Module Big Data & Small Data
Data collection and features of data
Assignment III: unsupervised algorithms

K-means

[1] Implement the K-means clustering algorithm. Your implementation should be put in a function named `k_means`.

The function must accept:

- `events`: an n-column event matrix
- `CLUSTERS`: the number of used clusters

The function must return a (n+1)-column matrix containing the event matrix and an additional column containing the cluster assignment, and the found cluster centroids.

Your function should properly initialize the cluster centroids, and perform the clustering of the events to the centroids. Hand in your code together with a document with a brief description of your implementation and motivation of your implementation choices.

You can test your implementation using the script `run_kmeans_assignment`. To run this script you should download the data-files `2017-12-street.mat` and `2017-12-stop-and-search.mat`. The data-file `2017-12-stop-and-search.mat` contains the coordinates of the streets where the British police "stopped and searched" somebody. The script will load the data and will call your function `k-means` to perform the clustering. Finally, the results will be displayed in a figure.

[2] Use your k-means function to determine the number of clusters in the dataset `2017-12-stop-and-search.mat`, and provide the location of the centroids. Show how you obtained your result.

[3] Consider the dataset shown in Figure 1. Explain why K-means will not be able to identify the two clusters visible in this dataset.

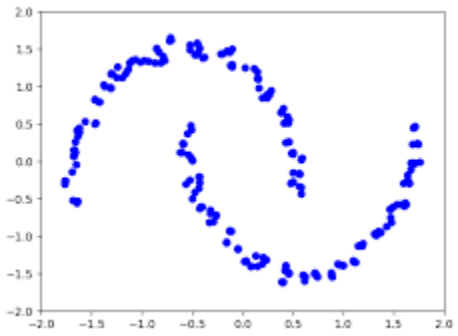

Figure 1 Example of a data set for which K-means does not work.

PCA

PCA can be used to reduce the dimension of a data set. This can be used to visualize the data. In this assignment we will apply PCA to the car-dataset used in the linear regression assignment.

The script `run_PCA_assignment` starts with loading the car data from the file `auto-mpg.mat`. The output column (first column), containing the mpg values, is discarded and will not be used. The goal is to project the 7 dimensional data onto the first two principle components.

[4] Add an implementation of the PCA-algorithm to the script. What are the first two principle components? Use the found principle components to reduce the dataset to two dimensions. The result should be assigned to Z. What percentage of the variance in the data is retained by the first two principle components?

In the second part of the script we use the earlier developed k-means algorithm to cluster the car-dataset, and use the found principle components to visualize the results.

[5] Apply your developed k-means algorithm to the car-data. Create the matrix Z-clustered that contains the projected data and the clustering assignment of each data point. Finally, project the found cluster centroids onto the principle components. What would be a suitable number of clusters for this data set?