Journal of Humanistic Mathematics

Volume 12 | Issue 1 January 2022

In Search of Star Clusters: An Introduction to the K-Means Algorithm

Marcio Nascimento
Federal University of Bahia

In Search of Star Clusters: An Introduction to the K-Means Algorithm

Marcio Luis Ferreira Nascimento

Department of Chemical Engineering and Institute of Humanities, Arts and Sciences
Federal University of Bahia, Salvador, Bahia, BRAZIL
mlfn@ufba.br

Synopsis

This article is a gentle introduction to K-means, a mathematical technique of processing data for further classification. We begin with a brief historical introduction, where we find connections with Plato's *Timæus*, von Linné's binomial classification, and the star clustering concept of Mary Sommerville and collaborators. Artificial intelligence algorithms use K-means as a classification methodology to learn about data in a very accurate way, because it is a quantitative procedure based on similarities.

Keywords: K-means, clustering, machine learning, stars.

1. Background

Clustering algorithms have become more visible and more important in day-to-day applications [2] as machine learning and data science have become established as core techniques in processing large volumes of data. Clustering is a computational technique that divides variables in a dataset into groups (clusters). K-means is one such technique, a mathematical technique of processing data for further classification, proposed by the American psychologist and mathematician James Buford MacQueen (1929–2014) in 1967 [10].

There are two types of clustering methods: hierarchical and non-hierarchical. The first repeatedly links pairs of data forming the first clusters until every data object is included in order. This method resembles a tree viewed from

its roots, growing upwards, or viewed downwards. In nonhierarchical clustering, such as the K-means algorithm, the relationship between groups is undetermined. Both types of algorithms are useful to classify things.

We can view classifying as a way of thinking that has been used since ancient times, a way of grouping objects by means of their similarities — a task for the first humans to survive.

From a literary point of view, the partitioning and classification problem can be traced back to the *Book of Genesis* 9:2, when God granted to Noah dominion over "every living creature on the earth, every bird of the air, every creature that crawls on the ground, and all the fish of the sea" (Figure 1). In *Leviticus* 11:2 one finds painstaking detail classifying animals according to whether they may be eaten.

If we are interested in scientific origins, it is possible to cite the Greek philosopher Plato (*c.* 428–348 BCE), who classified living creatures in nature in his masterpiece *Timæus*, separating them into two life forms: animals and plants. This work also presented, perhaps for the first time, how the world is *intelligible* using a scientific view. Many view this as a breakpoint in human history, when logic, a mathematical tool mastered by ancient Greeks, was applied to better understand the world in animated and non-animated forms.

The next most important scientific step in the classification and naming of organisms was made by the Swedish botanist, zoologist, and physician Carl von Linné (or *Carolus Linnæus*, 1707–1778), with his binomial nomenclature. He is known as the "father of modern taxonomy", due to his groundbreaking *Systema Naturæ* [8], where he considered three classical kingdoms: *animal*, *vegetable* and *mineral*. In his work it is also possible to find the first conceptual *dendrogram*, a systematic procedure to cluster objects ("*Clavis Systematis Sexualis*" or the "Sexual Key System" of all plants, separated into 24 groups), which gained immense influence around the world.

von Linné divided living beings into groups taking into account five criteria: *kingdom*, *class*, *order*, *genus* and *species*, following a type of hierarchical clustering [8]. After 1974, new modifications and insertions were made, following the hierarchy of the major taxonomic ranks: *domain*, *kingdom*, *phylum*, *class*, *order*, *family*, *genus* and *species* [11]. The first criterion is more general, and as the criteria advance, a more specific analysis is made, obtaining a better classification.

Figure 1: The Morgan Picture Bible, also known as Morgan Crusader's Bible or Maciejowski Bible (*c.* 1240). Morgan Library & Museum, New York: www.themorgan.org. MS M.638, *folium* 2v. There are Latin, Persian, and Hebrew inscriptions. In this page, following Genesis 6 to 9 and in obedience to the Lord, Noah built a huge ark, classified animals for salvation. After the storm, Noah searched for a dry land by releasing a dove and a raven. Then, on Mount Ararat, Noah, his family and all animals descend to promised land. Grateful, Noah's family offered sacrifices to the Lord, starting a new beginning. Public domain image from `Wikipedia`.

