Clustering and PCA

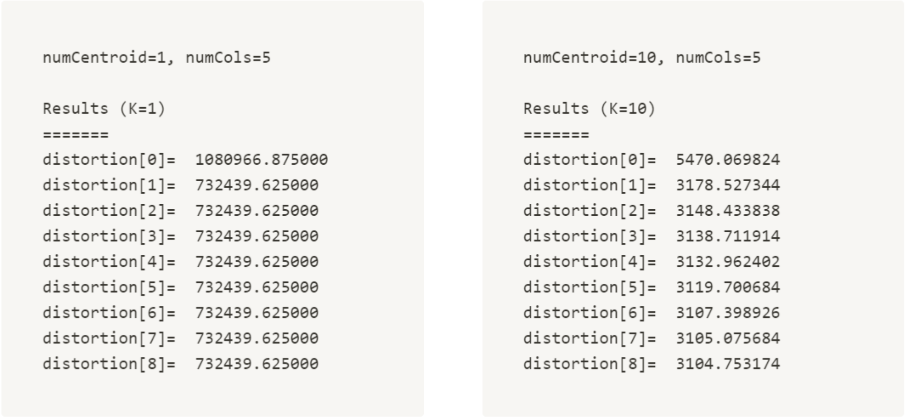
William L. and Andrea T.
Data Mining and Machine Learning

Clustering

Clustering can be applied to data using the given *agglom.c* and *k-means.c* executable files. *Agglom.c* gives us a set of K initial centroids that can be used for K-means clustering to locally optimize those centroids. Additionally, the distortion for the given set of centroids relative to data are also outputted. Using the command prompt, inside the lab2-2021 directory, we can use the format `agglom <dataFile> <centFile> <numCent>` and `k-means <dataFile> <centFile> <opFile> <numIter>` to execute *agglom.c* and *k-means.c*, respectively.

Distortion for Different Values of Initial Centroids

For one to ten initial centroids K, we executed $K = 1$: `agglom lab2Data lab2_1C 1`, until $K = 10$: `agglom lab2Data lab2_10C 10`. Having produced ten more files, nine iterations of K-means clustering was applied to them using the commands: `k-means lab2Data lab2_1C lab2_1op 9`,… , `k-means lab2Data lab2_10C lab2_10op 9`. The first and last result of the distortion is shown in Fig. 1.



```
numCentroid=1, numCols=5

Results (K=1)
=======
distortion[0]=  1080966.875000
distortion[1]=  732439.625000
distortion[2]=  732439.625000
distortion[3]=  732439.625000
distortion[4]=  732439.625000
distortion[5]=  732439.625000
distortion[6]=  732439.625000
distortion[7]=  732439.625000
distortion[8]=  732439.625000
```

(a)

```
numCentroid=10, numCols=5

Results (K=10)
=======
distortion[0]=  5470.069824
distortion[1]=  3178.527344
distortion[2]=  3148.433838
distortion[3]=  3138.711914
distortion[4]=  3132.962402
distortion[5]=  3119.700684
distortion[6]=  3107.398926
distortion[7]=  3105.075684
distortion[8]=  3104.753174
```

(b)

Fig. 1. distortion values for (a) K=1 and (b) K=10 initial centroids

The nine numbers appearing in Fig. 1, as well as every other output file, are the distortion values of the data after every iteration, when the centroid(s) are optimized and moved around. It can also be seen that after every iteration, the distortion value will decrease until it cannot decrease anymore. As for when $K = 1$, the distortion value stops decreasing after the second iteration because there is only one centroid, and after the second iteration, the algorithm was not able to locate a better location for the centroid to be in. But for $K = 10$, there are ten centroids being optimized and therefore, even subtle changes in one centroid might change the distortion values of the data.

