustered to the predicted-cluster 1. All of them belong to true-cluster 1.
2nd column: 11 images are clustered to the predicted-cluster 2; 5 of them belong to true-cluster 3; 6 of them belong to true-cluster 4.
etc

Running bipartite graph matching based using the graph edge weight matrix = confusion matrix,
we obtain the column permutation:
column-index = [1 4 3 2].

This means:
column-4 should be permuted to column-2, i.e., predicted cluster 4 should be labeled as cluster 2
column-2 should be permuted to column-4, i.e., predicted cluster 2 should be labeled as cluster 4
Column indexes 1 and 3 should remain unchanged.

After this column index permutation,
the confusion matrix becomes:

| | | | |
|---|---|---|---|
| 10 | 0 | 0 | 0 |
| 0 | 10 | 0 | 0 |
| 0 | 0 | 5 | 5 |
| 0 | 0 | 4 | 6 |

The accuracy = (10+10+5+6)/N = 0.775. Thus, out of 40 images, 31 are correctly clustered.

Explanation:

In any clustering algorithm, the final results are grouping, i.e., which data points are grouped to a group.
Data points (feature vectors) having the same cluster label are grouped into a cluster.
Exactly which group is labeled as Cluster 1, which group is labeled as cluster 2, etc, are completely un-determined.

But when assessing the quality of a clustering (a grouping), we need to re-label (permute predicted cluster labels)
so that a predicted cluster is maximally related to a true cluster. This is achieved using bipartite matching.

The confusion matrix is a bipartite graph. We use Hungarian algorithm to permute columns of
the confusion matrix to obtain the optimal matching.