The next step to deal with qualitative aspects of clustering, using distance as a similarity measure, was taken by the Scottish science writer and polymath Mary Somerville (*née* Fairfax, 1780–1872, Figure 2*a*). In her "Mechanism of Heavens" (1831) [13], she wrote on the relevant problem of star clusters, which had been discovered by the German amateur astronomer Johann Abraham Ihle (1627-*c.*1699) in 1665: "in some parts of the heavens, the stars are so near together as to form clusters, which to the unassisted eye appear like thin white clouds". Somerville was a close friend of the German astronomers Caroline Lucretia Herschel (1750–1848), his brother Friedrich Wilhelm Herschel (1738–1822, also a brilliant composer), as well as the English polymath John Frederick William Herschel (1792–1871, Wilhelm's son).[1] Wilhelm wrote many catalogues of nebulae and star clusters (the first in 1786 [5]) and coined the term "globular cluster" three years later [6]. This procedure seems to be close to a nonhierarchical clustering.

The first paper on clustering, still dealing with qualitative aspects, was published by the American anthropologists Harold Edson Driver (1907–1992) and Alfred Louis Kroeber (1876–1960, Figure 2*b*) in 1932 [3]. And the first book on cluster analysis was written by the American psychologist Robert Choate Tryon (1901–1967, Figure 2*c*) in 1939 [16]. They analyzed complex human aspects and tried to classify them. It is remarkable that the first paper and book on this subject were written by anthropologists and a psychologist, respectively, before any mathematical contributions.

2. A Gentle *K*-Introduction

The human eye has a keen sense of pattern, and we are good at classifying things roughly, but we are prone to various human shortcomings. The K-means technique reduces possible human subjectivity due to its precise mathematical algorithm. Its modelling is a feasible and efficient procedure when there are variations of certain characteristics among classes. It is an exploratory clustering method that partitions data into K-clusters where each group is nucleated by a *medoid* $\overline{m}_k$ (some authors prefer *centroid*) [4, 7].

[1] Editor's note: See "The Taste of Mathematics: Caroline Herschel at 31," a poem by Laura Long on the two siblings, published in the *Journal of Humanistic Mathematics* in July 2013 (Volume **3** Issue 2, page 148), available at https://scholarship.claremont.edu/jhm/vol3/iss2/14.

Figure 2: *a*) Mary Somerville (née Fairfax, 1780–1872), Scottish science writer and polymath. Portrait by the English painter John Jackson (1778–1831). *b*) Alfred Louis Kroeber (1876–1960), American anthropologist. *c*) Robert Choate Tryon (1901–1967), American psychologist. *d*) Władysław Hugo Dionizy Steinhaus (1887–1972), Polish mathematician. All pictures are in the public domain; see Acknowledgments for source information.

A medoid can be viewed as a special case of a mean, or average, of a cluster. In fact, MacQueen defined K-means as: "thus at each stage the K-means are, in fact, the means of the groups they represent (hence the term K-means)" [10].

Conceptually, K-means is actually quite simple: in each iteration, every datum is associated with the cluster that it is nearest to in terms of the cluster center. That center changes as new data are associated with the cluster until a convergence occurs.

How does one find the partition of a set? A division by two, halving, or mediation is a simple mathematical procedure. But mathematicians have other proposals to do the same, using for example the notion of distance as a similarity tool for classification.

Many sorts of data sets support K-means, but in the following we use examples specially selected so as to clarify and simplify the algorithm.

As our first example, let us partition the numbers between 1 and 6 into two clusters A and B ($K = 2$). Begin with our six elements:

$X_{1,6}$:	1	2	3	4	5	6

For this example, a monodimensional data set, $X_{i,p}$, is presented, where i is one (just one observation or case) and p, the number of labels (or variables, properties, or characteristics), is six, distributed by $K = 2$ groups.

