LM Data Mining and Machine Learning (2021)
Lab 2 – Clustering and PCA

Objectives
The objective of this lab is to use the methods described in lectures to discover the structure of a particular data set. At the end of the lab your task is to write down an intuitive textual description of the data. The techniques that you should apply are clustering and PCA.

What you will need
All of the files that you will need are in the zip archive `lab2.zip` which is on the Canvas page.

The Data
The data is stored in a text file called `lab2Data` (in the zip file `lab2.zip` on the Canvas page). The data consists of 1,000 points in 5 dimensional space. Each point appears as a row in the data file – have a look at the file to see its structure. There is a 'header' at the top of the file that specifies the number of columns and rows.

Part 1: Clustering

Your first task is to use clustering to try to determine whether there are natural clusters in the data, and if there are, how many. To do this you need to apply clustering to the data. You need two C programs `agglom.c` and `k-means.c`. Use the provided .exe files (or compile these two source C programs if needed).

The program `agglom.c` is an implementation of the agglomerative clustering algorithm described in lecture material. You should apply this to the data set to obtain a set of *K* initial centroids for k-means clustering (see the lecture notes to understand how). Then use `k-means.c` to locally optimize the centroids. As well as producing a locally optimized set of centroids, `k-means.c` returns the distortion for that set of centroids relative to the data. I recommend 9 iterations of k-means clustering.

Usage of `agglom` program: `agglom dataFile centFile numCent`
Runs agglomerative clustering on the data in `dataFile` until the number of centroids is `numCent`. Writes the centroid coordinates to `centFile`.

Usage of `k-means` program: `k-means dataFile centFile opFile numIter`
Runs `numIter` iterations of k-means clustering on the data in `dataFile` starting with the centroids in `centFile`. After each iteration writes distortion and new centroids to `opFile`.

You should use `agglom.c` and `k-means.c` to plot a graph of distortion as a function of *K*, the number of clusters. Plot distortion for values of *K* between 1 and 10. To clarify:

```
for K=1 to 10
•   Apply agglom.c to the data set to obtain K initial centroids
•   Apply 9 iterations of k-means clustering. A list of 9 numbers will appear on the screen.
    What are they? For each K make a note of the final number.
End
```
Plot a graph of these ten numbers against *K*.

Conclusion to Part 1: What does the graph tell you about the structure of the data?

Part 2: Principle Components Analysis (PCA)

To apply PCA to the data you will need to use MATLAB. MATLAB will complain about the header at the start of the data file `lab2Data`. Therefore I have created a version of this file without the header, called `lab2Data-matlab`. Use this file with MATLAB.

The procedure for applying and interpreting PCA is described in lecture material. In brief, the stages are as follows:

1. Load the data into a matrix, *X* say, in MATLAB.

2. Compute the covariance matrix of the data. You can either do this by implementing the formula for covariance given in the lectures, or you can simply use the MATLAB `cov` function:

```
>> C = cov(X)
```

3. Apply eigenvector/eigenvalue decomposition to the covariance matrix:

```
>> [U,D] = eig(C)
```

Write down the eigenvalues:

Conclusion to Part 2: What does the eigenvector/value decomposition of the covariance matrix *C* tell you about the structure of the data set?

Finally: Your summary

Summarize your findings. Write a short intuitive textual description of the data set:

END