Impact of Normalization on K-means Clustering

Raghda Al-taei

2024

1 Introduction

K-means is a popular clustering algorithm that partitions a dataset into k clusters by minimizing the sum of squared distances between data points and their respective cluster centers. However, the results of K-means can be influenced significantly by the scale of features in the dataset. This document explores the impact of normalization on K-means clustering and provides a comparison between clustering on raw and normalized data.

2 Problem Statement

We apply K-means clustering on a synthetic dataset using two different approaches:

1. Clustering without any preprocessing on the raw data.
2. Clustering after normalizing the data, i.e., scaling all features to the range [0, 1].

We investigate whether the clustering results from these two approaches are identical and discuss the reasons behind any differences.

3 Importance of Normalization

Normalization is a preprocessing step that scales the features of a dataset to a common range, typically [0, 1]. It ensures that each feature contributes equally to the distance calculations used in K-means clustering. Without normalization, features with larger ranges can dominate the distance metrics, leading to biased clustering results.

3.1 Normalization Formula

The formula for Min-Max normalization is:

$$X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Where:

- X is the original feature value.
- X_{min} is the minimum value of the feature.
- X_{max} is the maximum value of the feature.

4 Results

Figure 1 shows the clustering results of both approaches:

- **Left Plot**: Clustering without normalization. The clusters are biased towards features with larger ranges.
- **Right Plot**: Clustering with normalization. Each feature contributes equally, resulting in more balanced clusters.

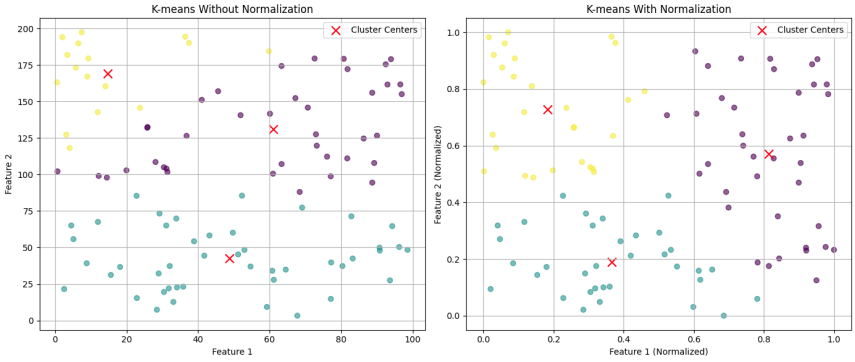

Figure 1: Comparison of K-means Clustering Results: Without Normalization (Left) vs. With Normalization (Right)

5 Conclusion

The results demonstrate that K-means clustering can produce significantly different outcomes depending on whether the data is normalized. Normalization ensures that all features have a balanced influence on the clustering process, leading to more meaningful clusters. Therefore, it is a crucial preprocessing step when applying K-means to datasets with features of varying scales.