The basic procedure would be to pick two numbers from the data (called seeds or medoids) and to calculate the numerical distances between these seeds from original data. Taking for example **1** and **6** as medoids, the distances of the elements to the medoids are:

x	distance to **1**	distance to **6**
1	$\lvert 1 - \mathbf{1}\rvert = \mathbf{0}$	$\lvert 1 - \mathbf{6}\rvert = \mathbf{5}$
2	$\lvert 2 - \mathbf{1}\rvert = \mathbf{1}$	$\lvert 2 - \mathbf{6}\rvert = \mathbf{4}$
3	$\lvert 3 - \mathbf{1}\rvert = \mathbf{2}$	$\lvert 3 - \mathbf{6}\rvert = \mathbf{3}$
4	$\lvert 4 - \mathbf{1}\rvert = \mathbf{3}$	$\lvert 4 - \mathbf{6}\rvert = \mathbf{2}$
5	$\lvert 5 - \mathbf{1}\rvert = \mathbf{4}$	$\lvert 5 - \mathbf{6}\rvert = \mathbf{1}$
6	$\lvert 6 - \mathbf{1}\rvert = \mathbf{5}$	$\lvert 6 - \mathbf{6}\rvert = \mathbf{0}$

Elements 1, 2, and 3 are closest to **1**, while 4, 5, and 6 are closest to **6**.

Therefore the first clusters A and B are

$$A = \{1, 2, 3\}$$
$$B = \{4, 5, 6\}$$

The next medoids would be a simple average between 1 to 3 (that is, $\overline{m}_1 = \mathbf{2}$) and between 4 to 6 (that is, $\overline{m}_2 = \mathbf{5}$). Calculating new distances from the original data considering these new medoids would give the same clusters A and B, giving the final result.

$A = \{1, 2, 3\}$	**2 as medoid**
$B = \{4, 5, 6\}$	**5 as medoid**

The simple algorithm we used was inspired by original work proposed by MacQueen [10], and can be applied to larger data sets [4, 7].

Others proposed similar algorithms independently, such as the Polish mathematician Władysław Hugo Dionizy Steinhaus (1887–1972, Figure 2*d*) in [14], and the American physicist Stuart Phinney Lloyd (1923–2007) in [9], both in 1957. However, due to confidentiality provisions in his contracts, Lloyd could only publish his results in 1982.

Let us now suppose a two-dimensional data described by pairs like stars in the sky. The five elements are

$$\{(\mathbf{1}, \mathbf{3}), (\mathbf{2}, \mathbf{3}), (\mathbf{1}, \mathbf{5}), (\mathbf{5}, \mathbf{3}), (\mathbf{6}, \mathbf{2})\}.$$

Taking (**1**,**3**) and (**6**,**2**) as the first seeds (medoids), the distances to the elements can be calculated as follows:

For $\overline{m}_1 = (\mathbf{1}, \mathbf{3})$ as initial medoid:

$$\text{distance of } (\mathbf{1}, \mathbf{3}) \text{ from } \overline{m}_1 \text{ is } \sqrt{(1-\mathbf{1})^2 + (3-\mathbf{3})^2} = 0.00,$$
$$\text{distance of } (\mathbf{2}, \mathbf{3}) \text{ from } \overline{m}_1 \text{ is } \sqrt{(2-\mathbf{1})^2 + (3-\mathbf{3})^2} = 1.00,$$
$$\text{distance of } (\mathbf{1}, \mathbf{5}) \text{ from } \overline{m}_1 \text{ is } \sqrt{(1-\mathbf{1})^2 + (5-\mathbf{3})^2} = 2.00,$$
$$\text{distance of } (\mathbf{5}, \mathbf{3}) \text{ from } \overline{m}_1 \text{ is } \sqrt{(5-\mathbf{1})^2 + (3-\mathbf{3})^2} = 4.00,$$
$$\text{distance of } (\mathbf{6}, \mathbf{2}) \text{ from } \overline{m}_1 \text{ is } \sqrt{(6-\mathbf{1})^2 + (2-\mathbf{3})^2} = 5.10.$$

