Unsupervised Learning Using K-Means Clustering on California Housing Data

Omid Saberi
Dept. of Industrial Automation
University West
Trollhättan, Sweden
https://orcid.org/0009-0005-3596-9182

***Abstract*—This report explores the application of K-Means clustering on the California Housing dataset to analyze geographic and socioeconomic patterns in the housing market. The dataset was clustered using latitude, longitude, and median income, with the optimal number of clusters determined by silhouette score. The results revealed two distinct clusters representing southern and northern California, which differ significantly in median income. These findings offer insights into the regional economic segmentation of the housing market, highlighting the utility of unsupervised learning for housing data analysis.**

***Index Terms*—K-Means clustering, unsupervised learning, California Housing dataset, silhouette score, geographic segmentation**

I. Introduction

K-Means clustering is a fundamental algorithm in unsupervised learning, designed to partition data into distinct groups or clusters based on their inherent similarities. Unlike supervised learning, where labeled data is used to train models, unsupervised learning aims to discover hidden patterns within unlabeled data [2]. The K-Means algorithm accomplishes this by minimizing the variance within each cluster while maximizing the variance between clusters, ultimately providing a segmentation of the dataset [3].

In this assignment, K-Means clustering is applied to the California Housing dataset, focusing on geographic (longitude and latitude) and economic (median income) features. The goal is to identify meaningful clusters in the data that correspond to different segments of the housing market. By optimizing the number of clusters using the Silhouette score, this analysis aims to reveal patterns in housing characteristics, providing valuable insights into how various regions and income levels impact the housing landscape in California.

II. K-Means Clustering and Evaluation

The K-Means clustering algorithm was applied to the California Housing dataset. The features selected for clustering were latitude, longitude, and median_income, which are relevant indicators of geographic and socioeconomic factors affecting housing prices.

A. Data Preprocessing

Before applying K-Means, the data was preprocessed. Missing values in the median_income feature were imputed using the median strategy, and a log transformation was applied to normalize the median_income data. This transformation ensures that the K-Means algorithm operates effectively on a dataset with a more normal distribution. The log-transformed values were then scaled to standardize the dataset, which helps to reduce the influence of different feature scales on clustering.

B. Hyperparameter Tuning

The optimal number of clusters was determined using GridSearchCV and silhouette score as the evaluation metric. The silhouette score measures how similar an object is to its own cluster compared to other clusters, with higher values indicating better-defined clusters. A range of clusters from 2 to 10 was evaluated, and the optimal number of clusters (k=2) was selected based on the highest silhouette score, which was approximately 0.666.

III. Results and Analysis

A. Cluster Visualization

The K-Means clustering was applied with the optimal number of clusters (k=2), and the results were visualized using a scatter plot as seen in Fig. 1. The plot shows the geographic distribution of the clusters based on the latitude and longitude of the housing data. The two clusters divide the dataset into two main geographic regions, which could represent different housing markets.

Cluster 0 primarily encompasses southern California, including areas like Los Angeles, where median income values are lower on average. Cluster 1 represents northern California, including regions around San Francisco, which are associated with higher median incomes.

B. Cluster Analysis

The cluster centers and mean values for each cluster provide additional insight into the socioeconomic and geographic segmentation of the housing market. The following observations were made:

- Cluster 0 is centered around a longitude of -118.006 and latitude of 33.939, representing southern California. The

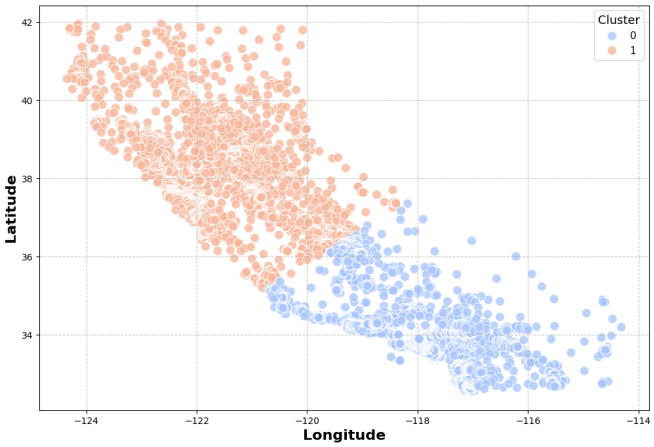

Fig. 1. Clusters of Housing Data

average log-transformed median_income for this cluster is slightly higher, indicating regions with relatively higher incomes.

- Cluster 1 is centered around a longitude of -121.715 and latitude of 37.955, representing northern California. This cluster has a lower average log-transformed median_income, suggesting that these regions have lower income levels.

These results highlight the geographic and economic divide in the housing market, where certain areas may be more affordable, while others are more expensive, aligning with the expected patterns based on the selected features. Table I displays the cluster centers calculated by the KMeans algorithm, including longitude, latitude, and median income for each cluster's central point. In contrast, Table II shows the average values for all members of each cluster, computed from all data points assigned to those clusters. The variations in median incomes for Cluster 0 and Cluster 1 between the two tables result from the fact that the cluster centers represent computed geometric means, while the averages reflect the mean values of all assigned data points, leading to slight discrepancies.

IV. Conclusion

The application of K-Means clustering to the California Housing dataset revealed distinct patterns in geographic and economic segmentation. The optimal number of clusters, determined by silhouette score, was 2, effectively dividing the dataset into two regions: southern and northern California. This clustering highlights important differences in median income across these regions and can be used as a basis for further analysis of the housing market. The results suggest that K-Means clustering is a useful tool for segmenting housing data based on geographic and socioeconomic features, allowing for deeper insights into housing affordability and market segmentation.

Future work could include comparing these results with other clustering algorithms such as DBSCAN, or further exploring the impact of additional features like housing price and population density on the clustering outcome.

TABLE I
Cluster Centers

Cluster	Longitude	Latitude	Median Income
0	33.939349	-118.006597	0.027841
1	37.955150	-121.715360	-0.038217

TABLE II
Average for the Whole Cluster

Cluster	Latitude	Longitude	Median Income
0	33.939467	-118.006805	0.027822
1	37.955450	-121.715501	-0.038199

References

[1] O. Saberi, "IAI600Lab6.ipynb," GitHub repository, GitHub, Oct. 23, 2024. [Online]. Available: https://github.com/omisa69/IAI600Lab6

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[3] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.