When should you use Minibatch

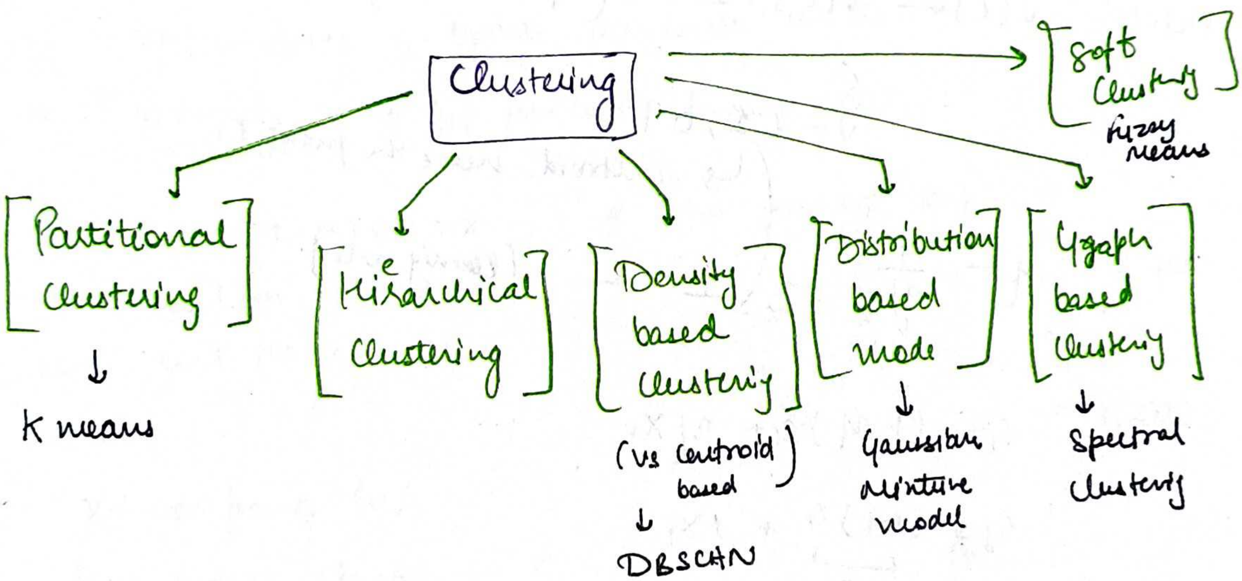
→ Speeds / bigger datasets

→ Online ML

downside

↳ accuracy low

Types of Clustering



Partitional Clustering

1. **Basic Concept:** Partitioning clustering algorithms divide a dataset into a set of non-overlapping subgroups or clusters, where each data point belongs to exactly one cluster.

2. **Example:** The most famous partitioning clustering algo is K-means. It assign data points to cluster in such a way that each point belongs to the cluster with the closest mean, which serves as a prototype of the cluster.

3. **Defining Number of Cluster:** A key requirement is pre-specifying the number of clusters (K). The selection of K-means, the process involves repeatedly assigning points to the nearest cluster centroid and then recalculating the centroids.

4. **Iterative Process:** These algorithm typically use an iterative refinement technique. For instance, in K-means, the process involves repeatedly assigning points to the nearest cluster centroid and then recalculating the centroids.

5. **Objective Function Optimization:** They aim to optimize an objective function, such as minimizing the total within cluster variance or the sum of squared distance betn data points and their respective cluster centroids.

6. **Suitability for Certain Data Shapes:** Partitioning methods are most effective when clusters are spherical or globular in shape. They assume homogeneity in cluster shapes and sizes.

7. **Sensitivity to Initial Condition**: These algo can be sensitive to the initial starting condition (like initial cluster centroid in k-means). Different initializations can lead to different clustering results.

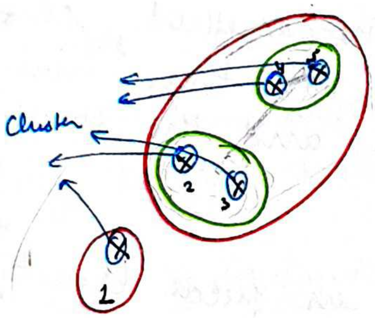
8. **Handling of outliers**: Partitioning algo can be influenced by outliers, as these can significancy skew the mean or centroid of a cluster.

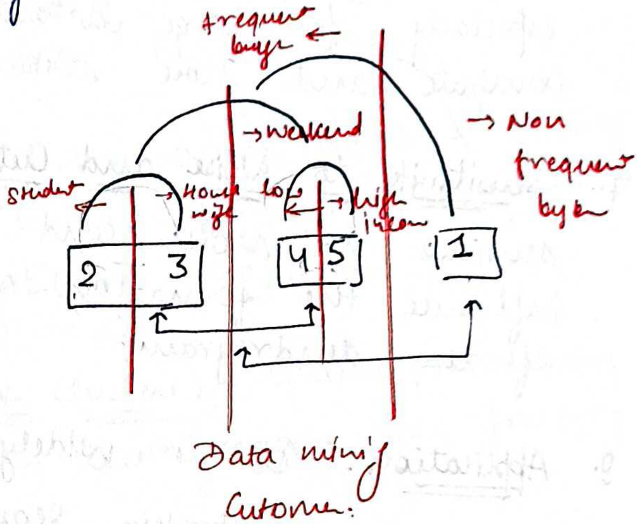
9. **Scalibility and Efficiency**: They are generally more scalable and efficient for larger datasets compared to hierarchical clustering, making them suitable for many practical applications.