For $\overline{m}_2 = (\mathbf{6}, \mathbf{2})$ as initial medoid:

$$\text{distance of } (\mathbf{1}, \mathbf{3}) \text{ from } \overline{m}_2 \text{ is } \sqrt{(1-\mathbf{6})^2 + (3-\mathbf{2})^2} = 5.10,$$
$$\text{distance of } (\mathbf{2}, \mathbf{3}) \text{ from } \overline{m}_2 \text{ is } \sqrt{(2-\mathbf{6})^2 + (3-\mathbf{2})^2} = 4.12,$$
$$\text{distance of } (\mathbf{1}, \mathbf{5}) \text{ from } \overline{m}_2 \text{ is } \sqrt{(1-\mathbf{6})^2 + (5-\mathbf{2})^2} = 5.83,$$
$$\text{distance of } (\mathbf{5}, \mathbf{3}) \text{ from } \overline{m}_2 \text{ is } \sqrt{(5-\mathbf{6})^2 + (3-\mathbf{2})^2} = 1.41,$$
$$\text{distance of } (\mathbf{6}, \mathbf{2}) \text{ from } \overline{m}_2 \text{ is } \sqrt{(6-\mathbf{6})^2 + (2-\mathbf{2})^2} = 0.00.$$

The minimum value between the respective distances defines the first clusters, named A and B.

$$A = \{(1,3), (2,3), (1,5)\}$$
$$B = \{(5,3), (6,2)\}$$

The next medoids would be a simple average over clusters A (that is, $\overline{m}_1 = (\mathbf{1.33}, \mathbf{3.67})$) and B (that is, $\overline{m}_2 = (\mathbf{5.50}, \mathbf{2.50})$). At this point, the distances can be recomputed.

For $\overline{m}_1 = (\mathbf{1.33}, \mathbf{3.67})$ as initial medoid, we get the following distances:

$$\sqrt{(1-\mathbf{1.33})^2 + (3-\mathbf{3.67})^2} = 0.75$$
$$\sqrt{(2-\mathbf{1.33})^2 + (3-\mathbf{3.67})^2} = 0.94$$
$$\sqrt{(1-\mathbf{1.33})^2 + (5-\mathbf{3.67})^2} = 1.37$$
$$\sqrt{(5-\mathbf{1.33})^2 + (3-\mathbf{3.67})^2} = 3.73$$
$$\sqrt{(6-\mathbf{1.33})^2 + (2-\mathbf{3.67})^2} = 4.96$$

For $\overline{m}_2 = (\mathbf{5.50}, \mathbf{2.50})$ as initial medoid, we get the following distances:

$$\sqrt{(1-\mathbf{5.50})^2 + (3-\mathbf{2.50})^2} = 4.53$$
$$\sqrt{(2-\mathbf{5.50})^2 + (3-\mathbf{2.50})^2} = 3.54$$
$$\sqrt{(1-\mathbf{5.50})^2 + (5-\mathbf{2.50})^2} = 5.15$$
$$\sqrt{(5-\mathbf{5.50})^2 + (3-\mathbf{2.50})^2} = 0.71$$

$$\sqrt{(6 - \mathbf{5.50})^2 + (2 - \mathbf{2.50})^2} = 0.71$$

These new distances give the same clusters A and B, illustrated in Figure 3.

