EPOCH

This is our Final Submission for EPOCH 2024 Hackathon Problem Statement 4

Challenges Encountered

1. **Understanding the Distance Expression**:
    - Understanding the distance expression posed an initial hurdle.We Had to go into the math and figure out what it meant for our problem.
2. **Understanding the Facts and Studying Various Clustering Models to Determine the Best Fit**:
    - Gaining a thorough understanding of the problem's underlying facts and evaluating different clustering models presented a notable challenge.
3. **Finding the Learning Rate for Latitude, Longitude, and Theta in Our Model**:
    - Identifying the appropriate learning rates for latitude, longitude, and theta parameters in our model posed another significant challenge.
4. **Dealing with a 5-Dimensional Matrix and Implementing it in Code**:
    - Managing and coding a 5-dimensional matrix presented a considerable hurdle. This required careful consideration of structures of Matrix and implement strategies to effectively handle the complexity.
5. **Initializing Multiple Loops and Estimating Time Requirements**:
    - The process of initializing multiple loops and estimating the time incurred before running the code was a challenge in itself, Because model was taking too much time to run.

Key Decisions Made

1. **Choosing Agglomerative Clustering**:
    - The decision to utilize Agglomerative Clustering was best in our approach. This clustering method was selected for its ability to effectively segment the data into distinct clusters based on proximity, which aligned well with our objective of region segmentation.
2. **Selection of Values for Theta, Latitude, and Longitude as Learning Rate Parameters**:

- Determining the appropriate values for the learning rate parameters, namely theta, latitude, and longitude, was crucial for the convergence and effectiveness of our model.

3. **Translation and Rotation Across the Entire Graph for Each Cluster**:
    - The approach involving translating and rotating along the entire graph for each cluster formed, enables a thorough search for the optimal solution.

Proposed Solution

Overview

The proposed solution aims to segment a set of datapoints into k regions and find optimized lines for each cluster. This process involves two main steps: data segregation using agglomerative clustering and line optimization through self regression model.

Steps

Data Segregation using Agglomerative Clustering

1. Given k as the number of lines to be laid down, the first step involves segregating the datapoints into k regions.
2. Agglomerative clustering, a hierarchical clustering technique, is employed for this purpose.
3. This method starts by considering each data point as a separate cluster and then iteratively merges the closest pairs of clusters until k clusters are formed.

Reasons to Use Agglomerative Clustering Agglomerative clustering is chosen for data segregation due to the following reasons:

- **Hierarchical Nature**: Agglomerative clustering creates a hierarchy of clusters, which can be useful for understanding the data structure at different levels of granularity.
- **Flexibility**: It allows for the incorporation of various distance metrics and linkage criteria, making it adaptable to different types of data and clustering objectives.
- **Ease of Implementation**: Agglomerative clustering is relatively straightforward to implement and interpret, making it suitable for prototyping and exploratory data analysis tasks.

Cluster Optimization for Line Placement

1. Once the data is segmented into k clusters, the next step is to find the optimal line for each cluster.
2. A self regression model is employed for this optimization process.
3. Equally spaced points are selected on the plane of points, denoted as b.
4. For each of these points, the line is rotated around it to explore various orientations.
5. Subsequently, the cost of each line is calculated.

Cost Calculation and Line Selection

1. The cost of each line configuration is determined based on predefined criteria dist
2. The line configuration that yields the minimum cost among all possibilities is selected as the target line for the respective cluster.

Conclusion

This approach ensures that each cluster is associated with an optimized line placement, facilitating effective data representation or separation based on the problem requirements.