Selection of K: Elbow Rule and Silhouette Analysis

Alberto Castro - Jose Melo

December 15, 2024

Abstract

Determining the optimal number of clusters, K, is a critical challenge in clustering analysis, particularly in the widely used K-means algorithm. In this document we explore two common approaches for K-selection: the Elbow Rule and Silhouette Analysis. The Elbow Rule provides a heuristic, graphical method based on minimizing the Within-Cluster Sum of Squares (WCSS), while Silhouette Analysis offers a data-driven approach that evaluates intra-cluster cohesion and inter-cluster separation. We detail the theoretical foundations, advantages, and limitations of each method, highlighting their applicability to different datasets and scenarios.

1 Introduction

Clustering is a very popular technique in unsupervised learning, widely used for organizing data into groups based on similarity and its patterns [3]. Among clustering algorithms, we will focus on one of most popular centroid-based clustering method known as K-means, which stands out as one of the most popular and computationally efficient methods [4].

Despite its advantages, one critical challenge of K-means algorithm is finding the optimal number of clusters, K. To address this challenge, we will discuss two techniques, the first one a heuristic approach, the elbow rule, and the second one, silhouette analysis, a more data driven approach.

The simplicity of the K-means make it suitable for a wide range of applications. For example, in market segmentation, K-means is used to group customers based on numerous features [4]. Another example could be in image compression or image segmentation, the algorithm clusters pixel values into a smaller number of representative colours in order to identify objects or compress the image [1] [2]. Moreover, K-means could be applied in other fields such as recommendation engines and document classification [4].

2 Theoretical Foundation

2.1 K-means Clustering

The K-means algorithm is based on a non-probabilistic approach. For this explanation we follow [1]. Let's consider a data set of N observations a D-dimensional variable $\mathbf{x}$,

$$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} ,$$

$$\mathbf{x} = (x_1, x_2, \ldots, x_D) .$$

The aim of the K-means algorithm, is to partition the previous data set into K clusters, in which a data point belongs just to one cluster $\mathcal{C}_k$ [4]. For that, let us also define K centroids of

each cluster, which are D-dimensional vectors that works as a prototype data point of the cluster $\mathcal{C}_k$,

$$\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K\} .$$

The main task performed by the K-means is to minimize an objective function called the *distortion measure*, which definition is

$$\Delta = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2 , \tag{1}$$

where r_{nk} is a binary indicator which assigns 1 if $\mathbf{x}_n$ belongs to the cluster $\mathcal{C}_k$ and 0 otherwise.

We just have defined the *distortion measure* as the sum of the squares of euclidean distance from each data point $\mathbf{x}_n$ to its centroid $\boldsymbol{\mu}_k$. This is also known as the Within Cluster Sum of Squares (WCSS).

2.1.1 *K*-means algorithm

In this section, we will define and algorithm to optimize the distortion (1). The algorithm consists in an iterative procedure of two steps, that ends when either a number of iterations is reached or when the change of two consecutive $\boldsymbol{\mu}_k$ is under a certain tolerance.

The algorithm could be defined by the following steps:

1. Initialize the centroids $\boldsymbol{\mu}_k$ to random values.

2. Minimize Δ keeping $\boldsymbol{\mu}_k$ fixed.

This step consists in assigning each $\mathbf{x}_n$ to the closest cluster $\mathcal{C}_k$, that means, actualize r_{nk} in order that the quantity $\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$ is minimum. This could be written in a more formally way as

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 , \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

3. Minimize Δ keeping r_{nk} fixed.

First we derive (1) with respect to each $\boldsymbol{\mu}_k$, and we get K equations, one for each k.

$$2 \sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0 , \qquad \longrightarrow \qquad \boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}} . \tag{3}$$

Let us take a look into the expression of $\boldsymbol{\mu}_k$ in (3). We fall directly that each $\boldsymbol{\mu}_k$ is the mean value of the data points in the cluster $\mathcal{C}_k$.

4. Repeat steps 2 and 3 until the convergence is terminated.

Note that in this algorithm, each step reduces Δ, so the convergence is assured, but it may converge to a local minimum rather than a global minimum. Consequently, this algorithm is dependent on the initial values of $\boldsymbol{\mu}_k$.

