Ilana ivan: 205634272
Agathe Benichou: 345428312S
Due: July 16th, 2023
CS3946: Advanced Machine Learning

Final Project: Explainable AI

The field of machine learning is dominated by a vast array of complex algorithms such as clustering algorithms who are able to group together similar data points. These algorithms, such as k-means and k-medians, often present significant challenges. The lack of explainability in their decisions leave humans in the dark about the underlying reasons for specific data clustering. In the paper, Explainable k-Means and k-Medians Clustering, the authors present the need for explainability in these traditional clustering algorithms. They introduce the Iterative Mistake Minimization (IMM) algorithm as an innovative approach to achieve explainability by minimizing the mistakes made during clustering, independent of the datasets dimensionality or the number of points.

Background

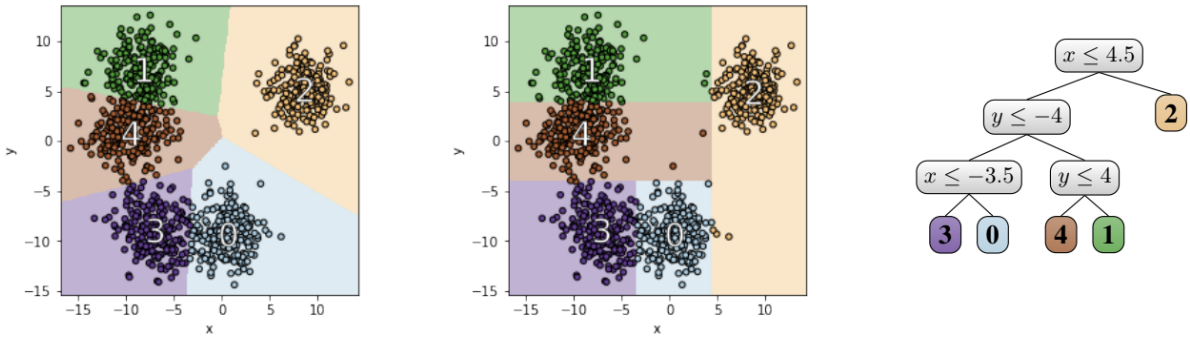
Clustering is a technique used to categorize similar data points into distinct groups in order to uncover patterns within complex datasets. Algorithms such as k-Means and k-Medians have the ability to extract the inherent structure of a dataset, segmenting it into distinct and interpretable groups. The k-Means algorithm is an iterative approach that divides N data points into k distinct clusters. Its objective is to minimize the sum of squared distances between each point and its assigned centroid. The quality of the clustering solution is quantified by the cost function, within-cluster sum of squares. Minimizing this cost function allows k-Means to find optimal cluster assignments. A variant of k-Means is k-Medians, which determines the centroid using the median value within each cluster. This approach is more robust to outliers, making it suitable for datasets that may contain outliers.

However, k-Means and k-Medians are not inherently explainable since the clusters are determined by all features (coordinates) of the data. The identification of cluster groups is influenced by numerous complex features, making it challenging to explain the outcomes generated by many clustering algorithms. However, there is a growing interest in incorporating explainability into traditional k-means clustering. As model complexity rises, interpretability plummets, leading to a black box phenomenon where a models decision making process becomes elusive and opaque. Research has shown that finding optimal clusters is non trivial and NP-hard (Aloise et al., 2009; Dasgupta, 2008). This lack of concise explanations for cluster assignments has led to the development of methods like LIME (Local Interpretable Model-Agnostic Explanations). LIME approximates the decision boundary of a complex model locally using a simple model that is more easily understood. Nonetheless, LIME has limitations, as it does not directly provide insights into the dataset and its explanations depend on the model being used. The ultimate goal is to develop more principled approaches that offer a comprehensive understanding of complex models and their decision-making processes.

In high-dimensional data, traditional clustering methods can result in clusters characterized by intricate and intertwined feature relations that are challenging to unravel. Seeking simplicity and interpretability without compromising clustering quality is a crucial consideration. Traditional approaches like dimensionality reduction or feature selection may not necessarily enhance interpretability since they may overlook important features or fail to carry over the same intuitive meaning. Striking the right balance between interpretability and cost is essential in constructing effective and understandable clustering models. One approach involves constructing a threshold tree with k leaves that partitions the dataset into clusters. Each internal node of the tree represents a question about a specific feature and threshold value, allowing us to understand how different features contribute to the cluster assignments. By integrating this explainable approach, we can shed light on the underlying factors that determine cluster identities and enhance our understanding of the clustering process.

IMM Algorithm

The Iterative Mistake Minimization (IMM) algorithm is a procedure designed for multi-cluster scenarios (where k > 2). It leverages the concept of a mistake, which is when a data point within one group is actually closer to the center of another group. IMM introduces an approximation method that operates independent of the data dimensionality and number of points. It uses dynamic programming to efficiently resource over all possible threshold cuts. IMM begins by running a traditional clustering algorithm, after which each data point is labeled according to its corresponding cluster. A top down threshold tree is constructed, partitioning the data based on thresholds and minimizing mistakes during the process. The end result is a tree with k leaves, where each leaf represents a cluster.



(a) Optimal 5-means clusters (b) Tree based 5-means clusters (c) Threshold tree

The algorithm delivers efficient runtime performance, robustness in cluster identification and proxies a provable approximation to optimal clustering. It has a runtime complexity of $O(k * d * n * log(n))$ which is contingent on the number of clusters (k), the dimensionality of the data (d) and the total number of points (n). IMM offers a mathematically provable guarantee of performance, being a $O(k^2)$ approximation to the optimal clustering (the ideal partitioning of the dataset into distinct groups or clusters) - this is independent of the datasets size or number of dimensions. In terms of k-means explainability, the price of explainability (the cost incurred to attain model interpretability) is between 3 and 4 when k = 2 and falls between log k and k^2 for k > 2. For k-medians, the price of explainity is exactly 2 for k=2 and ranges from log k to k for k > 2. This indicates a feasible tradeoff between interpretability and computational efficiency.

	k-medians		k-means	
	$k=2$	$k>2$	$k=2$	$k>2$
Lower	$2-\frac{1}{d}$	$\Omega(\log k)$	$3\left(1-\frac{1}{d}\right)^2$	$\Omega(\log k)$
Upper	2	$O(k)$	4	$O(k^2)$

Compared to techniques such as ID3, the IMM algorithm compares favorably while prioritizing explainability. The findings of the paper hold significant implications for the field of machine learning. The IMM algorithm provides a fresh perspective on achieving explainability in machine learning models. It serves as an explanation that efficiency and accuracy can be balanced with the pursuit of explainability. The algorithm's practical application in complex, high dimensional data encourages further research into methods that balance computational efficiency and explainability. The IMM algorithm provides an efficient and explainable solution to clustering problems, as it guarantees a robust performance and a provable approximation to optimal clustering. Despite its strengths, the IMM algorithm has some limitations. The approximation bounds depend on the height of the tree, influencing the algorithm's performance. Additionally, datasets with complex or overlapping distributions can lead to higher numbers of mistakes, affecting the cost and effectiveness of the resulting clusters. Another limitation is the requirement of a predetermined number of clusters, limiting flexibility when dealing with data where the optimal number of clusters is unknown or can vary.