10. **Use cases and Limitations**: While widely used in various fields like market research, pattern recognition and image processing these algo have limitations in handling non spherical clusters, varying cluster sizes, and datasets. Advanced versions and variation of partitioning algo have been developed to address some of these limitations.

Hierarchical Clustering

1. **Nature of Clustering:** Hierarchical clustering builds a hierarchy of clustering either by successively merging smaller cluster into larger ones (agglomerative approach) or by successively splitting larger cluster into smaller ones (divisive approach).





2. **No need to Specify Number of Clustering:** Unlike partioning algo like k-means, hierarchical clustering does not require pre-specifying the number of cluster. The Number of cluster can be determined by analyzing the dendrogram.

3. **Dendrogram Visualization:** It provides a tree like diagram called a dendrogram, which is a visual representation of the clustering process showing the order of clustering combination and the distance at which cluster are merged.

4. **Distance Metrics and Linkage Criteria:** Hierarchical clustering uses various distance metrics (like Euclidean or Manhattan distance) and linkage criteria (like single linkage, complete linkage, average linkage, and Ward's method) to decide which cluster to merge or split.

5. **Flexibility in Indentifying cluster shapes**: Hierarchical clustering can identify clusters with various shapes and sizes, unlike partitioning method that generally assume spherical cluster.

6. **Computational complexity**: It is generally more computationally instensive than partitioning methods, especially for large datasets, due to the need to compute and store distance betn all pairs of points.

7. **Sensitivity to Noise and Outliers**: The method can be sensitive to noise and outliers, as these can influence the formation of clusters and the structure of the dendrogram.

8. **Application**: It is widely used in fields like biology (for gene and protein sequencing), social science and linguistics and is particularly useful for exploratory data analysis where understandiy the hierarchical relationship between objects is important.

Density Based Clustering

Principle: Density-based clustering groups data points based on the density of data points in a region. It defines cluster as areas of higer density separated by areas of low density. The algo identifies cluster as region where data point are densely packed together, with areas of low density or noise between them.

Example: DBSCAN is one of the most popular density-based clustering algo. It is known for its efficiency and ability to find cluster of arbitrary shapes.

No Need to specify Number of Clusters: Unlike partitioning method, density-based clustering doesn't require pre-specifying the number of clusters.

Handling Noise and Outlier: It is robust to outlier and noise, as these are typically not part of the dense regions that form clusters.

Ability to find Arbitrary shapes: Density-based clustering can discovering cluster of arbitorary shapes, unlike methods like k-means which are biased towards spherical cluster.

Parameter Sensitivity: The performance of these algo is sensitive to the input parameters, like the radius of neighborhood (ε) and the minimum number of points required to form a dense region (MinPts in DBSCAN)

Scalability Issue: Some density-based algo is sensitive to the input parameters, like the radius of may struggle with very large datasets due to computational and memory constraints.

Applications: Widely used in fields such as anomaly detection, geospatial data analysis (like identifying geographic regions of interest) and image processing, especially where the shape of the cluster is not known in advance or the data contains noise.

Distribution/Model Based Clustering

1. **Statistical Distribution Model:** The central concept of distribution-based clustering is that data points in a cluster follow a certain statistical distribution, most commonly Gaussian or normal distributions.

2. **Parameter Estimation:** These algo focus on estimating the parameter (like mean, variance) of the assumed distribution for each cluster. The fit of these parameter to the actual data determines the quality of the clustering.

3. **Expectation-Maximization (EM) Algo:** A key algo used in distribution-based clustering is EM, which alternate both assigning data points to the most likely distribution parameters to maximum data fit (Maximization step).

4. **Handling of Complex Cluster Shapes:** Unlike methods such as K-Means, distribution based clustering can identify cluster of various shapes and sizes, making it more flexible in handling real-world data complexities.

5. <u>Computational Intensity</u>: The process of estimating distribution parameter and assigning data points can be computational demanding, especially for large datasets and when the number of feature (dimension) is high.

6. <u>Handling of Outliers</u>: These methods can be more robust to outliers, as outliers are less likely to significantly affect the parameter of the overall distribution.

7. <u>Scalability Issue</u>: While effective for small to medium sized dataset, scalability to very large dataset can be challenging due to the computational complexity of the algo and the need for more sophisticated optimization technique.

8. <u>Soft Clustering</u>: In soft clustering each data point is associated with a prob distribution across different cluster, indicating the degree of belonging to each cluster. This is in contrast to hard clustering, where each data point is assigned to exactly one cluster.