2.2 Selection of *K*

In the K-means algorithm, the number of clusters K is defined as an input. Since the data is unlabelled, it is natural to inquire about the optimal number of clusters for performing the K-means. In this section we will explore two different methods to approach this challenge.

2.2.1 Elbow Rule

The elbow rule is an heuristic method for finding the number of clusters, K. It takes advantage of the fact that k is an integer and the K-means usually is not computationally expensive. Therefore, we are able to compute the WCSS (1) (also known as inertia) for different values of K. This equation represents the sum of the distance between each observation $\mathbf{x}_n$, inside the cluster $\mathcal{C}_k$, and its centre ($\boldsymbol{\mu}_k$). One thing to notice is that the number of groups can be as large as the number of observations we have. Because of this, one would expect that the distance between the centre and the observations reduces as K increases. In section 2.3 we explore more in detail this simple approach.

2.2.2 Silhouette Analysis

In this second method, we explore a more data driven approach to the optimal selection of K. This is based in the analysis of the Silhouette score when performing the K-means algorithm for some values of K [6].

The Silhouette score, $\bar{s}(K)$ is the mean value of the Silhouette index of each data point $\mathbf{x}_n$ [6]. The Silhouette index was first defined in [5] as

$$s(\mathbf{x}_n) = \frac{b(\mathbf{x}_n) - a(\mathbf{x}_n)}{\max\{b(\mathbf{x}_n), a(\mathbf{x}_n)\}}\,, \tag{4}$$

where $a(\mathbf{x}_n)$ is the average distance from $\mathbf{x}_n$ to the other elements in its cluster $\mathcal{C}_k$ and $b(\mathbf{x}_n)$ is the average distance from $\mathbf{x}_n$ to all the elements to the closest neighbour cluster [7]. This idea is shown in Figure 1.

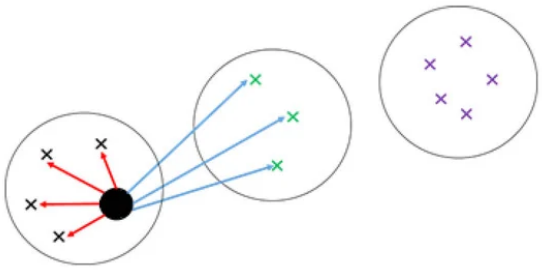

Figure 1: Procedure to calculate the Silhouette index for a data point. $a(\mathbf{x}_n)$ is the average of the red lines and $b(\mathbf{x}_n)$ is the average of the blue ones. [7]

In a more formal way, we define $a(\mathbf{x}_n)$ as

$$a(\mathbf{x}_n) = \frac{1}{N_k - 1} \sum_{\mathbf{x}_i \in \mathcal{C}_k} d(\mathbf{x}_i,\, \mathbf{x}_n)\,, \qquad i \neq n\,, \tag{5}$$

where N_k is the number of data points in the cluster $\mathcal{C}_k$ and $d(\cdot, \cdot)$ is the euclidean distance between two points. In order to construct $b(\mathbf{x}_n)$, we define the average distance of a data point $\mathbf{x}_n$ (which is in $\mathcal{C}_k$) to a neighbour cluster $\mathcal{C}_l$ as follows:

$$D_l(\mathbf{x}_n) = \frac{1}{N_l} \sum_{\mathbf{x}_i \in \mathcal{C}_l} d(\mathbf{x}_i,\, \mathbf{x}_n)\,. \tag{6}$$

Owing to the fact that $b(\mathbf{x}_n)$ is the average distance to its closest neighbour, then we can define $b(\mathbf{x}_n)$ as

$$b(\mathbf{x}_n) = \min_l \{D_l(\mathbf{x}_n)\}\,, \qquad l \neq k\,. \tag{7}$$

By just taking a look into equation (4), we found that its possible a values are in the interval $[-1, 1]$. If the Silhouette index is positive and near to 1, then we can infer that $\mathbf{x}_n$ is in the right

cluster, since $b(\mathbf{x}_n) \gg a(\mathbf{x}_n)$. On the other hand, if the index is negative and near to -1, then the data point may not be in the right cluster. Additionally, for values around 0 means that the data point is in between the two clusters, so we should not draw any conclusions. [7]

Now we recall the Silhouette score $\bar{s}(K)$, which is the arithmetic mean of the Silhouette index for all the data set and for K clusters. Therefore, we can perform the K-means for different values of K and obtain the Silhouette score. Then it would be appropriately to choose the value of K in which the Silhouette score is maximum [5]. Note that even though there is always a maximum, it may exists more natural values for K.

2.3 Advantages and Disadvantages

In this section, we analyse the strengths and limitations of the two methods discussed for selecting the number of clusters, K: the Elbow Rule and Silhouette Analysis.

As we have seen, the Elbow Rule is a simple and intuitive method for determining the optimal number of clusters in K-means clustering. Its ease of implementation makes it accessible, even to users with minimal technical expertise. By plotting the WCSS against different K values, the method allows for a straightforward visual interpretation of where the rate of decrease in WCSS slows down. This graphical approach provides insight into the diminishing importance of adding more clusters, making it especially useful for gaining a quick understanding of data structure. Moreover, the method's computational requirements are relatively low, as it only involves running K-means multiple times for various K, making it suitable for moderate-sized datasets.

However, the Elbow Rule has important limitations. The primary disadvantage is its reliance on subjective visual inspection to identify the "elbow" point, which can be ambiguous in many real applications. Additionally, the method lacks a formal mathematical criterion for determining K, which can lead to inconsistent results, for example across different users. Finally, the Elbow Rule focuses solely on WCSS reduction, without considering the quality of the resulting clusters, which may result in poorly defined clusters going unnoticed.

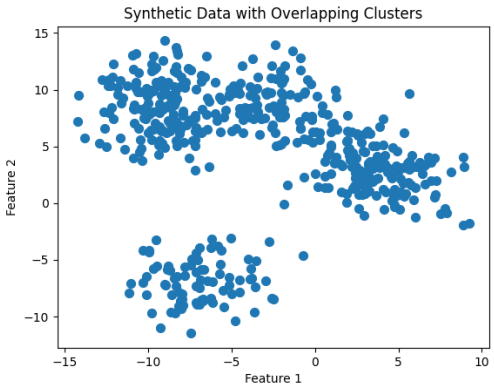

Figure 2: Clear Elbow Dataset.

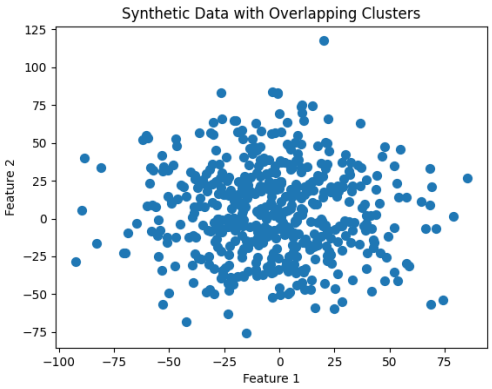

Figure 3: Unclear Elbow Dataset.

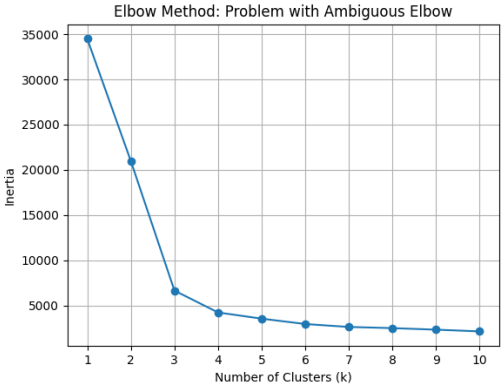

Figure 4: Clear Elbow.

