What is K-means Clustering?

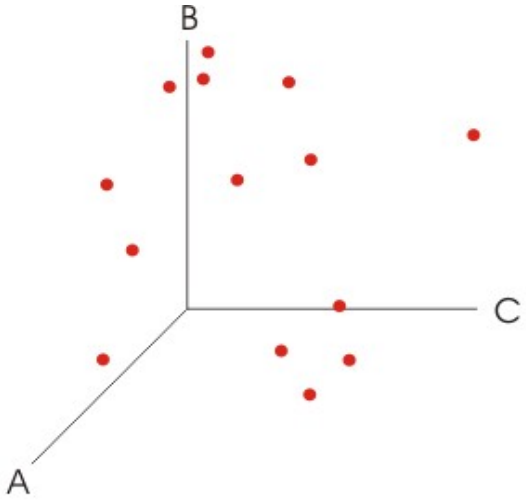
Introduction

K-means clustering is an example of a partitioning (bottom-up) algorithm. Data points are grouped based on similarity, but the degree of homogeneity of the clusters that are formed is dependant largely on how many clusters the algorithm is told to find.

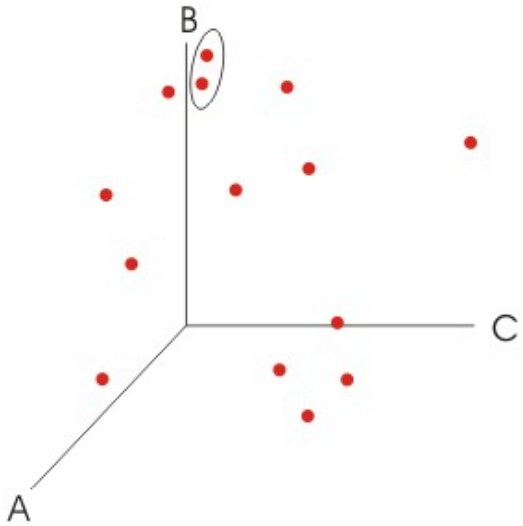
If the algorithm is told to find too few clusters then datapoints (or genes) that are quite different from others in the cluster may be included simply because that was the closest cluster to it - even though it may have been very far away. On the other hand, choosing too many final clusters can mean that more than one cluster may have very similar profiles.

Example

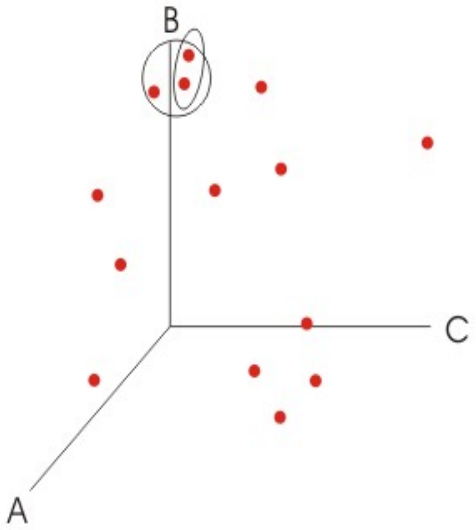
Step 1: Each of the data points are plotted in multi-dimensional space. For this example shown below, the data is plotted in 3-dimensional space.



Step 2: The clustering algorithm then looks for the closest two data points and groups them into a cluster.



Step 3: The next closest data point is then selected - in this case it is grouped into the same cluster. Distances between data points are compared to one another and to the centre (centroid) of each of the previously established clusters. The closest two points, centroids or one point and one centroid are chosen.



This process continues, and additional clusters are formed.

The user tells the algorithm how many clusters should be obtained in the end. Notice that some points may never be included into clusters with the other data points. The end result is dependant on the number of clusters decided upon by the user.