**Clustering Technique: k-means clustering**

**Topic:** k-mean clustering

**About k-mean clustering:**

k-means clustering is one of the best ways of doing data mining is by utilizing clusters. It is a technique used to uncover categories. “k” represents the number of categories identified, with each category’s average (mean) characteristics being appreciably different from that of other categories.

**How does k-mean clustering work:**

“It is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are: the centroids of the K clusters, which can be used to label new data & Labels for the training data (each data point is assigned to a single cluster).”

* Source: <https://www.datascience.com/blog/k-means-clustering>

**When is k-means clustering used :**

The K-means clustering is used to find organized groups not having been explicitly labeled in the data. It can be used to confirm business assumptions about what types of groups exist or to identify unknown groups in complex data sets. Once the algorithm has been run and the groups are defined, any new data can be easily assigned to the correct group.

There are various segments where this technique can be utilized:

1. Sorting sensor measurements - Detect activity types in motion sensors, Group images and Separate audio.
2. Detecting bots or anomalies: Separate valid activity groups from bots and Group valid activity to clean up outlier detection
3. Behavioral segmentation: Segment by purchase history, Segment by activities on application, website, or platform and Define personas based on interests.
4. Inventory categorization: Group inventory by sales activity and Group inventory by manufacturing metrics

**What kind of data is required by the technique :**

In general, k-means clustering requires numerical data with a relatively lower number of dimensions. One would use numerical data (or categorical data converted to numerical data with other numerical features scaled to a similar range) because mean is defined on numbers. One cannot calculate the mean of names of lab tests or fruits or the car one drives.

**Resources required :**

* Libraries required to perform k-means are car, summarytoold, Rcmdr and anymore for data updations.
* “kmeans” itself is a function in R to produce the results for the clustering technique analysis.

The algorithm randomly assigns each observation to a cluster, and finds the centroid of each cluster. Then, the algorithm iterates through two steps:

* Reassign data points to the cluster whose centroid is closest.
* Calculate new centroid of each cluster.

**Tutorials:**

**Link:** https://www.youtube.com/watch?v=sAtnX3UJyN0

Influxity. (2013, September 29). How to Perform K-Means Clustering in R Statistical Computing. Retrieved from youtube.com: <https://www.youtube.com/watch?v=sAtnX3UJyN0>

**Link:** <https://www.datacamp.com/community/tutorials/k-means-clustering->r

Jaiswal, S. (2018, March 18). K-Means Clustering in R Tutorial. Retrieved from Data Camp: <https://www.datacamp.com/community/tutorials/k-means-clustering-r>

**Word Count -- 505**