K-MEANS FROM SCRATCH

By Sayed Mohammad Fatemi

SUMMER 2024

Isfahan University of Technology

TECH STACK AI BRANCH

Sayed Mohammad Fatemi
Student Number : 40127793
+989140015340
Telegram : @Lucimad
smf007smf007@gmail.com

IMPLEMENTING K-MEANS CLUSTERING FROM SCRATCH

History

The term "*k*-means" was first used by James MacQueen in 1967,[1] though the idea goes back to Hugo Steinhaus in 1956.[2] The standard algorithm was first proposed by Stuart Lloyd of Bell Labs in 1957 as a technique for pulse-code modulation, although it was not published as a journal article until 1982.[3] In 1965, Edward W. Forgy published essentially the same method, which is why it is sometimes referred to as the Lloyd–Forgy algorithm.[4]

Explanation

k-means clustering is one of the simplest and most commonly used clustering algorithms. It tries to find cluster centers that are representative of certain regions of the data. The algorithm alternates between two steps: assigning each data point to the closest cluster center, and then setting each cluster center as the mean of the data points that are assigned to it. The algorithm is finished when the assignment of instances to clusters no longer changes. The following example illustrates the algorithm on a synthetic dataset[5]:

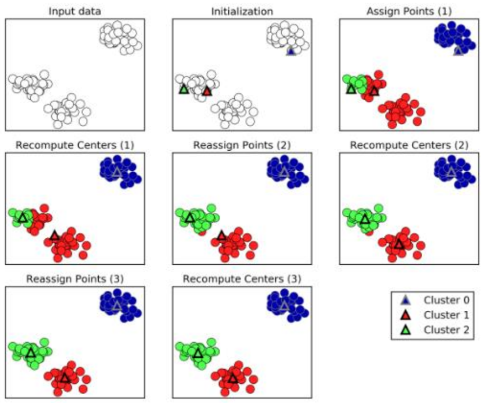

Figure 1 - Input data and three steps of the k-means algorithm

Sayed Mohammad Fatemi
Student Number : 40127793
+989140015340
Telegram : @Lucimad
smf007smf007@gmail.com

IMPLEMENTING K-MEANS CLUSTERING FROM SCRATCH

Code Example using scikit-learn[6]

```
from sklearn.cluster import KMeans
from sklearn.datasets import make_blobs
X, y = make_blobs([...]) # make the blobs: y contains the
cluster IDs, but we
# will not use them; that's what we want to predict
k = 3
kmeans = KMeans(n_clusters=k, random_state=42)
y_pred = kmeans.fit_predict(X)
```

Implementing from scratch

The source code is also included in a separate python file[7]. Some steps of this algorithm are :

