Johns Hopkins
Whiting School of Engineering

Algorithms for Data Science

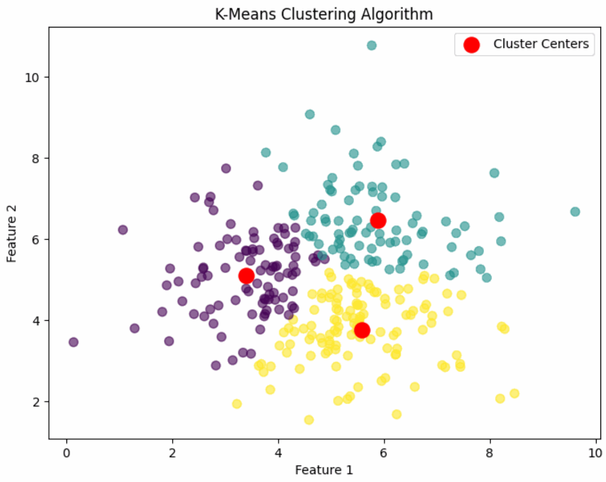
Unsupervised Learning: K-Means Clustering

K-Means Clustering Overview

K-Means is an iterative algorithm that partitions a dataset into k clusters by minimizing the within-cluster sum of squares (WCSS).

Key Steps

1. Assign each data point to the nearest cluster centroid.
2. Update centroids to be the mean of assigned points.



Applications

Document Clustering and Market Segmentation

K-Means: Mathematical Formulation

- **Objective Function:**
  - Minimize the WCSS*:

$$WCSS = \sum_{i=1}^{N} \min_{j \in \{1, \ldots, k\}} \left\| x_i - c_j \right\|^2$$

Where:

- x_i: Data point i
- c_j: Centroid of cluster j
- k: Number of clusters

- **Centroid Update Formula:**

$$c_j = \frac{1}{|S_j|} \sum_{x_i \in S_j} x_i$$

Where:

- S_j: Set of points in cluster j

***Minimizing WCSS ensures that points within clusters are as similar as possible.**

K-Means Clustering Algorithm Analysis

1. **Initialize:** Choose k centroids randomly from the dataset. → $O(k)$
2. **Repeat Until Convergence:**
   i. **Assign:** For each data point x_i, assign it to the nearest centroid. →

      For N points, compute distance to k centroids: $O(k \times d)$
      Total: $O(N \times k \times d)$

   ii. **Update:** For each cluster j, calculate the new centroid c_j as the mean of all points assigned to it. →

      For k centroids, calculate mean of assigned points: $O(N \times d)$
      Total: $O(k \times N \times d)$

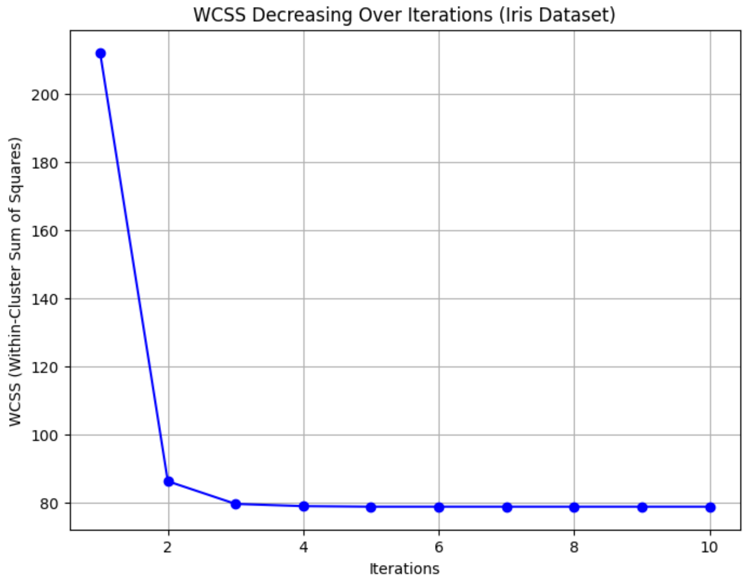
3. **Convergence:** Stop when centroids no longer move or cluster assignments stabilize.

Repeat over I iterations: $O(I \times N \times k \times d)$

K-Means Clustering: Correctness

Theorem: K-Means converges to a local minimum of WCSS.

- **Proof:**
  - The assignment step reduces WCSS by assigning points to the nearest centroid.
  - Update step recalculates centroids to further minimize WCSS within clusters.
  - WCSS decreases monotonically, ensuring convergence.