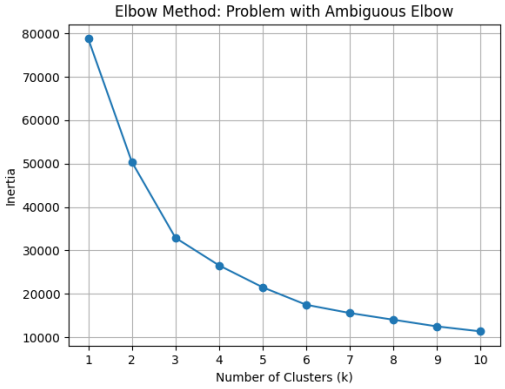

Figure 5: Unclear Elbow.

On the other hand, Silhouette Analysis provides a more data-driven and rigorous approach to selecting K by evaluating both intra-cluster cohesion and inter-cluster separation. Unlike the Elbow Rule, this method quantitatively measures how well each data point is assigned to its cluster compared to other clusters, ensuring a comprehensive assessment of clustering quality. The average silhouette score serves as an objective criterion for selecting K, thereby reducing subjectivity and enhancing reliability. Furthermore, this method adapts well to datasets with varying densities and separations, making it suitable for diverse clustering scenarios.

Despite its advantages, Silhouette Analysis also has some limitations. One significant drawback is its computational cost, as the method requires calculating pairwise distances for all data points, which can become prohibitive for very large datasets. Additionally, the approach assumes clusters are spherical or convex, meaning it may struggle to accurately evaluate datasets with irregularly shaped or overlapping clusters. Moreover, while the silhouette score provides a clear maximum, there may be cases where multiple values of K yield similar scores, leaving room for subjective judgment in the final decision.

As a demonstration, we generated data for 4 groups and apply the silhouette technique. In this case, 6 (Figure 8 is not a correct choice because the variance between groups silhouette coefficients variances in high. At the same time, selection 2 groups (Figure 7) produces groups which their silhouette plots thickness varies too much. We seek the number of plots which produces a balance silhouette plot, as shown in Figure 6, where we select 4 groups.

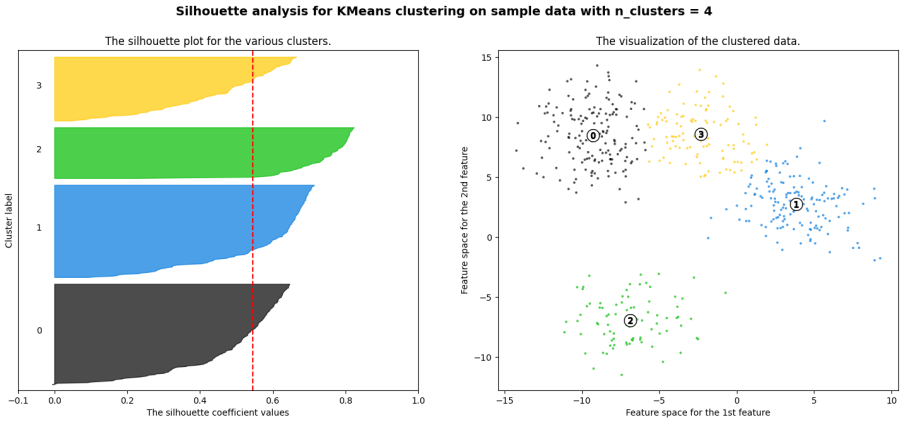

Figure 6: 4 groups.

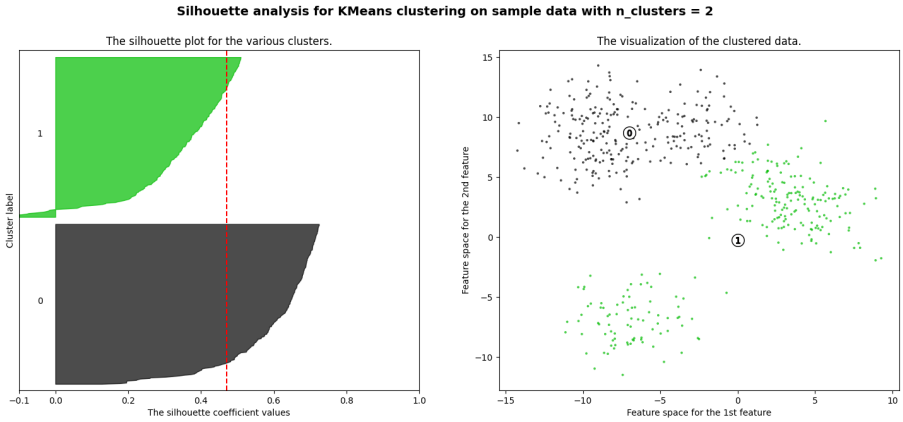

Figure 7: 2 groups.

