Clustering

Clustering is a statistical analysis of some data, where there is no predetermined classes to fit the data onto. Instead the data is compared to itself in search for either similarities or patterns. In contrast to supervised learning with classification, the result doesn't tell anything about the data, only the relation between the data objects within the data.

Unsupervised Learning

Clustering is unsupervised. This means that there is no external knowledge to guide or to supervise the clustering process.

- We cannot learn rules to sort objects into clusters.
- We do not know how the clusters are characterized.
- We do not know how many clusters there are.
- There is no unique criterion to judge on the quality of a derived clustering solution (evaluation).

Categories of Clustering Approaches

Partitioning

Model: Cluster are compact sets of object (points).

Parameter: (usually) number k of clusters or distance measure.

Looks for at flat partitioning into k cluster with maximal compactness.

Partitional clustering partitions datasets into k clusters, typically minimizing some cost function (compactness criterion). Central assumptions for approaches in this family are typically:

- Number of k clusters.
- Clusters are characterized by their compactness.
- Compactness measured by some distance function (e.g., distance of all objects in a cluster representative is minimal).
- Criterion of compactness typically leads to convex or even spherically shaped clusters.

Typically in search for the global optimal solution, we optimize for the local optimal. The clusters are either represented by their centroid, medoid or by some gaussian distribution model.

Density Based

Model: Cluster are areas of high density, separated by areas of low density.

Parameter: Minimal density in some cluster or distance measure.

Looks for flat partitioning into clusters exceeding some minimal density.

Hierarchical

Model: Compactness, density, ...

Parameter: Distance measure for points and for cluster.

Looks for a hierarchy of clusters (e.g. given as a tree), joins most similar clusters at a given level of the hierarchy.

Flat clusters can be derived by cutting the tree on some level.

K-means Clustering

Place centroids at random positions and iterate through with distance calculations at each iteration.