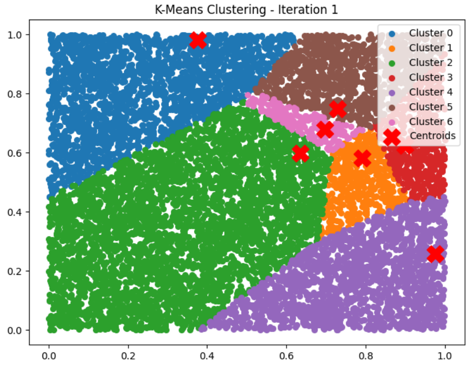

Figure 2 - step 1

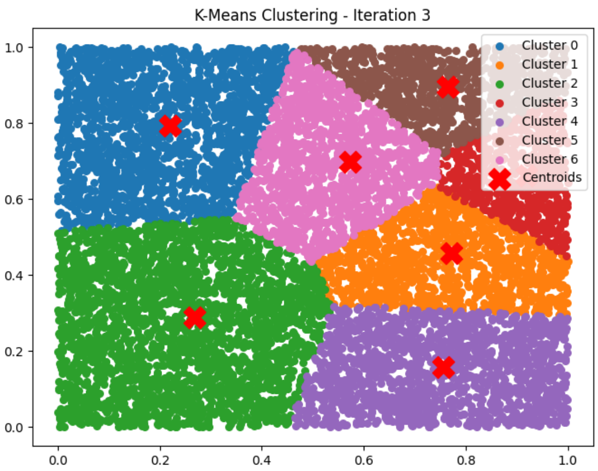

Figure 3 - step 3

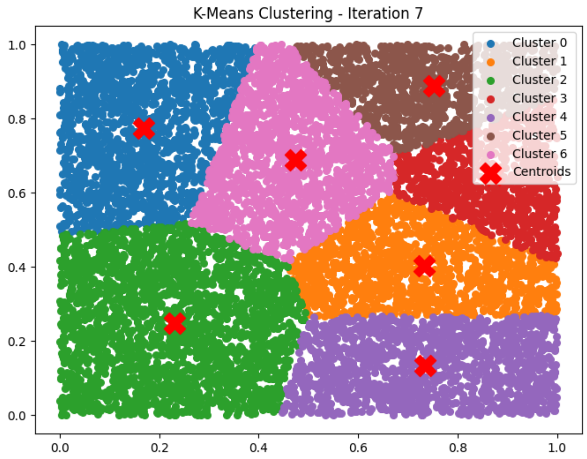

Figure 4 - step 7

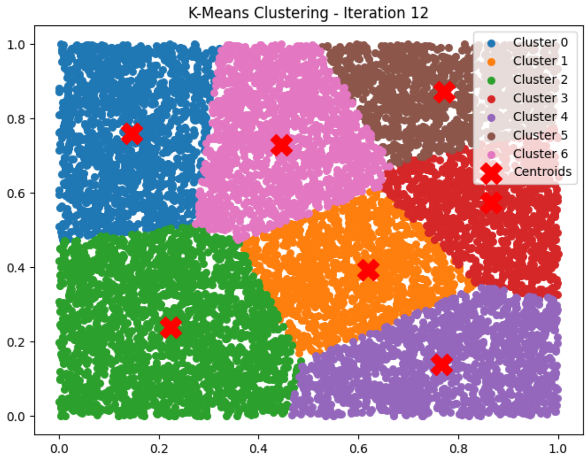

Figure 5 - step 12

TECH STACK AI BRANCH

Sayed Mohammad Fatemi
Student Number : 40127793
+989140015340
Telegram : @Lucimad
smf007smf007@gmail.com

IMPLEMENTING K-MEANS CLUSTERING FROM SCRATCH

IMAGE COMPRESSING USING K-MENAS

By changing the code and preparing it for loading images, we can use it to transform RGB images to 16-color images (so called compressing them). The source code is available in attached files. The outputs are shown in following pictures (iterations are limited to avoid very long time of execution) :