Table 1: Final distortion values for K initial centroids

K initial centroids	Distortion[8]
1	732439.6
2	317257.2
3	78126.7
4	32335.8
5	13455.1
6	5021.8
7	4581.7
8	3987.8
9	3529.1
10	3104.8

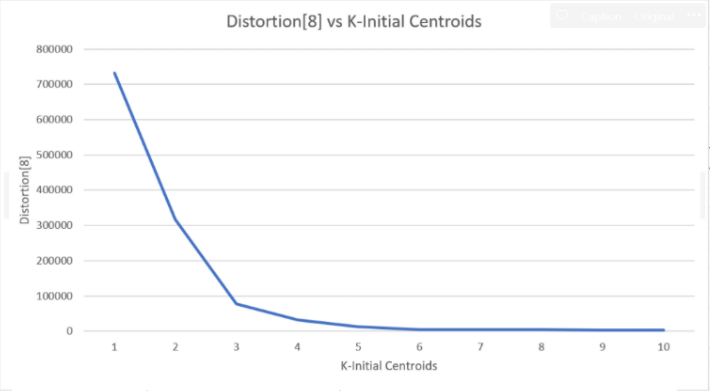


Fig. 2. Plotted final distortion values for K initial centroids

Table 1 and Fig, 2 shows the final distortion values after nine iterations as the K initial centroid increases. With there being more centroids to cover a lot more spaces in the data, the distances between each data point to the closest centroid becomes smaller, therefore decreasing the distortion. But the huge decrease in distortion is mainly due to the centroids covering all the important groups and discovering the structure of the data.

From Fig. 2, it can be deduced roughly that there are around five to six groups of data rather than just being one big group. This is because even after increasing the initial centroids from six to seven, or eight, or nine, the amount of distortion starts to decrease smoothly, supposedly saying that the centroids have found the centre of the different groups. At low initial K centroids, the distortion values are very big because the groups are probably far apart, and some centroids might even be in between two or three groups at once.

Principal Component Analysis (PCA)

MATLAB was used to apply PCA to the data. The data was loaded using the line `X = load('lab2Data-matlab');`. To do PCA, we first require the covariance matrix of the data, which can be obtained using the line `C = cov(X);`. The result of the covariance matrix C is

$$\mathrm{C} = \begin{bmatrix} 9.592 & 16.081 & 42.364 & 26.461 & 31.764 \\ 16.081 & 49.688 & 130.243 & 81.421 & 97.703 \\ 42.364 & 130.243 & 344.999 & 215.608 & 258.717 \\ 26.461 & 81.421 & 215.608 & 134.799 & 161.747 \\ 31.763 & 97.703 & 258.717 & 161.747 & 194.093 \end{bmatrix}$$

It is hard to deduce much about the data using the covariance matrix C, so to get more information on the structure of the data we can use eigenvalue decomposition on the covariance matrix using the line `[U,D] = eig(C)`. The results were eigenvectors U and eigenvalues D.

$$\mathrm{U} = \begin{bmatrix} 0.0017 & 0.0040 & 0.0191 & -0.9962 & 0.0850 \\ 0.0027 & -0.0292 & -0.9651 & 0.0036 & 0.2600 \\ -0.0111 & -0.6928 & 0.2066 & 0.05989 & 0.6882 \\ 0.7746 & 0.4497 & 0.1046 & 0.04185 & 0.4302 \\ -0.6324 & 0.5629 & 0.1205 & 0.04762 & 0.5162 \end{bmatrix}$$

$$\mathrm{D} = \begin{bmatrix} 0.0046 & 0 & 0 & 0 & 0 \\ 0 & 0.0661 & 0 & 0 & 0 \\ 0 & 0 & 0.4724 & 0 & 0 \\ 0 & 0 & 0 & 4.3584 & 0 \\ 0 & 0 & 0 & 0 & 728.2710 \end{bmatrix}$$

The eigenvectors U tells us the rotation of the matrix and transforms the data so that it becomes the new matrix D, which is our eigenvalues. While the eigenvalues from D tells us that most of the variation of the data is contained in the new 5th dimension, the principal component. And the least variation is in the 1st dimension, which could just be noise. The first three dimensions show negligible variations along their eigenvectors so it might be better to ignore them.

Conclusion

K-means clustering is very useful for finding groups of data that are unlabelled, and we have found out that there are approximately five to six different groups in the data set. From the PCA, it can be concluded that the first three dimensions are most likely useless due to them having very low variance. With PCA, we can reduce the dimension of the data and decrease its complexity so it can be more easily interpreted. Because the first three dimensions do not hold useful information, when further analysing the data, most of the important information will be in the fifth dimension, with a direction vector of the 5th column in the U matrix.