

SPPU-BE-COMP-CONTENT - KSKA Git

ASSIGNMENT-4

Q1

| FEATURE | HIERARCHICAL CLUSTERING | K-MEANS CLUSTERING |
|----------------------|--|--|
| Type | Agglomerative or divisive | Partition based |
| Number of clusters | Does not require specifying clusters upfront | Requires specifying k in advance. |
| cluster shape | can detect clusters of arbitrary shape | Assume - clusters are roughly spherical |
| Scalability | Poor for very large datasets | Good for large datasets |
| Algorithm complexity | Higher $O(n^2 \log n)$ for naive implementations | Lower $O(n * k * t)$ where t is iterations |

Q2

- K-means iteratively updates clusters' centroids & assignment until a stopping condition is met. common stopping criteria include.
1. Convergence of centroids: Stop when centroids do not change significantly b/w iterations.
 2. Minimal change in objective function: Stop when the total within cluster sum of squares or distortion function changes very little b/w iterations.

SPPU-BE-COMP-CONTENT - KSKA Git

3. Maximum iterations:

Stop after pre designed no. of iterations, even if convergence hasn't occurred.

4. Stable cluster Assignments

Stop if no points change their cluster assignment in current iterations.

Q3

1. Feature Scaling: Standardize or normalize features to ensure equal weighing

ex: standardScaler in python.

2. Handling missing values: Impute missing data because K means cannot handle NaNs.

3. Removing outliers: outliers can distort cluster centroids consider removal or capping

4. Categorical variables: convert to numeric form (one hot encoding) is needed.

5. Dimensionality Reduction:

Use PCA or t-SNE for high dimensional data to reduce noise & speed up computation.