BASIC ISSUES IN CLUSTERING

- The cluster membership may change over time due to dynamic shifts in data.
- Handling outliers is difficult in cluster analysis.
- Clustering struggles with high-dimensional datasets.
- Multiple correct answers for the same problem.
- Evaluating a solution's correctness is problematic.
- Clustering computation can become complex and expensive.
- Assumption of equal feature variance in distance measures.
- Struggle with missing data (columns and points)

Clustering Challenges from Data Changing

Clustering struggles with time series data since you will have to pick a point in time to do your analysis. Data points could have belonged to any number of clusters at any time, and doing one analysis at one point in time may be misleading.

Clustering Challenges from Outliers

Since most clustering algorithms use distance-based metrics, outliers in our datasets can completely change the clustering solution. The presence of just one outlier will cause cluster centroids to update, possibly moving points that should exist in one cluster to another.

Clustering Challenges from High Dimensional Data.

High-dimensional data affects many machine learning algorithms, and clustering is no different. Clustering high-dimensional data has many challenges. These include the distance between points converging, the output becoming impossible

to visualize, correlation skewing the location of the points, and the local feature relevance problem.

Clustering Challenges from Multiple Solutions.

Many clustering algorithms will generate random centroids to start the computation. This methodology creates a scenario where different solutions can be found depending on how many clusters are chosen and where the centroids are initially placed. Since many local solutions can be found, the reproducibility of your analysis suffers.

Clustering Challenges in Evaluating Accurate Solutions

Since labelled data is missing in clustering algorithms, evaluating solutions becomes tough. This is because we have nothing to compare our clusters against. There are tactics to evaluate solutions from clusters (like the silhouette method), but utilizing distance in our metrics can cause problems.

Clustering Challenges Due To Computation Limits.

In situations where there are very large data sets or many dimensions, many clustering algorithms will fail to converge or come to a solution. For example, the time complexity of the K-means algorithm is O(N^2), making it impossible to use as the number of rows (N) grows.

Clustering Challenges Based On Initial Clustering Assumptions.

Clustering algorithms that only utilize distance metrics assume that all features are equal in terms of relevance to the problem. This creates problems, as some dimensions are much more relevant to our specific situation than others, potentially weakening our analysis.

Clustering Challenges from Missing Values and Data.

In clustering, missing values and dimensions can completely change the solution. While this problem exists in all subsets of machine learning, clustering is highly sensitive to it. This is due to the distance nature of clustering algorithms that recomputed centroids based on available data.