Unsupervised learning
Part 1: K-means clustering

By: Noureddin Sadawi, PhD

University of London

Clustering

- ‘Clustering is a technique to divide data into different groups, where the records in each group are similar to one another…
- ‘A goal of clustering is to identify significant and meaningful groups of data…’
- ‘The groups can be used directly, analysed in more depth, or passed as a feature or an outcome to a predictive regression or classification model.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020 p.294).

K-means clustering

- ‘K-means divides the data into K clusters by minimising the sum of the squared distances of each record to the mean of its assigned cluster…’
- ‘ …referred to as the within-cluster sum of squares or within-cluster SS.’
- ‘ K-means does not ensure the clusters will have the same size but finds the clusters that are the best separated.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

K-means clustering

‘K-means was the first clustering method to be developed…

…still widely used, because of the relative simplicity of the algorithm and its ability to scale to large data sets:

KEY TERMS FOR K-MEANS CLUSTERING

Cluster

A group of records that are similar.

Cluster mean

The vector of variable means for the records in a cluster.

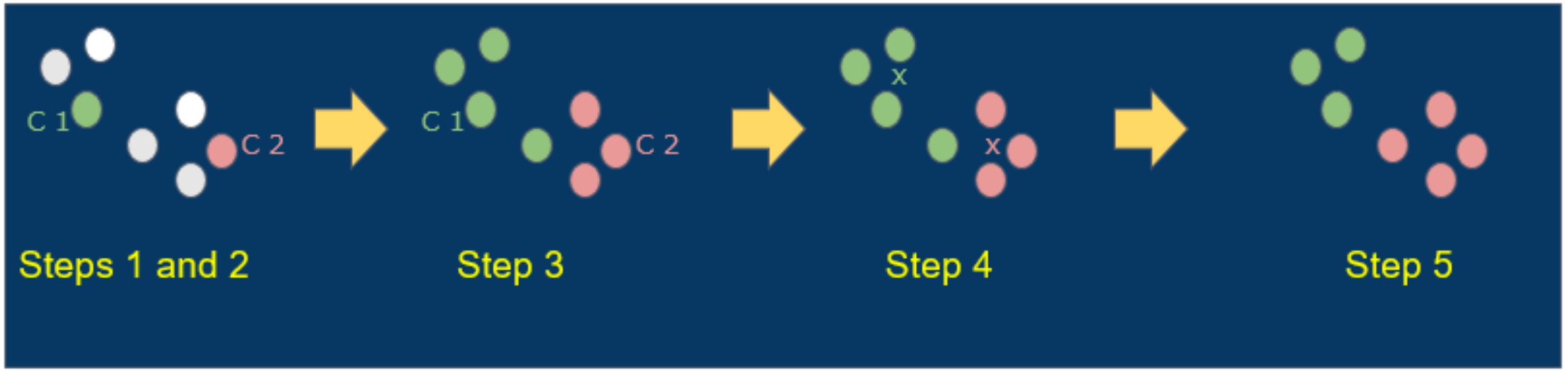
K

The number of clusters.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition 2020).

K-means algorithm

- Step 1: Select the number of clusters *k.*
- Step 2: Randomly select k points from the data as centroids.
- Step 3: Assign each data point to its closest cluster centroid.
- Step 4: Recompute the centroids of newly formed clusters.
- Step 5: Repeat steps 3 and 4 until centroids stop changing.



