A Comparative Analysis of K-Means using Pythagorean Means

Ahmed Abdelsalam; Advisor: David Allen, Ph.D.

Borough of Manhattan Community College

INTRODUCTION

- K-means is an unsupervised algorithm used for data clustering based on similarities.
- In most applications the arithmetic mean is used in the update rule in the algorithm.
- The study aims to compare and contrast different means (Geometric, Harmonic, and Arithmetic means) when the included in the K-means algorithm
- Rigorous mathematical analysis and testing are conducted to assess the effectiveness of each Pythagorean mean.
- The goal is to uncover the strengths and limitations of each Pythagorean mean in the context of K-means clustering.

DATASET

Gas Turbine CO and NOx Emission Data Set[1]

- The dataset contains 36,733 instances of 11 sensor measures aggregated over one hour from a gas turbine located in Turkey's northwestern region.
- The purpose of this dataset is to study flue gas emissions, specifically carbon monoxide (CO) and nitrogen oxides (NOx), which are the sum of nitrogen oxide (NO) and nitrogen dioxide (NO2).
- The data covers a period from January 1, 2011, to December 31, 2015, and includes gas turbine parameters (e.g., Turbine Inlet Temperature, Compressor Discharge pressure) along with ambient variables.
- Principal component analysis (PCA) was used on this dataset to reduce the dimensionality of its features from 11 features to two principal components for better visualization.

MATHEMATICAL BACKGROUND

$$x_1, x_2, \ldots x_n \in \mathbb{R}^d$$

Arithmetic Mean

$$\mu_A = \frac{1}{n}(x_1 + x_2 \ldots + x_n)$$

Geometric Mean

$$\mu_G = \sqrt[n]{x_1 x_2 \ldots x_n}$$

Harmonic Mean

$$\mu_H = \frac{n}{(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n})}$$

CHOOSING THE OPTIMAL K: GAP STATISTICS

The gap statistics method[4] is a popular approach used to determine the optimal number of clusters, K, in the K-means algorithm. The method compares the within-cluster dispersion of data with an expected null reference distribution to assess if the clusters obtained with a particular K are significantly better than randomly distributed data.

Gap Statistics Formula[3]

$$Gap(k) = \frac{1}{B}\sum_{i=1}^{B} log(W_k^{(i)}) - log(W_k)$$

- ***K*** **= number of clusters being evaluated**
- ***W*** **= within-cluster dispersion of the original data for the given *K***
- W_b **= within-cluster dispersion of the randomly generated reference data for the given *K*, averaged over *B* random reference datasets**

Decay

$$\text{Decay(k)} = \log(W^{(i)}_k) - \log(W_k)$$

K-MEANS ALGORITHM

Goal: Compute

$$\underset{i}{\operatorname{argmin}} \parallel x_i - \mu_j \parallel : j = A, H, G$$

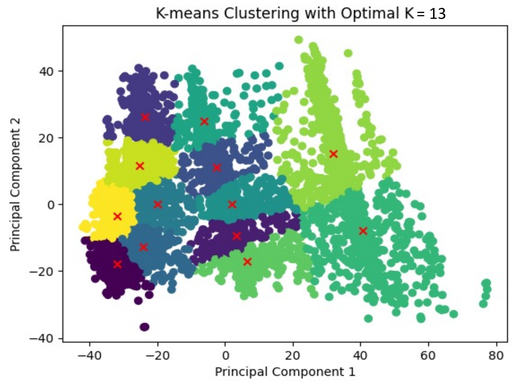
The primary objective of the algorithm is to find the best possible partitioning of the data, where each data point is assigned to the cluster whose centroid (mean) is closest to it. K-means[2] is an iterative algorithm and typically converges to a stable clustering solution.

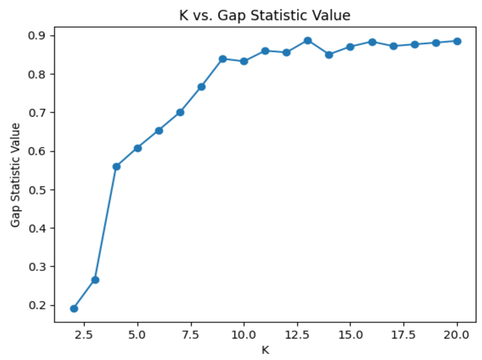
Algorithm Steps:

1. Gap Statistics for K Selection: Before the K-Means algorithm initializes, we use the Gap Statistics to determine an optimal K.
2. Initialization: With the optimal K determined using Gap Statistics, we randomly select K data points as the initial centroids. These centroids will serve as the centers of the initial clusters.
3. Assignment: Each data point is assigned to the nearest centroid based on computing $\underset{i}{\operatorname{argmin}} \parallel x_i - \mu_j \parallel : j = A, H, G$. This step ensures that each data point belongs to the cluster whose centroid is closest to it.
4. Update Centroids: After the assignment step, we recalculate the centroids of each cluster. The new centroid of a cluster is computed as the mean (average) of all the data points currently assigned to that cluster.
5. Convergence Check: Steps 3 and 4 are iteratively repeated until the centroids no longer change significantly or until a predetermined number of iterations is reached. This process continues until the centroids stabilize, indicating convergence, and yielding the final clustering solution.

