Prepared by: Mohammad Rkieh
Instructor: Dr. Yasser Alginahi

University of Windsor

Fuzzy Clustering

Intro

Unsupervised learning is a machine learning approach where algorithms are trained on unlabeled data without specific target outputs. Its primary purpose is to discover patterns, structures, or relationships within the data autonomously. Common applications include clustering, dimensionality reduction, and anomaly detection.

Clustering is a machine learning technique that involves grouping similar data points together into distinct categories (clusters). It is used to discover underlying patterns or structures in a dataset, making it valuable for tasks such as customer segmentation and image categorization.

Limitation of Hard Clustering

Hard clustering assigns each data point to a single cluster, making it exclusive to one group. This results in a clear-cut, discrete partition of data but may not capture underlying nuances in the data. (binary)

Hard clustering is computationally less complex and typically used when data points are distinctly separable.

Example of problem: Classifying animals at a zoo. Instead of assigning each animal to a single exhibit, it allows for the possibility that an animal might belong to multiple areas, like a panda that could be part of both the "Bears" and "Endangered Species" exhibits, reflecting the idea that some animals don't fit into just one category neatly. This approach provides a more flexible and nuanced way to describe the zoo's animal population.

Soft Clustering

Soft clustering is suitable for scenarios where data points overlap multiple clusters. It allows data points to belong to multiple clusters simultaneously. Each data point is associated with a probability distribution over clusters, providing a more fine-grained distinction representation of data relationships.

Soft clustering provides richer information about data structures and is often employed in customer segmentation, where customers may exhibit behaviors across multiple segments.

Fuzzy C-Means (FCM) algorithm

"Fuzzy" is a term often used to describe something that is unclear, vague, or imprecise.

Fuzzy C-Means is the most popular fuzzy clustering algorithm used to partition data into clusters with varying degrees of membership. Based on the traditional K-Means clustering algorithm by allowing data points to belong to multiple clusters with membership degrees ranging between 0 and 1. Fuzzy Clustering is ideal for when data overlaps multiple clusters.

Applications

FCM can be used to identify patterns in large datasets by grouping similar data points together.

FCM can be used to segment customers based on their preferences and purchasing behavior, allowing for more targeted marketing campaigns.

FCM can be used to diagnose diseases by grouping patients with similar symptoms together.

FCM can be used to segment images by grouping pixels with similar properties together, such as color or texture.

Advantages

1. Flexibility: FCM allows for overlapping clusters, which can be useful when the data has a complex structure or when there are ambiguous or overlapping class boundaries.
2. Robustness: Fuzzy clustering can be more robust to outliers and noise in the data, as it allows for a more gradual transition from one cluster to another.
3. No Preset Number of Clusters: Unlike K-Means, FCM does not require specifying the number of clusters in advance. It can adapt to the data and determine the optimal number of clusters based on the input.

Disadvantages

1. Complexity: Fuzzy clustering methods can be slower and require more computing power compared to regular clustering methods because they must do extra work to figure out how much each data point belongs to multiple clusters.
2. Model selection: Choosing the right number of clusters and membership functions can be challenging and may require expert knowledge of the field or trial and error (inefficient).
3. Determining the optimal number of clusters in fuzzy clustering can be challenging. Various methods, such as the Fuzzy C-Means (FCM) algorithm and validation metrics like the Dunn index, can help you make this decision.

Dunn Index

The Dunn Index is a metric used to evaluate the quality of clustering in a dataset. It helps determine how well the data points are grouped into clusters.

- Max_intra-cluster_distance: This represents the maximum distance between any two points within the same cluster. It measures the compactness of individual clusters. A smaller max_intra-cluster_distance indicates that the points within a cluster are closer to each other, which is generally desirable in clustering.
- Min_inter-cluster_distance: This represents the minimum distance between any two points belonging to different clusters. It measures the separation between clusters. A larger min_inter-cluster_distance indicates that the clusters are well-separated from each other, which is also generally desirable in clustering.

In summary, a higher Dunn Index is associated with better clustering quality, where the clusters are tight and well-separated. Conversely, a lower Dunn Index suggests that the clusters are either not compact enough within themselves or not well-separated from each other, indicating poorer clustering results.

$$DI = \frac{min\_inter{-}cluster\_distance}{max\_intra{-}cluster\_distance}$$

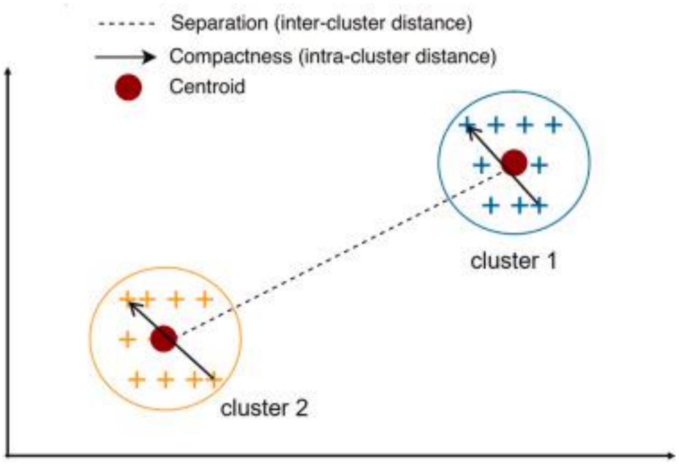
Separation (inter-cluster distance)
Compactness (intra-cluster distance)
Centroid
cluster 1
cluster 2

Figure 1 Dunn Index Visual Representation

Resources

Video presentation: https://www.youtube.com/watch?v=65G5xkWXTTk

References

[1] Rajesh Khadka, “Machine Learning Types #2 - Towards Data Science,” *Medium*, Sep. 07, 2017. https://towardsdatascience.com/machine-learning-types-2-c1291d4f04b1

[2] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer Science & Business Media, 2013.

[3] Witold Pedrycz, *An Introduction to Computing with Fuzzy Sets*. Springer Nature, 2020.

[4] “Day 71 - Fuzzy C-Means Clustering Implementation,” *www.youtube.com*. https://www.youtube.com/watch?v=W-3ZYGmLJ-4&list=PPSV (accessed Oct. 17, 2023)

[5] “Dunn index and DB index - Cluster Validity indices | Set 1,” *GeeksforGeeks*, May 09, 2019. https://www.geeksforgeeks.org/dunn-index-and-db-index-cluster-validity-indices-set-1/

[6] “Fuzzy c-means,” *www.youtube.com*. https://www.youtube.com/watch?v=zr50h_91gOw&ab_channel=jeff (accessed Oct. 20, 2023).

[7]“Day 70 - Fuzzy C-Means Clustering Algorithm,” *www.youtube.com*. https://www.youtube.com/watch?v=VhYt7nxOKKs&ab_channel=DataSciencewithSharan (accessed Oct. 20, 2023).

[8]“Fuzzy C Means Clustering Algorithm Solved Example | Clustering Algorithm in ML & DL by Mahesh Huddar,” *www.youtube.com*. https://www.youtube.com/watch?v=X7co6-U4BJY&t=230s&ab_channel=MaheshHuddar (accessed Oct. 20, 2023).