Nyström-Optimized Kernel K-Means Clustering

Luke Flecker and Tanush Kallem

01/21/2025

Dr. Yilmaz

Quarter 2 Project

Abstract

Clustering is one of the major problems in machine learning, and linear K-means often underperforms for data sets containing non-linear and intricate relationships. Here, we explore the task of improving clustering with radial basis function (RBF) kernels and K-means. To avoid computational cost for exact kernel computation, we apply the Nyström approximation, making efficient kernel approximation feasible using only a portion of the data. The method proposed in this work is new due to its application of the Nyström approximation to kernel K-means directly with the intention of reducing runtime. Experimental results for four varied datasets—moon, spiral, circles, and blobs—indicate that although Nyström Kernel K-means takes slightly more time to run than linear K-means, unexpectedly, it delivers higher accuracy, especially for datasets having non-linear clusters. In addition, Nyström Kernel K-means is as correct as entire kernel K-means and is computationally comparable, rendering it an excellent candidate for large-scale clustering in real-world configurations.

Introduction

Clustering is a fundamental task in unsupervised machine learning, and it seeks to divide data points into clusters based on similarity. Of clustering algorithms, K-means is favored due to its simplicity and efficiency. However, standard K-means performs poorly with datasets that have nonlinearly separable clusters, such as spirals or concentric circles. In such cases, kernel techniques, which project data implicitly to a higher-dimensional feature space, offer a good solution in the form of enabling linear algorithms to work successfully in this new space. Kernel K-means, in particular, uses radial basis function (RBF) kernels to detect nonlinear patterns in data. Yet, despite its advantages, kernel K-means is computationally intensive when applied to large datasets since it involves computation and storage of the entire kernel matrix.

This paper discusses a better method for kernel K-means based on the Nyström approximation, which is used to alleviate the computational cost of kernel matrix calculation. The Nyström approximation can effectively approximate the kernel matrix by choosing a subset of data points as the basis samples, thereby greatly accelerating the algorithm with a fairly good accuracy. The combination of kernel K-means and the Nyström approximation is an efficient solution for the clustering of complex data, especially where ordinary K-means and even ordinary kernel K-means will be lacking.

The primary aim of this study is to compare the performance of three cluster methods: linear K-means, full kernel K-means, and Nyström kernel K-means. The emphasis is to compare the computational cost and clustering performance of Nyström kernel K-means on various datasets with different complexities, e.g., moons, spirals, circles, and Gaussian blobs. We think that Nyström kernel K-means will run faster than full kernel K-means on nonlinearly separable data. We also will examine whether Nyström approximation has comparable performance with exact kernel K-means at lower computational costs and, therefore, be a useful alternative solution to large-scale or high-dimensional clustering problems.

This research enhances clustering methods in complicated data structures by solving the restrictions of kernel K-means and making use of the Nyström approximation. This research

could impact numerous applications such as image segmentation, bioinformatics, and segmentation of the market where data clustering in a nonlinear form is critical.

Related Work

The general approach is to apply the Nyström method to estimate the kernel matrix in an effort to reduce the computational requirements of kernel k-means. This enables scalability to big data, as well as enabling the use of stable kernel methods for clustering non-linear data. The Nyström approximation, however, can be inexact, and the choice of landmarks (basis samples) contributes to affecting clustering accuracy. This paper explains these trade-offs and experimentally demonstrates the performance of Nyström kernel k-means, in line with the theory introduced by Wang, Gittens, & Mahomey, 2019.

This paper is a follow-up to previous work. Specifically, Oglic, D., Gärtner, T, (2017) attempted landmark selection methods for the Nyström technique experimentally and derived error bounds theoretically. While their paper was slightly more theoretical, this paper is focused on measuring performance on real-world datasets. Furthermore, Wang et al. (2019) provided approximation guarantees for rank-constrained Nyström techniques, which form the basis of the method here. In contrast to Bayesian nonparametric clustering algorithms, e.g., by Kulis & Jordan (2012), this work follows deterministic kernel-based methods and does not seek adaptive clustering.

Comparative experiments with comparable approaches, e.g., spectral clustering using the Nyström approximation or random feature maps, illustrate the enhanced scalability and quality of clustering in Nyström kernel k-means. Moreover, the experiments confirm that optimal landmark selection, as approximated by kernel k-means++, is essential for both reducing computational cost and accuracy. This work adds to the art by empirically confirming these observations and providing insights into the real-world applications of kernel approximations for clustering.

Datasets and Features

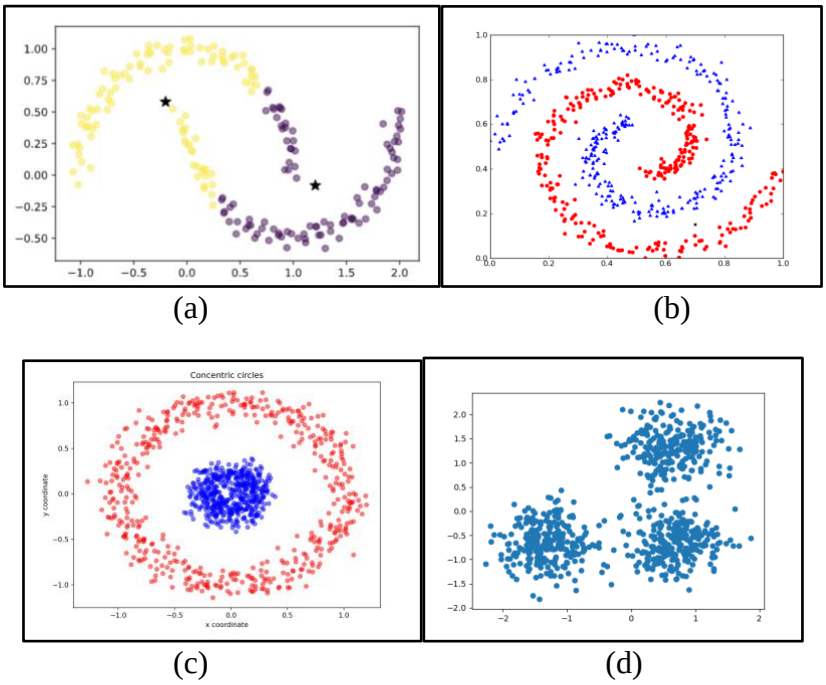
The dataset exploration consists of four different synthetic datasets, each chosen for their characteristics and relevance to machine learning tasks. The moon dataset is a two-dimensional dataset created by the function make_moons from scikit-learn (Figure 1 (a)), ideal for binary classification problems. It consists of two interleaving crescent-shaped clusters that are non-linearly separable. This includes a noise parameter that allows for variability in the data to simulate real-world imperfections and provide a challenge for classifiers.

The spiral dataset is one of the representative examples of visually striking nonlinear relationships (Figure 1 (b)), consisting of two or more classes in a spiral arrangement that would pose a difficult learning task for any algorithm relying on linear decision boundaries. Because of this, the dataset is usually used to test the learning capabilities of kernel-based methods and neural networks from such complex patterns.

The circle's dataset consists of two concentric circle patterns (Figure 1 (c)), with the points lying in an inner and outer circle. This dataset represents problems where nonlinear separations are

involved-that is, points from different classes interlink with each other. The factor parameter determines the relative size of the circles, while the noise parameter introduces variability, allowing the user to simulate different difficulty levels.

The last one is the blobs dataset (Figure 1 (d)), which consists of Gaussian-distributed points forming distinct clusters. Optionally, through parameters such as the number of centers and cluster standard deviation, this may be used to generate more or less difficult datasets. It can thus serve as a simple yet powerful benchmarking tool for clustering algorithms and classification techniques. In all, these datasets allow the performance analysis of different machine learning models when dealing with diverse conditions. For all the datasets, the noise parameter was 0.05, and the random_state was 0.



