Global Scientific JOURNALS

GSJ: Volume 11, Issue 9, September 2023, Online: ISSN 2320-9186
www.globalscientificjournal.com

Unraveling the Intricacies of K-Means Clustering in Machine Learning: From Fundamentals to Real-world Applications

Sandun Jayathilake, *PG Student in Data Science*

Abstract

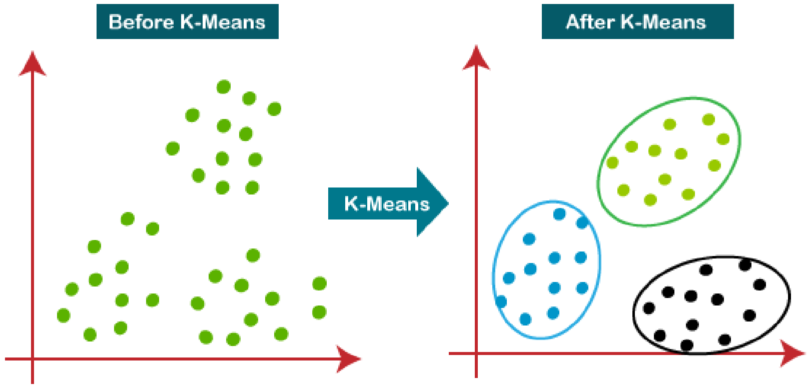
Machine learning is a field that revolves around uncovering hidden patterns and structures in data. One of the fundamental techniques used for this purpose is clustering, which groups data points based on their similarities. K-Means Clustering is one of the most widely used clustering algorithms in machine learning. This comprehensive article will delve into the intricacies of the K-Means Clustering algorithm, covering its underlying concepts, the step-by-step process, and real-world examples. By the end of this article, you will have a deep understanding of how K-Means works and how to apply it to various data analysis tasks.

Keywords

Machine Learning, Unsupervised Learning, Clustering, K-Means, Data Analysis, Centroids, Euclidean Distance, Inertia, Elbow Method, Real-world Applications.

Introduction

K-Means Clustering is a foundational unsupervised learning algorithm, invaluable in uncovering patterns within datasets used in machine learning and data mining. It falls under the umbrella of clustering algorithms, whose primary objective is to group similar data points together. "K" in *K-Means* represents the number of clusters we aim to form, and the algorithm works by iteratively refining cluster centroids to minimize the sum of squared distances(Euclidean distances) between data points and their assigned centroids. This process continues until the centroids no longer change significantly. K-Means Clustering was introduced by Stuart Lloyd in 1957 as a technique to quantize signals in the context of pulse-code modulation. However, it was popularized by John MacQueen in 1967, who provided a more formal algorithmic description. Since then, it has found applications across diverse domains, from image compression to document clustering.



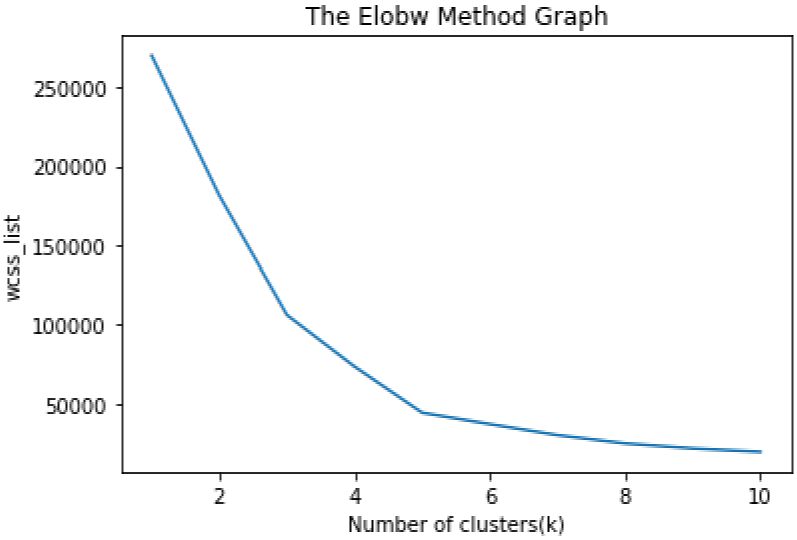
How the K-Means Algorithm works

Step 1: Select the number K to decide the number of clusters.