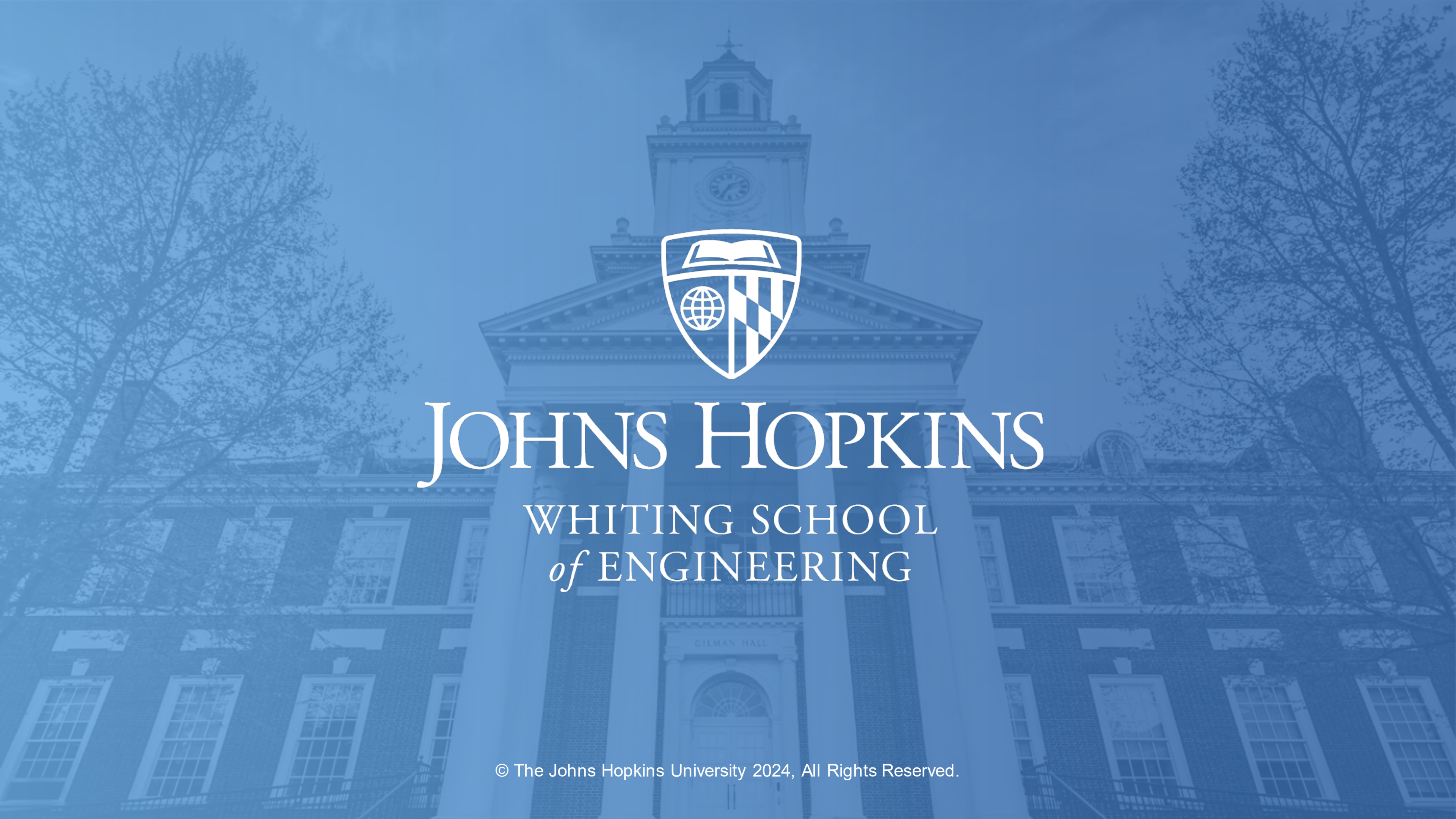
Advantages and Limitations

Advantages

- **Simplicity:** Easy to implement and understand.
- **Scalability:** Works efficiently for moderate-sized datasets.
- **Flexibility:** Applies to diverse data types.

Limitations

- **Initialization Sensitivity:** Results depend on initial centroids.
- **Cluster Shape Assumption:** Assumes clusters are spherical.
- **Fixed Clusters:** Requires k to be predefined.


JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING