Problem: Road Trajectory Prediction

In the rapidly evolving process of urbanization, traffic problems are becoming increasingly prominent. Accurately predicting vehicle trajectories is crucial for urban traffic management, reducing congestion, preventing accidents, improving transportation efficiency, and reducing energy consumption and environmental pollution. By analyzing vehicle trajectory data, we can understand driver behavior patterns, identify traffic hotspots, and optimize traffic signal control. These data can also support autonomous driving technology, shared mobility services, and smart city planning.

In recent years, with the proliferation of sensor technology and mobile devices, a vast amount of traffic data has been collected and stored. GPS trajectory data, as an essential component, records the movement paths of vehicles on urban roads. By analyzing these data, we can effectively predict the future paths of vehicles, providing decision support for traffic management departments and mobility services.

The goal of this competition is to leverage historical GPS trajectory data and external data to build a predictive model that accurately forecasts vehicle movement trajectories over a specified future period. Participants will process and analyze these data, applying machine learning and time series analysis methods to solve real-world traffic problems and explore how data-driven approaches can enhance the intelligence of urban transportation.

To accomplish this task, the following datasets are provided:

1) "*Microsoft T-Drive Trajectory Dataset*": This dataset, collected by Microsoft Research Asia, contains GPS trajectory data of 10,357 taxis in Beijing, recording vehicle movements over a specific period. The dataset includes a total of 15 million GPS points. Each trajectory record contains the following information:
    - Taxi ID: A unique identifier for each taxi
    - Timestamp: The time of the GPS point
    - Latitude: The latitude of the GPS point
    - Longitude: The longitude of the GPS point

2) "*Beijing_sevenpart Dataset*": The dataset is comprised of seven parts of data that were extracted from the GPS trajectories of taxicabs, road networks, POIs of Beijing, and video clips recording real traffic on roads

3) "*PeMS04*": The dataset is a traffic dataset collected from 307 sensors over a continuous period of 59 days, starting from January 1, 2018. The sensors recorded traffic data every 5 minutes. The shape of the raw traffic data file, data.npz, is (307, 16992, 3).

4) "*PeMS08*": The dataset contains traffic data from eight highways in San Bernardino collected between July and August 2016. The data were recorded by 1979 sensors every 5 minutes, primarily capturing the number of vehicles passing by. The dataset also includes a 3x107 adjacency matrix file, representing the connectivity and distances between 107 intersections.

Tasks

Utilize historical GPS trajectory data and any additional external data to predict vehicle movement trajectories over a future period. The goal is to build a predictive model capable of forecasting vehicle movements over a specified future period.

- Data Cleaning: Clean the data to remove noise and outliers, handle missing values, and remove duplicates.
- Data Integration: Integrate multiple data sources, including historical GPS trajectory data and external data (such as road networks and traffic conditions).
- Data Encoding: Properly encode data elements like time, latitude, and longitude to ensure data consistency.
- Feature Extraction: Extract useful features from the raw data, including temporal features (e.g., hour of the day, day of the week, holidays) and spatial features (e.g., road network).
- Feature Engineering: Further optimize and select features to enhance model performance.
- Model Selection: Choose appropriate machine learning or deep learning models for prediction, including but not limited to linear regression, random forests, gradient boosting, LSTM, etc.
- Model Training: Train the model using the training set and fine-tune the model's hyperparameters to improve prediction accuracy.
- Model Validation: Validate the model using the validation set, evaluate its performance, and make necessary adjustments.
- Model Testing: Evaluate the final performance of the model using the test set and compute evaluation metrics.

Evaluation Method

- Evaluation Metrics: Use metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R-squared (R^2) to evaluate the performance of the predictive model.
- Evaluation Process: The dataset split into training, testing and validation sets. Establish and

evaluate the model using these datasets. The evaluation results should include the specific values of each evaluation metric.

Submission Requirement

Your PDF solution of no more than 20 total pages should include:

- One-page Summary Sheet.
- Table of Contents.
- Your complete solution.
- One page Article for *Optimal Road Trajectory* magazine.
- Reference List.

Note: Each team must submit a **20-page limit report** summarizing their model and methodology, explaining the application scenarios and potential impact of the prediction results. All aspects of submission count toward the 20-page limit (Summary Sheet, Table of Contents, Reference List, and any Appendices). You must cite the sources for your ideas, images, and any other materials used in your report.

Attachments

We provide the following four data files for this problem.

1. **T-drive Taxi Trajectories.zip**
2. **beijing_sevenpart.zip**
3. **PeMS04.rar**
4. **PeMS08.rar**

Descriptions

1. T-drive Taxi Trajectories

- **Taxi ID:** A unique identifier for each taxi.
- **Timestamp:** The time of the GPS point.
- **Latitude:** The latitude of the GPS point.
- **Longitude:** The longitude of the GPS point.

2. beijing_sevenpart

- **Real-time traffic conditions on each road segment:** Real-time traffic conditions on each road segment in different time slots of a day.
- **Real-time traffic conditions in each region:** Real-time traffic conditions in each region in different time slots of a day.
- **Road network features and POI features:** Road network features of each road segment and POIs around each road segment.
- **Road network:** Road network connections of Beijing.

- **Traffic volume ground truth:** Traffic volume ground truth on different levels of road segments at different time slots.

3. PeMS04

- **PeMS04.rel:** Contains relational data describing the connections or distances between sensors.
- **PeMS04.geo:** Provides the geographical locations (latitude and longitude) of each sensor.
- **PeMS04.dyna:** Contains the dynamic traffic data (flow, speed, occupancy) recorded by the sensors every 5 minutes.

4. PeMS08

- **PeMS04.rel:** Contains relational data describing the connections or distances between sensors.
- **PeMS04.geo:** Provides the geographical locations (latitude and longitude) of each sensor.
- **PeMS04.dyna:** Contains the dynamic traffic data (flow, speed, occupancy) recorded by the sensors every 5 minutes.