Selecting the appropriate value of K, which represents the number of clusters in K-Means Clustering, is a critical decision in the clustering process. Choosing an incorrect value of K can result in suboptimal clustering results. The *Elbow Method* is one of the most commonly used techniques for selecting K. It involves plotting the sum of squared distances (inertia) between data points and their assigned centroids for a range of K values. As K increases, the inertia typically decreases because the data points are closer to the centroids of their clusters. However, beyond a certain point, the rate of decrease in inertia starts to slow down, forming an "elbow" shape in the plot.

This method uses the concept of WCSS value. WCSS stands for Within Cluster Sum of Squares, which defines the total variations within a cluster. The formula to calculate the value of WCSS (for 3 clusters) is given below

$$\text{WCSS}= \sum_{\text{Pi in Cluster1}} \text{distance}(P_i\ C_1)^2 +\sum_{\text{Pi in Cluster2}}\text{distance}(P_i\ C_2)^2+\sum_{\text{Pi in CLuster3}} \text{distance}(P_i\ C_3)^2$$



$\sum_{\text{Pi in Cluster1}} \text{distance}(P_i\ C_1)^2$: It is the sum of the square of the distances between each data point and its centroid within a cluster1 and the same for the other two terms. To measure the distance between data points and the centroid, we can use any method such as Euclidean distance or Manhattan distance.

To find the optimal value of clusters, the elbow method follows the below steps:

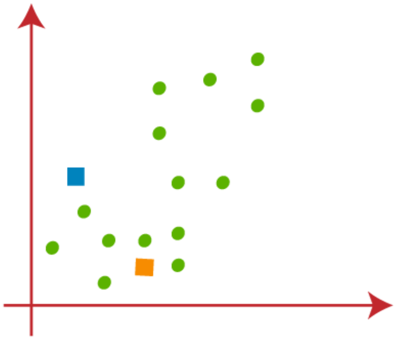
- It executes the K-means clustering on a given dataset for different K values (ranges from 1-10).
- For each value of K, calculate the WCSS value.
- Plots a curve between calculated WCSS values and the number of clusters K.
- The sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.

Since the graph shows the sharp bend, which looks like an elbow, it is known as the elbow method. To calculate the variance explained by different k values while looking for an “elbow” – a value after which higher k values do not influence the results significantly. This will be the best k value to use.

Step 2: Select random K points or centroids.

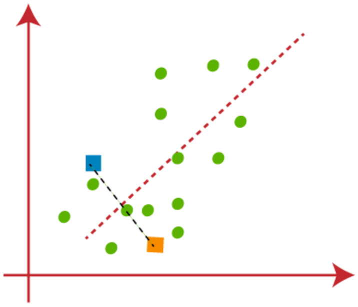
Consider the number k of clusters, i.e., K=2, to identify the dataset and to put them into different clusters. This means here we will try to group these datasets into two different clusters.

Required to choose some random k points or centroid to form the cluster. These points can be either the points from the dataset or any other point. So, here we are selecting the below two points as k points, which are not part of our dataset.



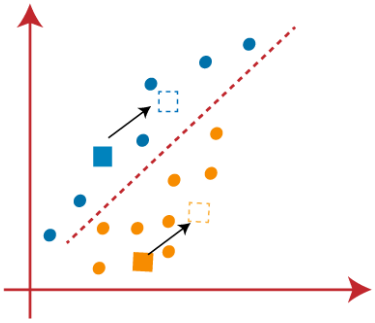
Step 3: Assign Data Points to Closest Centroids

We will compute it by applying some mathematics that we have studied to calculate the distance between two points. So, we will draw a median between both the centroids.



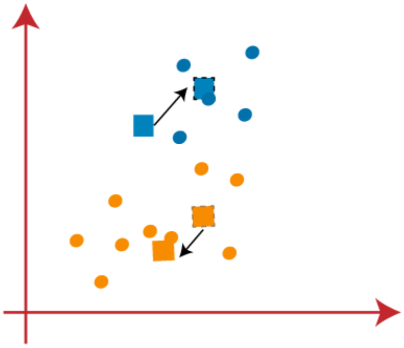
Step 4: Recalculate Centroids

Calculate the variance of the data points within that cluster. The variance of a cluster is a measure of how spread out the data points within that cluster are.