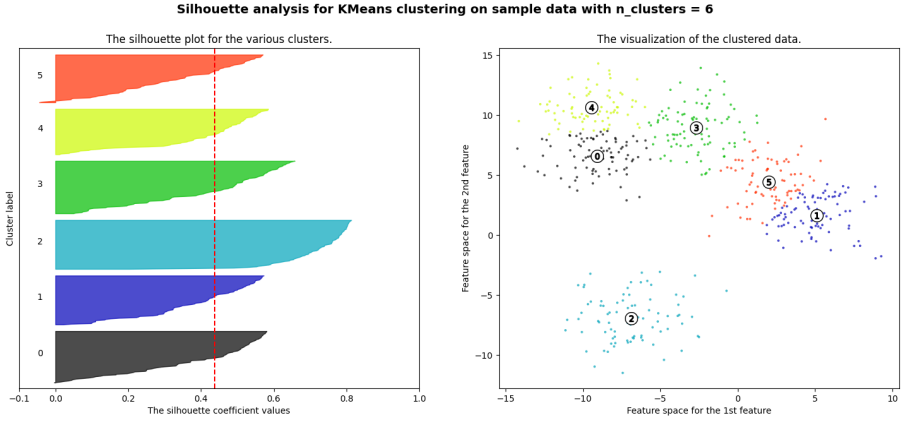

Figure 8: 6 groups.

A brief summary of the advantages and disadvantages of each method is presented in Table 1.

Aspect	**Elbow Rule**	**Silhouette Analysis**
Advantages	• Simplicity and ease of implementation. • Low computational cost for moderate datasets.	• Data-driven and rigorous evaluation. • Objective selection of K. • Adapts well to various datasets.
Disadvantages	• Subjectivity in interpreting the "elbow" point. • Lacks a quantitative criterion. • Insensitive to data characteristics and clustering quality.	• Computationally intensive for large datasets. • Sensitive to assumptions about cluster shapes. • Ambiguity in results for similar silhouette scores across K.

Table 1: Comparison of Elbow Rule and Silhouette Analysis for K-selection in K-means clustering.

3 Conclusions

The number of clusters in K-means is a crucial piece of information. It so relevant that it allows the method to perform as expected. That is why in this document we asses which tools -the Elbow Method or Silhouette Analysis- helps us get the correct answer to this question.

After what we have shown in this presentation, we can say that the Elbow Rule is best suited for scenarios requiring quick, intuitive insights into clustering, especially when computational resources are limited. However, its reliance on visual inspection can lead to inconsistent results. Silhouette Analysis, on the other hand, offers a more rigorous approach with objective metrics, making it preferable for datasets where accurate cluster evaluation is critical. Nonetheless, its higher computational cost and limitations with complex cluster shapes must be considered when scaling to large or irregular datasets.

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN: 0387310738.

[2] Nameirakpam Dhanachandra, Khumanthem Manglem, and Yambem Jina Chanu. "Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm". In: *Procedia Computer Science* 54 (2015), pp. 764–771. ISSN: 1877-0509.

[3] IBM. *What is clustering?* URL: https://www.ibm.com/topics/clustering.

[4] IBM. *What is k-means clustering?* URL: https://www.ibm.com/topics/k-means-clustering#f13.

[5] Peter J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. ISSN: 0377-0427.

[6] Ketan Rajshekhar Shahapure and Charles Nicholas. "Cluster Quality Analysis Using Silhouette Score". In: *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. 2020, pp. 747–748.

[7] Meshal Shutaywi and Nezamoddin N. Kachouie. "Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering". In: *Entropy* 23.6 (2021).