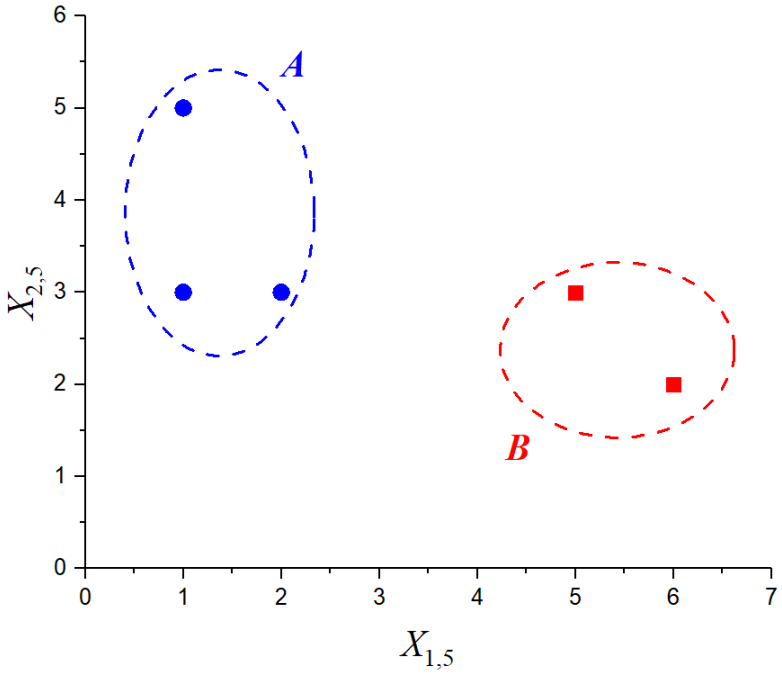

Figure 3: Two-dimensional K-means cluster distribution example.

There are also ways to choose an optimal K from a data set. As far as I know, the first way to do this was proposed by the American psychologist Robert Ladd Thorndike (1910–1990) in 1953 [15]. This is the *a posteriori* method, named today as *elbow* or critical point procedure, that was labeled as such by Thorndike based on an old science fiction TV show named *Captain Video and His Video Rangers*. This show aired in the United States between 1949 and 1955 [15]. Thorndike discussed clustering in terms of family members, looking for ideal grouping numbers. More precisely, he had in mind the ratings of each of twelve Air Force job categories with respect to nineteen dimensions (or categories). He observed that the average within-cluster distance changed for different numbers of clusters but did not explain the reason, waiting for next episodes of *Captain Video* and its characters [15]. Now there is a vast literature about procedures to decide on the optimal K; see for example [4, 7].

However, Thorndike's conclusion is not surprising in exploratory analysis. For instance, in 2015, Raphael Silberzahn (*b.* 1984) and Eric Luis Uhlmann (*b.* 1978) recruited twenty-nine research teams and asked them to answer the same research question with the same data set [12]. The question was the following: "are football (soccer) referees more likely to give red cards to players with dark skin than to players with light skin?" Using different exploratory techniques, of the twenty-nine teams, twenty found a statistically significant correlation between skin colour and red cards. Thus, results and conclusions were not the same. This is a simple way to note that no convergence occurred. But, that's OK: classification is a huge task since Plato and Linné, still in progress [1].

3. Final Words

Historically, clustering was used to solve complex problems by philosophers as Plato, physicians as Linné, astronomers as Sommerville, anthropologists as Driver and Kroeber, psychologists as Tryon, Thorndike, and MacQueen, and then many mathematicians and physicists, as Steinhaus and Lloyd, for example.

Today, artificial intelligence and machine learning algorithms use classification methodologies such as K-means to learn about data in a very accurate way, because these offer a quantitative procedure based on similarities. As explained by Pedro Morais Delgado Domingos (*b.* 1965) in his interesting book *The Master Algorithm* [2], computer machines learn quite similarly to humans, considering lots of lessons (or data) from last decades (and even centuries) by trial and error but faster.

As far as I know, Sommerville was the first to note the mathematical problem related to stars that are close enough to one another to form clusters. She anticipated by a century and half the promise of distance as a similarity tool.

Inspired by local astronomical observations, some national flags display certain constellations. For example, the Southern Cross or Crux, visible in the Southern Hemisphere, is represented in Australia, New Zealand, Papua New Guinea, and Samoa flags. However, there is one, the Brazilian flag, that represents more that just one constellation. One can observe five star clusters, including the Crux, Canis Major, Hydra, Triangulum Australe, and Scorpius (Figure 4), that can be grouped by machines using K-means clustering in the same way the first humans did so long ago.