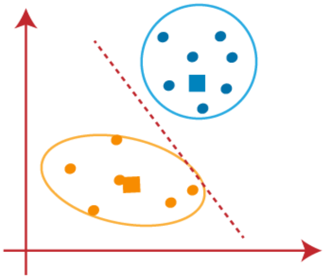
Step 5: Repeat Assignment and Recalculation

Repeat the third step, which means reassigning each data point to the new closest centroid of each cluster**.**



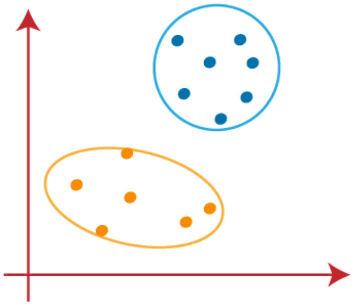
Step 6: Convergence Check

If any reassignment occurs, return to step 4; otherwise, proceed to the final step.



Step 7: **The model is ready**

K-Means model is ready or has converged is essential to stop the iterative process and finalize the clustering. Convergence means that the centroids have stabilized, and the assignments of data points to clusters are no longer changing significantly.



Applications of K-Means

Clustering may be used to tackle a wide range of real-world challenges. Below are a few examples of the applications:

- Clustering customers: Companies can use clustering to group their customers for better target marketing and understanding of their customer base. And tailoring products and services to meet specific customer preferences, ultimately leading to improved customer satisfaction and loyalty.

- Document classification: Cluster analysis of text documents in order to arrange enormous volumes of data into useful and topic-specific clusters or groupings.

- Image segmentations: These clusters are created from some of the pixels or data points comparable or common properties. Each pixel has a characteristic such as RGB value, etc. Similar pixels are clustered together using distances such as the Euclidian distance.

- Prediction of student's academic performance: Students may be divided into high achievers and ordinary performers, and this information can be used to improve the learning experience.

Advantages of K-Means

- Simple and easy to implement: K-Means is relatively easy to understand and implement, making it a popular choice for clustering tasks. It primarily relies on the Euclidean distance metric to measure similarity between data points, and it doesn't require prior knowledge or labeled data for training.
- Fast and efficient: K-Means is computationally efficient and can handle large datasets with high dimensionality, making it suitable for real-world applications
- Scalability: It can be easily scaled to tackle even larger datasets, making it suitable for real-world applications with extensive data volumes
- Flexibility: K-Means is versatile and can be adapted to various applications and data types. It works well with different distance metrics and initialization methods, allowing for customization to suit specific needs.

Disadvantages of K-Means

- Sensitivity to initial centroids: K-Means is sensitive to the initial selection cluster centroids. Different initializations can lead to different results, including suboptimal or varying cluster assignments.
- Requires specifying the number of clusters: The number of clusters k needs to be specified before running the algorithm, which can be challenging in some applications and an incorrect choice may result in poor clustering outcomes. Various techniques, such as the Elbow Method or silhouette score, can help, but they are not foolproof.
- Sensitive to outliers: Outliers in the dataset can significantly impact K-Means clustering. A single outlier can pull the centroid of a cluster far from the actual cluster's data points, leading to less meaningful clusters.
- Dependence on Distance Metrics: The choice of distance metric (e.g., Euclidean distance) can impact K-Means results. In cases where a different distance measure is more appropriate, manual adjustments are required.

Conclusion

K-Means Clustering is a strong unsupervised learning algorithm that plays a pivotal role in the realm of machine learning and data analysis. Its ability to group similar data points together has a wide range of applications, making it an indispensable tool for researchers, data scientists, and analysts. From customer segmentation to image compression, K-Means continues to prove its worth by extracting valuable insights from data, ultimately driving informed decision-making across various industries. Understanding its principles and nuances is essential for harnessing its full potential in the world of machine learning.

References

Logunova, I. (2023, January 10). K-Means Clustering in Machine Learning. https://serokell.io/blog/k-means-clustering-in-machine-learning

JavaTpoint.K-Means Clustering Algorithm. https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

Kanungo, T., Mount, D. M., Silverman, R., Netanyahu, N. S., & Wu, A. Y. (Year). The Analysis of a Simple K-Means Clustering Algorithm.

Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of K in K-means clustering. Volume 219(1).