Clustering is a way that classify the raw data reasonably and searches the hidden patterns that may exist in datasets. A cluster is a collection of data points where each observation is similar to other observations in the same cluster & dissimilar to observations in other clusters

Pseudo code of K-means

1) Select K points as the initial centroids.

2) Repeat Until convergence {

Form K clusters by assigning all points to the closest centroid.

Recompute the centroid of each cluster }

K-means is a numerical, unsupervised, non deterministic, iterative method.

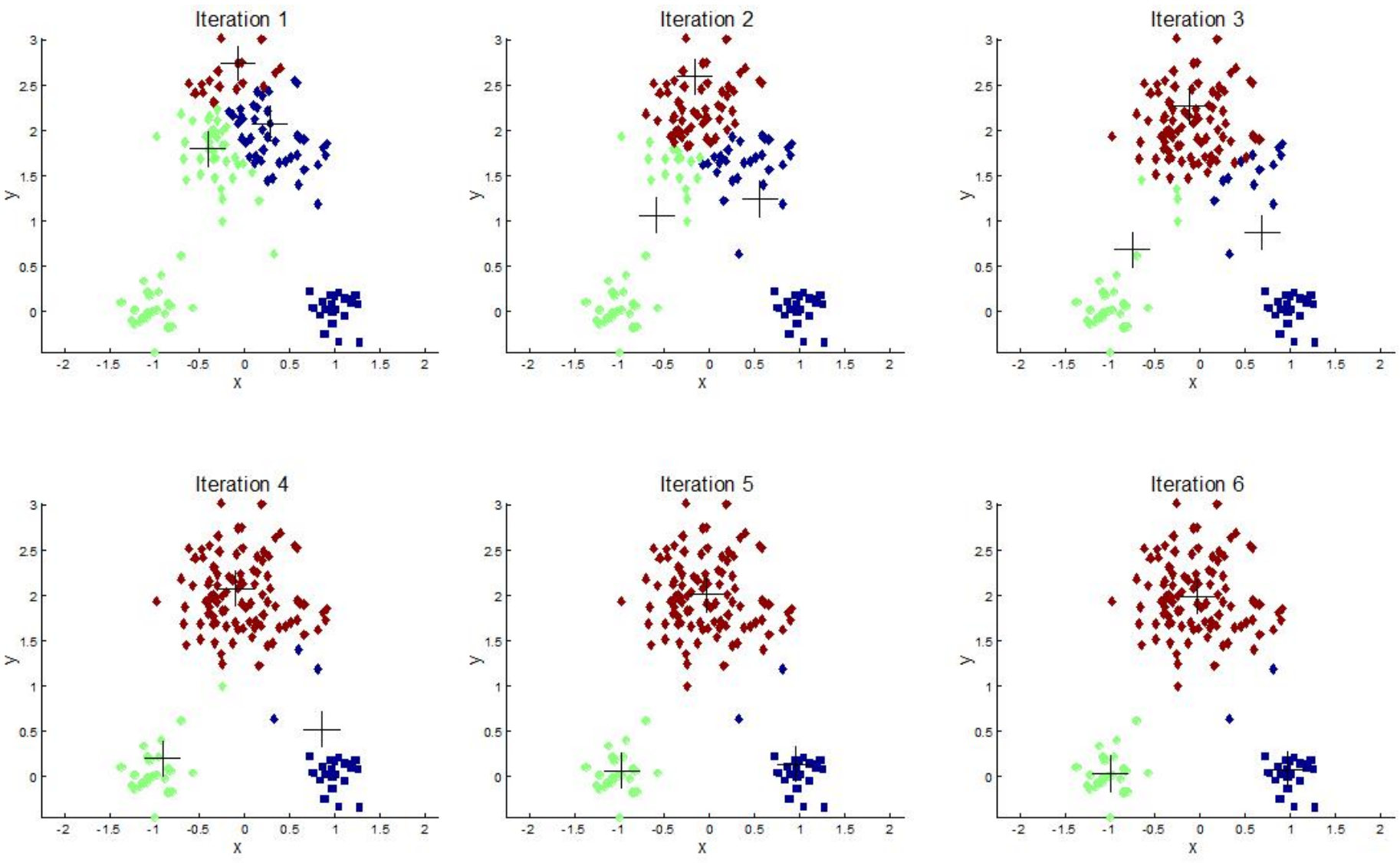
1) Partitional Clustering approach

2) Each cluster is associated with a centroid

3) Each point is assigned to the cluster with the closest centroid

4) Number of clusters, K, must be specified.

Convergence of K-means Algorithm



Details of K-means

1) Initial centroids are often chosen randomly.

-*Clusters produced vary from one run to another*

3) 'Closeness' is measured by Euclidean distance, cosine similarity, correlation

4) K-means will converge for common similarity measures mentioned above.

Limitations of K-means

K-means has problems when clusters are of differing

Sizes

Densities

Non-globular shapes

K-means has problems when the data contains outliers.