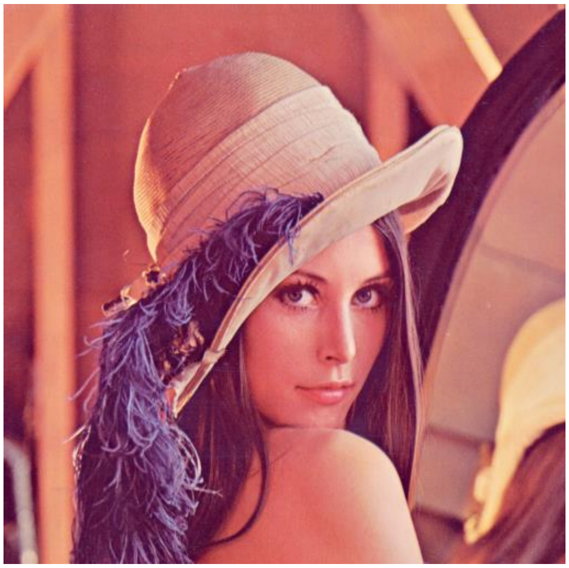

Figure 6 - original image

IMPLEMENTING K-MEANS CLUSTERING FROM SCRATCH

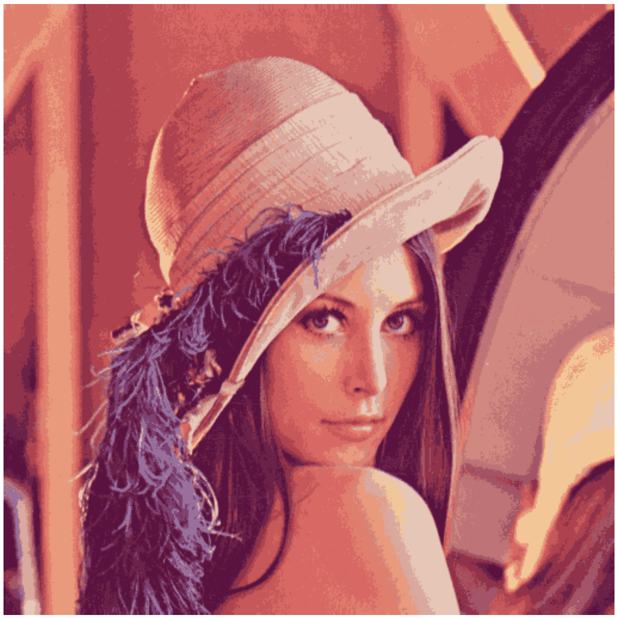

Figure 7 - compressed image

Execution Time : 1min 36s

Max Iterations : 300

L2 Norm : 3412782.115732

IMPLEMENTING K-MEANS CLUSTERING FROM SCRATCH

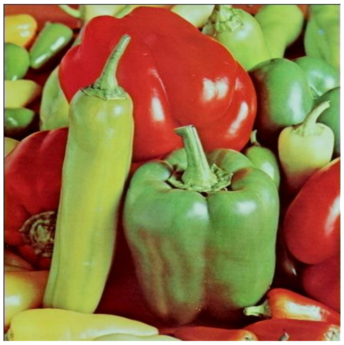

Figure 8 - original image

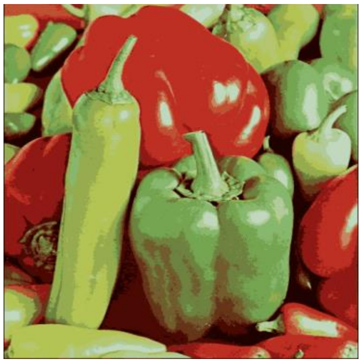

Figure 9 - compressed image

Execution Time : 42s, Iterations : 300, L2 Norm : 4562158.853435

IMPLEMENTING K-MEANS CLUSTERING FROM SCRATCH

OPTIMIZATION

There are a lot of ways that we can improve the execution time or quality of the this algorithm, however it is important to note that sometimes it is a trade-off between these two. Some optimization techniques are as follows:

- Initialization Techniques
  - E.g. : K-Means++: A smarter initialization method that spreads out initial cluster centers. This significantly improves the convergence speed and accuracy.[8]
- Distance Measures
  - E.g. : Euclidean Distance: The default, but other distance measures can be used depending on the data.[9]
- Efficient Computation

  E.g. : Elkan's Algorithm: This optimizes the distance computation by using the triangle inequality, which reduces the number of distance calculations.[10]
- Cluster Updating
  - E.g. : Mini-Batch K-Means: Processes a small random subset of data in each iteration, leading to faster convergence and scalability to larger datasets.[11]
- Stopping Criteria
  - Fixed Iterations: Run for a set number of iterations.
  - Convergence Threshold: Stop when the centroids no longer move significantly (i.e., the change in centroid positions is below a threshold).
  - Centroid Convergence: Stop when a certain percentage of centroids do not change between iterations.[12]
- Dimensionality Reduction
  - Principal Component Analysis (PCA): Reduces the dimensionality of the data while preserving as much variance as possible.
  - t-SNE: For non-linear dimensionality reduction, often used in visualizing clusters in high-dimensional data.[13]

IMPLEMENTING K-MEANS CLUSTERING FROM SCRATCH

- Parallel and GPU Computation
  - Parallel Processing: Distribute the computation across multiple processors.
  - GPU Acceleration: Leverage GPUs for parallel computation to accelerate K-Means clustering.[14]
- Avoiding Local Minima
  - Multiple Runs: Run K-Means multiple times with different initializations and choose the best result.
  - Simulated Annealing: Combine K-Means with simulated annealing to escape local minima.
  - Genetic Algorithms: Use genetic algorithms to optimize the selection of centroids.[15][16]
- Post-Processing
  - Refining Clusters: Post-process clusters to merge or split them based on additional criteria.
  - Cluster Refinement via Hierarchical Clustering: After initial K-Means clustering, refine clusters by applying hierarchical clustering techniques.[17]

OPTIMIZED IMPLEMENTATION

Now let’s implement the optimized version of K-means, using max iterations and K-means++ optimization. However remember that in small data, the difference won’t be that much different.

