Unsupervised learning

Unsupervised learning is the result of a dataset not having pre-assigned labels or scores for the training data. As a result, unsupervised learning algorithms must first discover any naturally occurring patterns within the dataset. A common method is clustering, where the algorithm groups the dataset into categories with similar features. Another method is principal component analysis, where the algorithm finds ways to compress the data by identifying which features are most useful, to differentiate between the different data points, and discarding the rest.

Clustering analysis

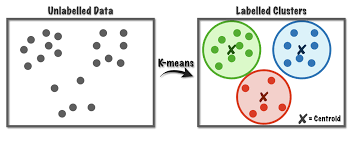
Clustering is a method which forms together objects together into separate groups (clusters) such that those objects are more similar to each other than ones from other groups (clusters).

K-Means algorithm

A clustering algorithm whereby the k number refers to the number of partitions. The k number is set by the used. The centroid is essentially the ‘mean’ of the data points which represents the centre of the cluster.

How K-means works

To process the data, the K-means algorithm starts with a group of randomly selected centroids, which are used as the starting points for every cluster. Then iterative calculations are performed to optimise the positions of the centroids. The creating and optimising of clusters may stop when either the centroids have been stabilised or the defined number of iterations have been achieved.



This method can be used to group the data points from the respiratory dataset, into a specified number of clusters (k=2 is recommended). Once the data points are partitioned into clusters, future labels may be assigned to the formed clusters. The respiratory dataset can also potentially be used as a test dataset which predictions can be made onto.

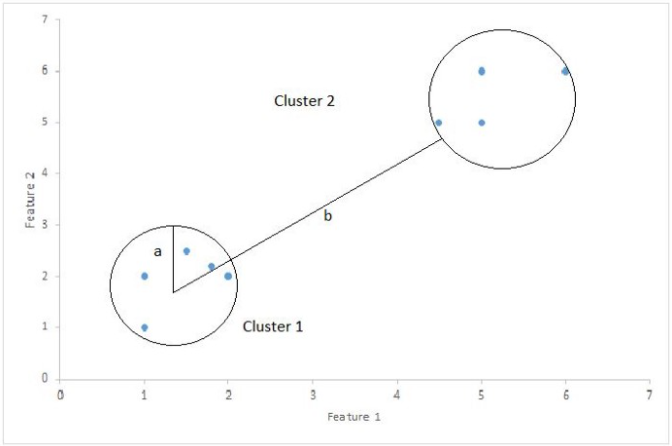
Measure goodness of fit

Silhouette Coefficient or Silhouette score is used to calculate how well the clustering is fitted. Its values range from -1 to 1.

1: Means clusters are well apart from each other and clearly distinguished.

0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.

-1: Means clusters are assigned in the wrong way.



Silhouette score = (b-a)/max(a, b)

a: average intra-cluster distance (average distance between each point within the cluster)

b: average inter-cluster distance (average distance between all clusters)