

Bachelor Thesis Topic: Deterministic Initialization for X-Means

Winter Semester 2026/27

1 Introduction and Background

K-Means [9] can be considered to be the most successful clustering algorithm. It relies on a straightforward approach that iteratively assigns data instances to cluster centroids and updates those centroids to fit the assigned data. Despite its success, k-Means has several flaws: the first is that it heavily depends on its initialization [6], in which the initial centroid candidates used for assignment are set. The second is that it requires the concrete number of clusters k to be specified, which may not be known beforehand.

Still, given its importance, there are already many approaches to address these challenges. There exist a wide variety of different initialization strategies. The most frequently used ones are a uniformly random selection from the pool of instances [10, 3] and k-Means++ [1], which relies on a random selection weighted by the distance to other centroids. However, both methods are non-deterministic because they rely on a random selection of centroids. This typically means that their execution is repeated several times across different random states. However, as clustering evaluation is non-trivial [15], selecting among multiple clusterings may be difficult if only one is desired. As such, there have also been multiple deterministic approaches to k-Means initialization [3, 16]. These methods usually rely on a variety of tradeoffs to produce initialization sets for k-Means. Some major ones include PCA-Part [13] and Var-Part [14], which iteratively split the data according to the eigenvector with the largest eigenvalue and the dimension with the highest variance, respectively. Later variants also include more sophisticated splitting positions at the cost of an additional hyperparameter [2]. Katsavounidis' variant [8] of the Maximin approach [4] selects the next centroid so that it maximizes the distance to all prior centroids, starting with the instance with the highest norm. More recently, Density K-means++ [11] relies on the data density to select the initial centroids.

Similarly, there are also multiple k-estimation approaches [5, 12, 7] that try to determine the most suitable number of clusters for the input data. Here, we specifically want to focus on X-Means [12], a popular k-estimation strategy. X-Means works by iteratively running k-Means and testing whether splitting the

found centroids would be advantageous. The algorithm starts with an initial k-Means run that selects the first centroid candidates.

While both issues in k-Means are well-researched and relevant, there has been little research on combining their solutions. The goal of this thesis is to perform this combination, utilizing a selection of deterministic initialization strategies together with X-Means. For X-Means, there are two places where deterministic initialization can be inserted: the most straightforward is to apply the initialization to the k-Means call in the first step; however, the iterative k-Means calls that split the centroid candidates can also be made deterministic. Importantly, the best-suited deterministic initialization method may differ between the two phases, necessitating a differentiating investigation.

2 Research Question

How suitable are different deterministic initialization methods for the X-Means setting?

3 Tasks & Goals

- **Literature Review:** Understanding the primary literature regarding k-Means, X-Means, as well as different initialization strategies for k-Means
- **Candidate Selection:** Well-structured preliminary selection of viable deterministic initialization methods
- **Implementation of Candidates:** Implementation of different initialization strategies
- **Integration of deterministic initialization strategies into initialization of X-Means:** Combination of the implemented initialization strategies with pre-existing X-Means code for both the initial k-Means run and subsequent centroid splits
- **Experimental Evaluation:** extensive experiments across different datasets and combinations (including a comparison against a greedy max-distance split in the centroid-splitting steps and against k-Means using the same initialization method with the estimated final k -value)
- **Analysis & Discussion:** Investigation of the advantages and disadvantages of different initialization approaches in this X-Means-based setting.

4 Expected Outcomes

- An in-depth analysis of the suitability of different deterministic k-Means initialization strategies for X-Means

- A proper comparison of the X-Means results with corresponding k-Means results
- A well-documented thesis with reproducible code.

5 Resources

Python implementations for X-Means (<https://github.com/collinleiber/ClustPy/tree/main>) and k-Means (<https://scikit-learn.org>) are available

6 Requirements

- Study in the field of computer science
- Prior programming experience in Python
- Beneficial: understanding of clustering

7 Contact

If you are interested, please send your CV and transcripts to jahn@dbis.fh-lmu.de

References

- [1] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In Nikhil Bansal, Kirk Pruhs, and Clifford Stein, editors, *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007*, pages 1027–1035. SIAM, 2007.
- [2] M. Emre Celebi and Hassan A. Kingravi. Deterministic initialization of the k-means algorithm using hierarchical clustering. *Int. J. Pattern Recognit. Artif. Intell.*, 26(7), 2012.
- [3] M. Emre Celebi, Hassan A. Kingravi, and Patricio A. Vela. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Syst. Appl.*, 40(1):200–210, 2013.
- [4] Teofilo F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theor. Comput. Sci.*, 38:293–306, 1985.
- [5] Sibylle Hess and Wouter Duivesteijn. k is the magic number - inferring the number of clusters through nonparametric concentration inequalities. In Ulf Brefeld, Élisabeth Fromont, Andreas Hotho, Arno J. Knobbe, Marloes H. Maathuis, and Céline Robardet, editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2019*,

Würzburg, Germany, September 16-20, 2019, *Proceedings, Part I*, volume 11906 of *Lecture Notes in Computer Science*, pages 257–273. Springer, 2019.

- [6] Abiodun M. Ikotun, Absalom E. Ezugwu, Laith Abualigah, Belal Abuhaija, and Heming Jia. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Inf. Sci.*, 622:178–210, 2023.
- [7] Argyris Kalogeratos and Aristidis Likas. Dip-means: an incremental clustering method for estimating the number of clusters. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 2402–2410, 2012.
- [8] Ioannis Katsavounidis, C.-C. Jay Kuo, and Zhen Zhang. A new initialization technique for generalized lloyd iteration. *IEEE Signal Process. Lett.*, 1(10):144–146, 1994.
- [9] Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–136, 1982.
- [10] J. MacQueen. Some methods for classification and analysis of multivariate observations. 1967.
- [11] N. Nidheesh, K. A. Abdul Nazeer, and P. M. Ameer. An enhanced deterministic k-means clustering algorithm for cancer subtype prediction from gene expression data. *Comput. Biol. Medicine*, 91:213–221, 2017.
- [12] Dan Pelleg and Andrew W. Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In Pat Langley, editor, *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, pages 727–734. Morgan Kaufmann, 2000.
- [13] Ting Su and Jennifer G. Dy. A deterministic method for initializing k-means clustering. In *16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2004), 15-17 November 2004, Boca Raton, FL, USA*, pages 784–786. IEEE Computer Society, 2004.
- [14] Ting Su and Jennifer G. Dy. In search of deterministic methods for initializing k-means and gaussian mixture clustering. *Intell. Data Anal.*, 11(4):319–338, 2007.
- [15] Ulrike von Luxburg, Robert C. Williamson, and Isabelle Guyon. Clustering: Science or art? In Isabelle Guyon, Gideon Dror, Vincent Lemaire, Graham W. Taylor, and Daniel L. Silver, editors, *Unsupervised and Transfer Learning - Workshop held at ICML 2011, Bellevue, Washington, USA*,

July 2, 2011, volume 27 of *JMLR Proceedings*, pages 65–80. JMLR.org, 2012.

- [16] Avgoustinos Vouros, Stephen Langdell, Mike Croucher, and Eleni Vasilaki. An empirical comparison between stochastic and deterministic centroid initialisation for k-means variations. *Mach. Learn.*, 110(8):1975–2003, 2021.