The Learning Machine

Statistics Machine Learning Applied ML Contribute

Clustering

Differences between clustering algorithms

Clustering is the task of dividing data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

Clustering Methods:

1. **Density-Based Methods:** These methods consider the clusters as the dense region having some similarity and different from the lower dense region of the space. These methods have good accuracy and the ability to merge two clusters. Example DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure) etc.
2. **Hierarchical Based Methods:** The clusters formed in this method forms a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two category
   --> *Agglomerative* (bottom-up approach)
   --> *Divisive* (top-down approach).
   Examples CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies) etc.
3. **Centroid (Partitioning) Methods:** These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter example K-means, CLARANS (Clustering Large Applications based upon randomized Search) etc.
4. **Distribution Methods:** These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution (For example: Normal, Gaussian). These models often suffer from overfitting. A popular example of these models is Expectation-maximization algorithm which uses multivariate normal distributions.

Density-Based Models

Clustering Model	Pros	Cons
DBSCAN (Density-Based Spatial Clustering of Applications with Noise)	• Clusters of arbitrary shape and size • Robust to noise • Does not need a pre-set number of clusters • Deterministic	• Requires connected regions of sufficiently high density • Data sets with varying densities are problematic
Mean-Shift Clustering	• Does not need a pre-set number of clusters • Simple to understand	• Selection of the window size/radius “r” can be non-trivial

Hierarchical Models

Clustering Model	Pros	Cons
Hierarchical Clustering	• The optimal number of clusters can be obtained by the model itself • Practical visualisation with the dendrogram	• Not appropriate for large datasets
BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies)	• Designed for clustering a large amount of numerical data • Works well only for spherical clusters	• Can handle only numeric data • Sensitive to the order of the data records

Centroid Methods

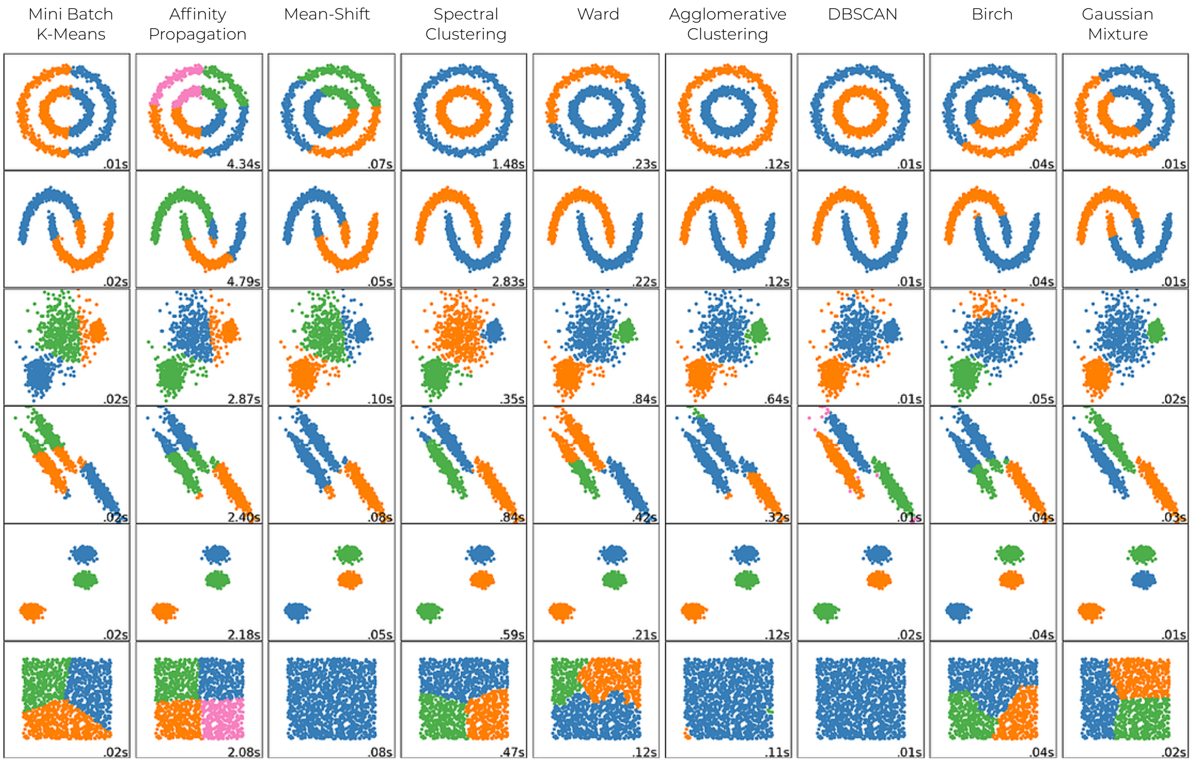
Clustering Model	Pros	Cons
K-Means (MiniBatch K-Means, K-Means ++)	• Simple to understand • Easily adaptable • Works well on small or large datasets • Fast, efficient and performant	• Need to choose the number of clusters • Assumes the clusters as spherical, so does not work efficiently with complex geometrical shaped data(Mostly Non-Linear) • Hard Assignment might lead to mis grouping.
Affinity Propagation	• Does not need a pre-set number of clusters • Works well on small or large datasets • Clusters of arbitrary shape and size • Does better clusters than K-Means	• Much slower than K-Means

	• Does better clusters than K-Means	
Spectral Clustering	• Elegant, and well-founded mathematically • Works quite well when relations are approximately transitive • Excellent quality under many different data forms	• Not appropriate for very noisy datasets • Much slower than KMeans

Distribution Methods

Clustering Model	Pros	Cons
GMM Expectation-Maximization using Gaussian Mixture Models	• A lot more flexible in terms of cluster covariance than K-Means • Have multiple clusters per data points, I.e GMMs support mixed membership • Does not assume clusters to be of any geometry. Works well with non-linear geometric distributions as well. • Does not bias the cluster sizes to have specific structures as does by K-Means (Circular).	• Slow convergence • Inability to provide estimation to the asymptotic variance-covariance matrix of the maximum likelihood estimator (MLE) • Difficult to interpret.

Overall Model Performance


Mini Batch K-Means
Affinity Propagation
Mean-Shift
Spectral Clustering
Ward
Agglomerative Clustering
DBSCAN
Birch
Gaussian Mixture
.01s
4.34s
.07s
1.48s
.23s
.12s
.01s
.04s
.01s
.02s
4.79s
.05s
2.83s
.22s
.12s
.01s
.04s
.01s
.02s
2.87s
.10s
.35s
.84s
.64s
.01s
.05s
.02s
.02s
2.40s
.08s
.84s
.42s
.32s
.01s
.04s
.03s
.02s
2.18s
.05s
.59s
.21s
.12s
.02s
.04s
.01s
.02s
2.08s
.08s
.47s
.12s
.11s
.01s
.04s
.02s