OLA TRIPS CLUSTERING

MINI-PROJECT REPORT

Submitted By :

HRISHITA KISHORE
Roll No. : 2201096
Group: CS31

COURSE : Machine Learning Lab(CS360)
INSTRUCTOR : Upasana Talukdar

Problem Statement

In urban areas, ride-sharing platforms like Ola generate a large volume of trip data every day. Analyzing these trips based on geographical locations and the time of day can provide valuable insights into the patterns of transportation. For instance, certain locations may have higher numbers of trips during specific times, such as school trips in the morning or market visits in the evening.

The goal of this mini-project is to apply machine learning clustering techniques to analyze and categorize trips in terms of their geographical locations and the time of day. By clustering trips, we can derive patterns that would help in identifying key trip types (e.g., school trips, office commutes, market visits, etc.). This report focuses on using clustering algorithms to group trips based on the number of trips in a particular geographical area and the time of day, and subsequently categorizing the trip types based on the dominant clusters.

Methodology

Data Collection and Preprocessing

The dataset for this project consists of 500 trip records from the Ola platform, which includes details such as Trip ID, Origin, Destination, Latitude and Longitude for both origin and destination, Start Time, and Trip Duration. To ensure the dataset was suitable for clustering, several preprocessing steps were performed:

- **Data Cleaning**: Missing or incomplete data was handled through imputation or removal of outliers.
- **Standardization**: Certain features like geographical coordinates and trip duration were standardized to ensure uniform scaling, which is crucial for K-Means clustering.
- **Time Feature Transformation**: The start time was extracted and converted into numerical features, such as the hour of the day. This allowed us to categorize the trips into time periods like morning, afternoon, and evening, which are significant in identifying trip patterns.

Feature Engineering

To effectively apply clustering, the following transformations were made to the dataset:

- **Geospatial Data**: The geographical coordinates of trips (origin and destination) were used to compute the Euclidean distance between each pair of coordinates, which helps in grouping trips geographically.

- **Time of Day**: The time at which the trip started (in 12-hour AM/PM format) was transformed into a 24-hour scale to represent different parts of the day (e.g., early morning, office hours, evening).

K-Means Clustering Algorithm

K-Means is an unsupervised learning algorithm used to partition data into K distinct clusters. Each cluster is represented by its centroid, and the goal is to minimize the within-cluster variance (i.e., the sum of squared distances between each point and its cluster center). The steps for applying K-Means clustering in this project are as follows:

1. **Initialization**: Initially, K centroids are chosen randomly, or using some heuristic.
2. **Assignment**: Each data point is assigned to the nearest centroid based on Euclidean distance.
3. **Update**: The centroids are recalculated by taking the mean of all data points assigned to each centroid.
4. **Iteration**: Steps 2 and 3 are repeated iteratively until the centroids no longer change significantly, or a maximum number of iterations is reached.

Determining the Optimal Value of K

One of the most critical steps in K-Means clustering is selecting the appropriate value for K (the number of clusters). To do this, we used the **Elbow Method**, a common technique for choosing the optimal number of clusters. Here's how it works:

Elbow Method: The Elbow Method involves plotting the **Within-Cluster Sum of Squares (WCSS)** or the **Inertia** (the sum of squared distances of samples to their closest centroid) for different values of K. As K increases, WCSS will decrease because there are more centroids to fit the data. However, after a certain point, the decrease in WCSS starts to slow down, forming an "elbow" shape in the plot. The optimal value of K is typically chosen at this "elbow," where the rate of improvement slows down, indicating a reasonable trade-off between the number of clusters and the clustering performance.

By plotting the inertia for different values of K, you can visualize the point where the elbow occurs and select that value as the optimal K.

Evaluating Cluster Quality with Silhouette Score

Once K-Means clustering is applied, the next step is to evaluate the quality of the resulting clusters. One of the most commonly used metrics for evaluating clustering quality is the **Silhouette Score**. The Silhouette Score measures how similar each point is to its own cluster (cohesion) compared to other clusters (separation).

Silhouette Score Formula: For each data point iii, the Silhouette Score s(i)s(i)s(i) is calculated as:

$$S(i) = b(i)-a(i)/max(a(i),b(i))$$

where:

- a(i) is the average distance from point iii to all other points within the same cluster (cohesion).
- b(i) is the minimum average distance from point iii to all points in any other cluster (separation).

  The Silhouette Score ranges from -1 to +1:

  - A score close to +1 indicates that the data point is well clustered and far from other clusters.
  - A score close to 0 indicates that the point is on or near the decision boundary between clusters.
  - A score close to -1 indicates that the point is incorrectly clustered, as it is closer to a different cluster than its own.

  The average Silhouette Score for all points in the dataset provides an overall measure of clustering quality. A higher average score indicates better-defined and well-separated clusters.

Cluster Labeling and Interpretation

Once the optimal K is selected and the clustering is complete, we label the clusters by analyzing the trip patterns in each one. Based on the geographical locations (origin and destination coordinates) and time of day, each cluster is associated with a particular type of trip:

- **Morning Clusters**: Typically have trips originating from residential areas and heading to schools or offices.
- **Evening Clusters**: Feature trips concentrated around shopping centers, markets, and leisure areas.
- **Peak Hour Clusters**: Show high concentrations of trips around office hubs during office hours, indicating peak demand.

These labels help to interpret the clusters in a meaningful way, providing insights into the typical trip types occurring at different times and locations.

Result

After applying the K-Means clustering algorithm and evaluating the results, the following key observations were made:

1. **Clustering Insights**: The trips naturally grouped into distinct clusters based on their geographical proximity and the time of day. The clusters with the highest

number of trips were associated with locations such as city centers, shopping areas, and residential neighborhoods.

2. **Trip Types**: Based on the clusters, it was possible to identify patterns:

   - **Morning Clusters**: These clusters showed a high concentration of trips originating from residential areas and heading to schools or offices, indicating school commutes and office rush hours.

   - **Evening Clusters**: Trips in the evening were concentrated around shopping malls, markets, and entertainment zones, indicating leisure or market visits.

   - **Peak Hour Clusters**: During peak office hours, trips were predominantly clustered around office hubs, signaling high demand for commutes to and from work.

3. **Application**: The insights gained from this clustering can be applied to optimize ride-sharing services, by predicting peak demand times and areas. This could help in better resource allocation, reducing wait times, and improving the overall user experience for Ola riders.
4. **Future Work**: Future improvements could involve refining the clustering process by incorporating additional features such as traffic patterns or weather conditions, which could affect trip durations and trip demand. Additionally, implementing more advanced clustering algorithms like DBSCAN could help in discovering more complex patterns and outliers.