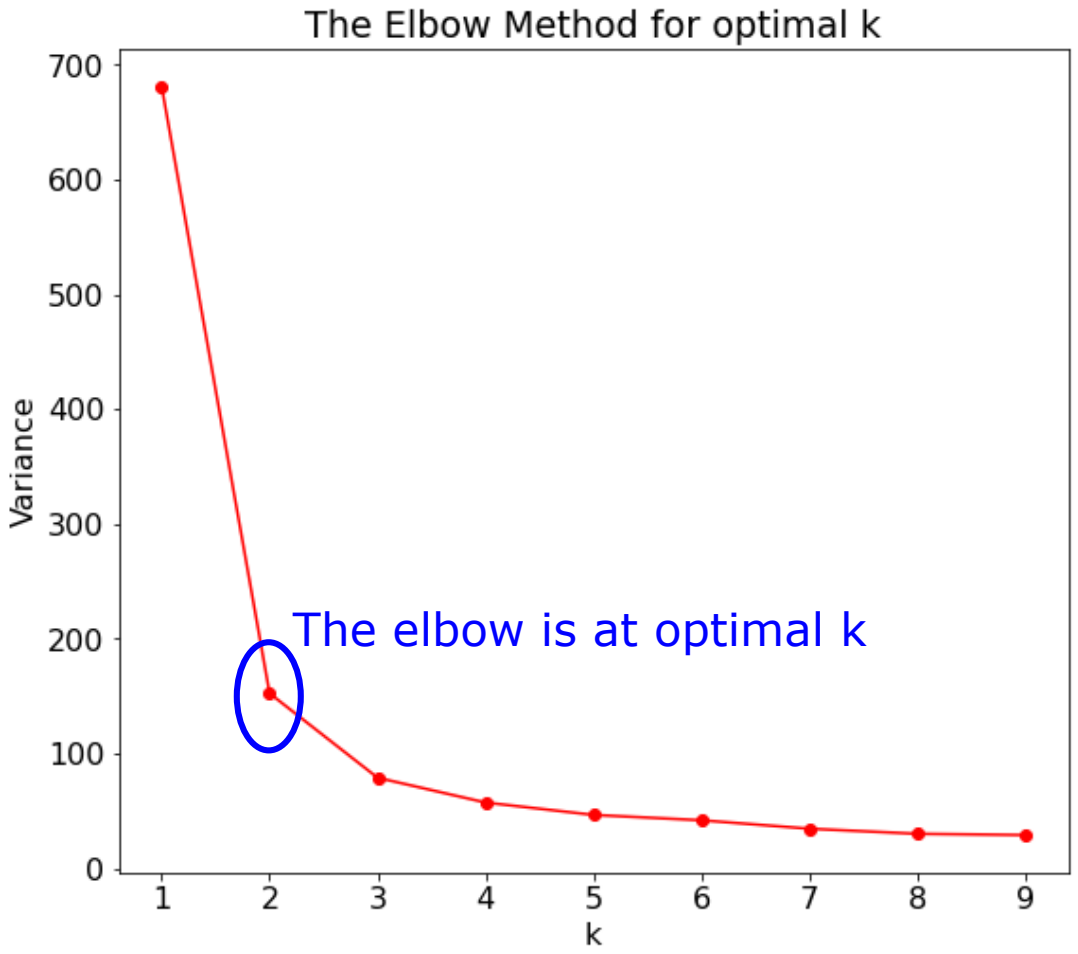
The number of clusters

- ‘The K-means algorithm requires that you specify the number of clusters K.’
- 'Sometimes the number of clusters is driven by the application.’
- 'In the absence of a cluster number dictated by practical or managerial considerations, a statistical approach could be used.’
- 'There is no single standard method to find the “best” number of clusters.’

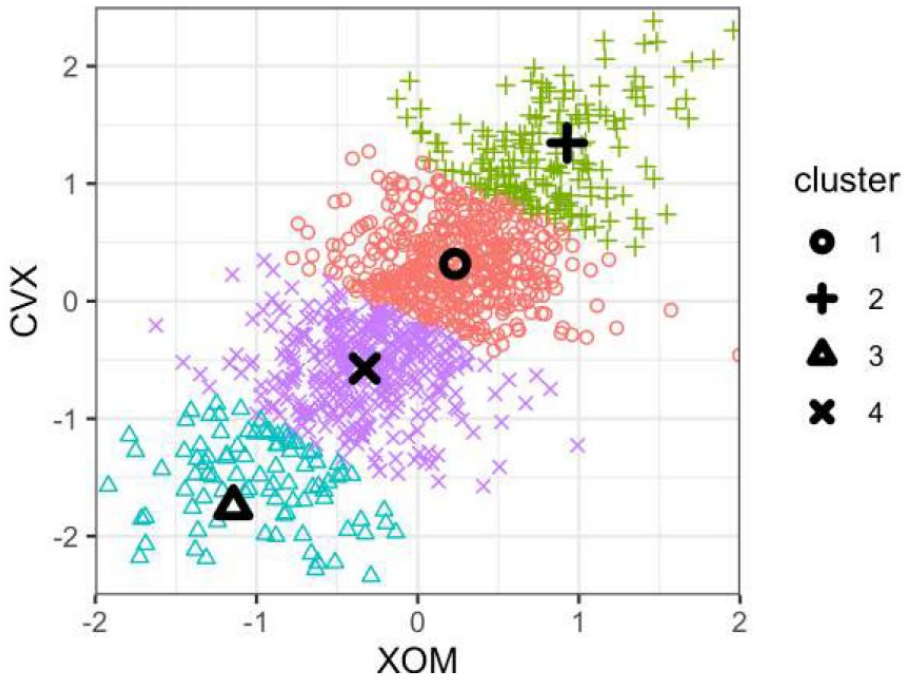
(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

The elbow plot

- Try different values for k.
- Identify when the number of clusters explains ‘most’ of the variance in the data.
- The point of inflection on the curve is at the optimal value of *k*.
- No elbow is an indication that the data that does not have well-defined clusters.



Example



'The clusters of K-means applied to daily stock returns for ExxonMobil and Chevron (the cluster centers are highlighted with black symbols).'

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).