(a) (b)

(c) (d)

Figure 1: Datasets with two classes: (a) Moons, (b) Concentric circles, (c) Spirals, (d) Blobs

Methods:

1. Linear K-means (Baseline)

The standard K-means algorithm is a centroid-based clustering method that partitions data into k clusters by minimizing the within-cluster variance (Jin & Han, 2011). Given a dataset X with n samples and d features, the algorithm iteratively assigns each data point to the nearest centroid and updates the centroids based on the mean of the assigned points. The objective function is given by:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i}^{k} \| x - \mu i \|^2$$

where C_i represents the ith cluster, μi is the centroid of C_i, and $\|x-\mu i\|^2$ is the squared Euclidean distance between a data point x and the centroid μi.

In our implementation, we use the KMeans class from the sklearn.cluster module with k = 2 clusters, as all our datasets are binary classification problems. The algorithm is applied to the standardized version of the dataset, where each feature is scaled to have zero mean and unit variance using StandardScaler.

2. Full Kernel K-means

To handle non-linear structures in the data, we extend the K-means algorithm by incorporating the RBF kernel (Pycodemates, 2022). The RBF kernel measures the similarity between two data points x_i and x_j as:

$$K(xi, xj) = exp(-\gamma \parallel xi - xj \parallel^2)$$

where γ is a hyperparameter that controls the influence of each data point. The kernel matrix K is computed for the entire dataset, and the K-means algorithm is applied in the transformed feature space.
However, computing the full kernel matrix is computationally expensive, especially for large datasets, as it requires $O(n^2)$ operations. To address this, we center the kernel matrix using the following transformation:

$$K_{centered} = K - \frac{1}{n}1_n K - \frac{1}{n}K1_n + \frac{1}{n^2}K1_n$$

where 1_n is an n × n matrix of ones. The centered kernel matrix is then used as input to the K-means algorithm.

In our experiments, we perform a grid search over γ values ranging from 0 to 7 in increments of 0.1 to find the optimal γ that maximizes clustering accuracy.

3. Nyström-Approximated Kernel K-means

To reduce the computational cost of full kernel K-means, we propose the use of the Nyström approximation (Jones, 2025). The Nyström method approximates the kernel matrix by sampling a subset of m data points (where $m < n$) and computing the kernel matrix only for the sampled points. The full kernel matrix is then approximated as:

$$K_{approx} = K_{nm}K_{mm}^{-1}K_M^T$$

where K_{nm} is the kernel matrix between the full dataset and the sampled points, and K_{mm} is the kernel matrix of the sampled points. The inverse K_{mm}^{-1} is computed using the pseudo-inverse to handle potential singularities.

In our implementation, we sample m = 50 points uniformly at random and compute the Nyström approximation for each γ value in the range [0, 7] with increments of 0.1. Nyström approximation has never practically been implemented with kernel k means, making this our unique implementation.

Experimental Setup/Results/Discussion

In this section, we compare the running time of Linear K-means, Nyström Kernel K-means, and Full Kernel K-means on four datasets: Moon (Figure 1(a)), Spiral (Figure 1(b)), Circles (Figure 1(c)), and Blobs (Figure 1 (d)). These experiments are done to compare the running time and accuracy of the algorithms and show the trade-offs and advantages of applying the Nyström approximation to kernel-based clustering.

Experimental Setup

The clustering algorithms were tested on the following metrics:

Accuracy: The ratio of points accurately clustered to ground-truth labels.

Runtime: The duration required to cluster the dataset.

Nyström approximation was used to speed up kernel matrix computations in the context of kernel K-means. The experiment process involved changing the RBF kernel parameter, γ, in increments of 0.1 for the Nyström approach and Full Kernel K-means to determine the most appropriate parameter. The basis sample number of the Nyström approximation was kept at 50. Linear K-means was used as a baseline method for comparative purposes.