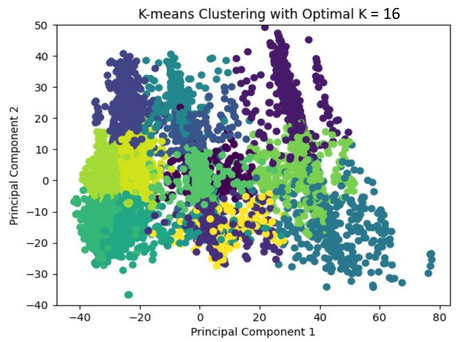
RESULTS

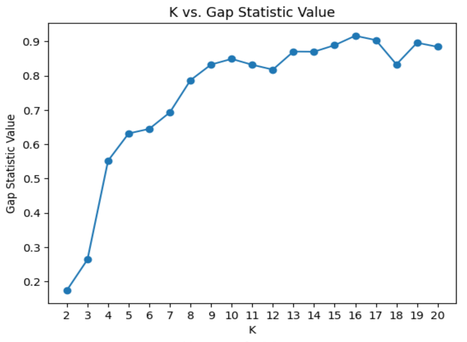
Arithmetic Mean



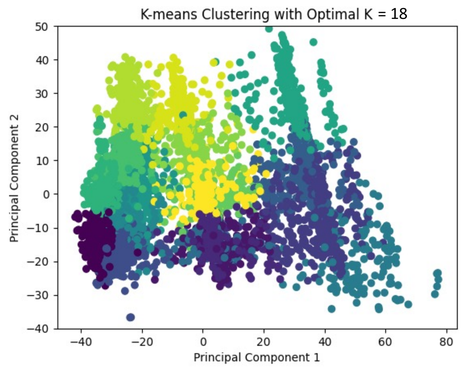


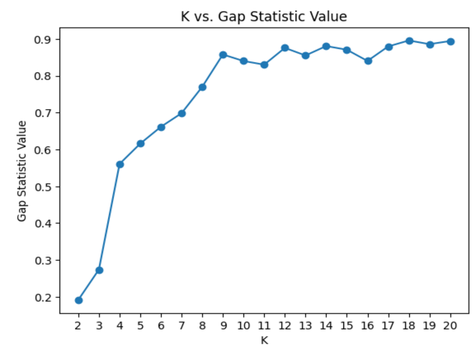
Geometric Mean





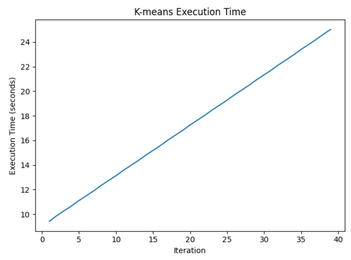
Harmonic Mean





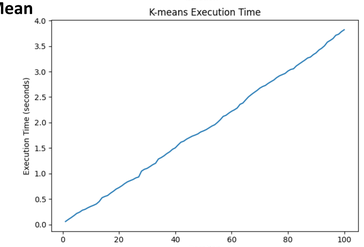
Runtimes

Arithmetic Mean



Geometric Mean

Harmonic Mean



CONCLUSIONS

In conclusion, our Comparative Analysis of K-Means using Pythagorean Means revealed that the Arithmetic means produces the most clearly defined clusters. These clusters demonstrated superior compactness and separation.

Furthermore, we observed that the Harmonic mean exhibits resilience to outliers, highlighting its potential in scenarios where outlier-robust clustering is essential. While it may not have matched the clustering quality of the arithmetic mean in our study, its immunity to certain outliers is a valuable characteristic.

Acknowledgements

I am immensely grateful for the invaluable guidance and mentorship provided by Professor David Allen throughout this project. Working with him has been a truly enriching experience, and I am thankful for the opportunity to learn and grow under his supervision.

References

[1] "Gas Turbine CO and NOx Emission Data Set," UCI Machine Learning Repository. https://doi.org/10.24432/C5WC95.

[2] A. McDonald, "How to use Unsupervised Learning to Cluster Well Log Data using Python," Medium, Jun. 03, 2021. https://towardsdatascience.com/how-to-use-unsupervised-learning-to-cluster-well-log-data-using-python-a552713748b5 (accessed Aug. 02, 2023).

[3] C. Yuan and H. Yang, "Research on K-Value Selection Method of K-Means Clustering Algorithm," Jun. 18, 2019.

[4] S. Hayasaka, "How Many Clusters?," *Medium*, Feb. 11, 2022. https://towardsdatascience.com/how-many-clusters-6b3f220f0ef5

[5] M. Garbade, "Understanding K-means Clustering in Machine Learning," Towards Data Science, Sep. 12, 2018. https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1