Ahmedabad University

CSE523 Machine Learning

Weekly Report - 2

Project 5: Identify Hard stop and momentary stop using vehicle trajectory dataset.

Submitted to faculty: Mehul Raval

Date of Submission: 16-03-24

Roll No.	Name of the Student
AU2140040	Kathan Thakkar
AU2140171	Harsh Pandya
AU2140224	Dhruvi Rajput
AU2140230	Yax Prajapati

AIM OF THE WEEK:

Data set cleaning, algorithm application and trying to find optimal values of the parameters of the DBSCAN.

Introduction:

In this week's report, our focus lies on the essential stages of data analysis: dataset cleaning, algorithm application, and parameter optimization. Specifically, we delve into the process of refining raw data sets to ensure accuracy and consistency, applying DBSCAN algorithm for clustering tasks, and employing techniques to determine optimal parameter values for enhanced performance. By addressing these critical components, we aim to streamline our data analysis pipeline and unlock valuable insights from complex datasets.

Data cleaning:

In our data cleaning phase, we meticulously selected the essential features required for our analysis, including track ID, x and y coordinates, and v_smoothened. To ensure data integrity, we systematically removed redundant entries characterized by identical values across all features, specifically focusing on instances indicative of hard stops. By eliminating these redundant data points, we aimed to refine our dataset, mitigating potential biases and optimizing the accuracy of subsequent analyses. This meticulous curation sets a robust foundation for our ongoing data exploration and algorithmic application, ultimately enhancing the reliability and efficacy of our insights.

Writing and testing code:

In this week's report, we undertook the implementation of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to extract clusters from our dataset. Leveraging Python code, we conducted iterative tests varying the crucial parameters of epsilon (ε) and minimum sample size to identify their impact on clustering performance. Through rigorous training and testing, we analyzed the resultant clusters and their corresponding ranges of coordinates. This comprehensive evaluation provided valuable insights into the algorithm's sensitivity to parameter settings, enabling us to fine-tune our approach for optimal clustering outcomes. By systematically probing the interplay between epsilon and minimum sample size, we gained a deeper understanding of their influence on cluster formation, facilitating informed decision-making for subsequent data analysis tasks.

Link of Code:

https://colab.research.google.com/drive/1is6kfMFN7XRONoCOnq10Ooa95XdqHRUR#scrollTo=ybzs4xnYJUmO

Inferences:

```
Cluster -1:
x range: 241 - 3437
y range: 70 - 1858

Cluster 2:
x range: 2648 - 2814
y range: 599 - 637

Cluster 3:
x range: 2058 - 2220
y range: 593 - 681

Cluster 0:
x range: 1444 - 1587
y range: 1267 - 1366

Cluster 1:
x range: 2401 - 2740
y range: 412 - 452

Cluster 5:
x range: 1855 - 1944
y range: 136 - 328

Cluster 4:
x range: 1845 - 1931
y range: 443 - 789
```