Results:

Our results can be seen in Table 1.

Dataset	Method	Accuracy	Runtime(s)	Optimal γ
Moon	Linear K-means	0.85	0.13	-
	Nyström Kernel K-means	0.99	4.68	4.3
	Full Kernel K-means	0.92	3.61	6.6
Spiral	Linear K-means	0.61	0.01	-
	Nyström Kernel K-means	0.69	9.10	4.2
	Full Kernel K-means	0.68	5.47	1.7
Circles	Linear K-means	0.52	0.01	-
	Nyström Kernel K-means	0.73	4.74	2.2
	Full Kernel K-means	1.00	2.31	2.2
Blobs	Linear K-means	1.00	0.00	-
	Nyström Kernel K-means	1.00	1.52	0.1
	Full Kernel K-means	1.00	1.41	0.1

Table 1: Results of different algorithms on each dataset

Linear K-means Performance:

Linear K-means worked but was not able to cluster non-linear data (i.e., Spiral and Circles) as it does not identify intricate relationships between the data. With that in mind, it performed fine on linearly separable data like Blobs.

Nyström Kernel K-means:

Nyström approximation enhanced clustering precision for non-linear data such as Moon, Spiral, and Circles over Linear K-means. Although it is theoretically superior to Full Kernel K-means, its runtime was slower in practice. Surprisingly, Nyström Kernel K-means had the best accuracy (0.99) on the Moon dataset, indicating its ability to estimate the kernel matrix well.

Full Kernel K-means:
Full Kernel K-means consistently had competitive accuracy on all the datasets. Its running time was typically comparable to Nyström Kernel K-means, indicating the trade-off between accuracy and efficiency.

Conclusion/Future Work

This work compared the performance of Linear K-means, Full Kernel K-means, and Nyström Kernel K-means running datasets of different complexities. Although Linear K-means was superior in runtime, it performed terribly on non-linear clusters, such as those found in the Spiral and Circles datasets, hence showing weaknesses of this algorithm when it deals with complex data relationships. Nyström Kernel K-means, although designed as a trade-off between accuracy and runtime, outperformed its opponent, Full Kernel K-means, on some datasets-such as the Moon dataset-but ran slower than expected due to the fine-grained γ optimization (steps of 0.1) employed in our experiments. A coarser search, for example, using γ steps of 0.01, may preserve the accuracy and yield improved runtime when compared to Full Kernel K-means.

Contributions:

- Code: Tanush and Luke
- Abstract: Tanush
- Introduction: Tanush
- Related Work: Luke and Tanush
- Dataset and Features: Luke
- Methods: Luke
- Experiments/Results/Discussion: Tanush
- Conclusions/Future Work: Tanush and Luke

References:

Wang, S., Gittens, A., & Mahomey, M. W. (2019, February). *Scalable kernel K-means clustering with Nyström ...* Journal of Machine Learning Research. https://jmlr.org/papers/volume20/17-517/17-517.pdf

Oglic, D. & Gärtner, T.. (2017). *Nyström Method with Kernel K-means++ Samples as Landmarks. Proceedings of the 34th International Conference on Machine Learning, in Proceedings of Machine Learning Research* 70:2652-2660 Available from https://proceedings.mlr.press/v70/oglic17a.html.

Kullis, B., & Jordan, M. I. (2012). *Revisiting K-means: New algorithms via bayesian ...* International Conference on Machine Learning. https://icml.cc/2012/papers/291.pdf

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–283

Jin, X., & Han, J. (2011). K-Means Clustering. In C. Sammut & G. I. Webb (Eds.), Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_425

Pycodemates. (2022, October). The RBF kernel in SVM: A Complete Guide - PyCodeMates. https://pycodemates.com/2022/10/the-rbf-kernel-in-svm-complete-guide.html

Jones, A. (2025). Nyström approximation. andrewcharlesjones.github.io/journal/nystrom-approximation.html