Figure 4: The Gold-Green Flag of Brazil. Public domain image from Wikipedia. All stars represent the sky at Rio de Janeiro, the second Brazilian capital, at 8:30 a.m. on 15 November 1889, the Proclamation of the Republic Day. There are five-star clusters (Canis Major, Hydra, Crux, Triangulum Australe and Scorpius), where each star represents all 26 Brazilian states plus the Federal District. This is also the unique flag with an inscription, or lemma: Order and Progress.

As an interpretative tool, one would also find different clusters from the Brazilian flag using the K-means routine. This is still fine because the astronomical groups were first determined arbitrarily, including cultural, social and historical contexts. This is the power and the curse of some exploratory tools—they can sometimes give us a new answer, or at least a new way to think about the problem in an interesting way.

Acknowledgments. Many thanks to the Brazilian physicist A. P. Ricieri, who introduced the author to this subject by means of his course at the Prandiano Museum (http://www.prandiano.com.br). This work was supported by the National Council for Scientific and Technological Development (CNPq), contracts 304705/2015-2, 404004/2016-4 and 305331/2018-3.

Readers can find the originals of the images used in Figure 2a at https://artuk.org/discover/artworks/mary-somerville-17801872-as-a-young--woman-223448; Figure 2b at https://alchetron.com/Alfred-L-Kroeber; Figure 2c at https://www.gf.org/fellows/all-fellows/robert-choate-tryon/; and Figure 2d at https://alchetron.com/Hugo-Steinhaus.

References

[1] E. S. Chang, M. Neuhof, N. D. Rubinstein, A. Diamant, H. Philippe, D. Huchon, P. Cartwright, "Genomic insights into the evolutionary origin of Myxozoa within Cnidaria," *Proceedings of the National Academy of Sciences*, Volume **112** (2015), pages 14912–14917.

[2] P. Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, Basic Books, New York 2015.

[3] H. E. Driver, A. L. Kroeber, "Quantitative Expression of Cultural Relationships," *University of California publications in American Archaeology and Ethnology*, Volume **31** (1932), pages 211–256.

[4] J. F. Hair, Jr, W. C. Black, B. J. Babin, R. E. Anderson, *Multivariate Data Analysis*. Cengage, Hampshire, 2019.

[5] W. Herschel, "Catalogue of One Thousand New Nebulae and Clusters of Stars," *The Philosophical Transactions of the Royal Society of London*, Volume **76** (1786), pages 457–499.

[6] W. Herschel, "Catalogue of a Second Thousand of New Nebulae and Clusters of Stars; with a Few Introductory Remarks on the Construction of the Heavens," *The Philosophical Transactions of the Royal Society of London*, Volume **79** (1789), pages 212–255.

[7] D. Kopec, *Classic Computer Science Problems in Python*, Manning Publications Co., Shelter Island, 2019.

[8] C. Linnæi, *Systema naturæ per regna tria naturæ, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis* [System of nature through the three kingdoms of nature, according

to classes, orders, genera and species, with characters, differences, synonyms, places], J. W. de Groot, Lugduni Batavorum 11 p. (*in Latin*), 1735.

[9] S. P. Lloyd, "Least Squares Quantization in PCM," *IEEE Transactions in Information Theory*, Volume **28** (1982), pages 129–137.

[10] J. B. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," pages 281–297 in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

[11] R. T. Moore, "Proposal for the Recognition of Super Ranks," *Taxon*, Volume **23** Number 4 (1974), pages 650–652.

[12] R. Silberzahn, E. L. Uhlmann, "Many Hands Make Tight Work," *Nature*, Volume **526** (2015), pages 189–191.

[13] M. Somerville, *Mechanism of the Heavens*. John Murray Ed., London, 1831.

[14] H. Steinhaus, "Sur la Division des Corps Matæriels en Parties [On the Division of Material Bodies into Parts]," *Bulletin L'Académie Polonaise des Science*, Volume **4** (1957), pages 801–804 (in *French*).

[15] R. L. Thorndike, "Who Belongs in the Family?" *Psychometrika*, Volume **18** (1953), pages 267–276.

[16] R. C. Tryon, *Cluster Analysis: Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality*, Edwards Brother Inc. Lithoprinters and Publishers, Ann Arbor, Michigan, 1939.