Using K-means++ we choose the initial centroids in a way that it converges better, so the final L2 Norm will improve, hence we get better image quality. Compare the following pictures with previous ones :

IMPLEMENTING K-MEANS CLUSTERING FROM SCRATCH

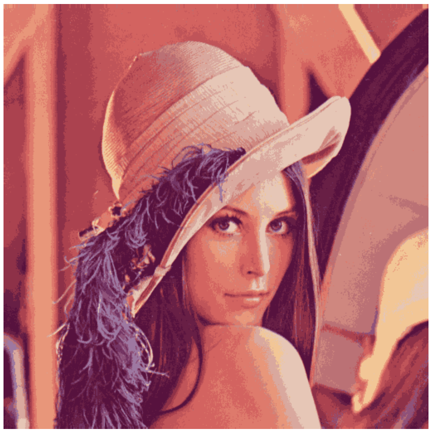

Figure 10 - lena optimized K-means

Execution Time : 36s
Max Iterations : 100
L2 Norm : 3415795.502945

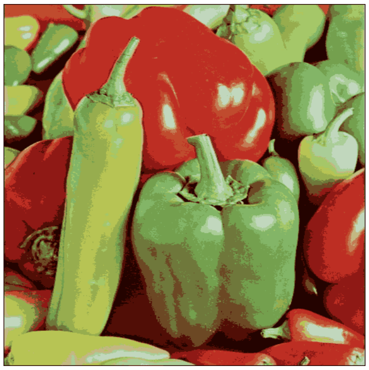

Figure 11 - peppers optimized K-means

Execution Time : 36s
Max Iterations : 100
L2 Norm : 4650520.628046

As you can see the L2 Norm for pictures has not changed a lot even though the iterations are reduced to 100 only, instead of 300.

Sayed Mohammad Fatemi
Student Number : 40127793
+989140015340
Telegram : @Lucimad
smf007smf007@gmail.com

IMPLEMENTING K-MEANS CLUSTERING FROM SCRATCH

Sources

1. *MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.*
2. *Steinhaus, Hugo (1957). "Sur la division des corps matériels en parties". Bull. Acad. Polon. Sci. (in French).*
3. *Lloyd, Stuart P. (1957). "Least square quantization in PCM". Bell Telephone Laboratories Paper.*
4. *Forgy, Edward W. (1965). "Cluster analysis of multivariate data: efficiency versus interpretability of classifications".*
5. [ML] Introduction to Machine Learning with Python (2017) book
6. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems
7. ChatGPT helped in this project too.
8. Arthur, D., & Vassilvitskii, S. (2007). *K-means++: The Advantages of Careful Seeding*. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms.
9. Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. (Chapter on clustering)
10. Elkan, C. (2003). *Using the Triangle Inequality to Accelerate k-Means*. Proceedings of the Twentieth International Conference on Machine Learning (ICML).
11. Sculley, D. (2010). *Web-Scale K-Means Clustering*. Proceedings of the 19th International Conference on World Wide Web (WWW '10).
12. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). *Data Clustering: A Review*. ACM Computing Surveys.
13. Jolliffe, I. T., & Cadima, J. (2016). *Principal Component Analysis: A Review and Recent Developments*. Philosophical Transactions of the Royal Society A.
14. Zhong, S., & Ghosh, J. (2005). *A Comparative Study of Generative Models for Document Clustering*. Machine Learning.

Sayed Mohammad Fatemi
Student Number : 40127793
+989140015340
Telegram : @Lucimad
smf007smf007@gmail.com

IMPLEMENTING K-MEANS CLUSTERING FROM SCRATCH

Sources

15. Kleinberg, J., Tardos, É. (2006). *Algorithm Design*. Pearson. (Section on optimization problems)
16. Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). *Optimization by Simulated Annealing*. Science.
17. Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis*. Wiley. (Chapter on hierarchical clustering)