Sea Surface Temperature Prediction using LSTM

Problem Statement

Develop a Machine Learning model that can take sequential data and generate the Sea Surface Temperature. The model will be trained on labelled data and used to predict unlabeled data.

Dataset

The dataset includes daily measurements of various atmospheric and oceanographic variables from multiple locations. The key variables include:

- Year, Month, Day
- Latitude, Longitude
- Zonal Winds, Meridional Winds
- Humidity, Air Temperature
- Sea Surface Temperature (Target Variable)

Approach

There are **2 jupyter notebooks** containing the project code :

1. train.ipynb (for training)
2. predict.ipynb (for prediction)

1. Data Preprocessing

- **Load and clean the data**: Fill missing values using forward fill.
- **Feature Selection**: Select relevant features excluding unnecessary lag features.

2. Model Training

- **Model Selection**: LSTM (Long Short-Term Memory) network is chosen due to its effectiveness in handling sequential data.
- **Data Preparation**: Reshape the data to be suitable for LSTM input.
- **Model Definition**: Define and compile an LSTM model with one LSTM layer followed by a Dense layer.
- **Training**: Train the model using the training dataset with a validation split to monitor performance.

3. Model Evaluation

- **Validation**: Evaluate the model on the validation dataset using Mean Absolute Percentage Error (MAPE) as the metric.

4. Prediction

- **Load new data**: Preprocess new data similarly to the training data.
- **Predict**: Use the trained model to predict SST for the new data.
- **Save Predictions**: Save the predictions to a CSV file.

Results

The trained model was used to predict SST on new data from the given two csv files. The model's performance was evaluated on the validation set using MAPE, providing a measure of prediction accuracy.

Key Metrics

- **MAPE on Validation Set**: The model achieved a MAPE of 91.34%.

Predictions

The predictions for the new data have been saved in the file as:

1. data_1997_1998.csv —> `data_1997_1998_predicted_sst.csv`.
2. evaluation.csv —> `evaluation_predicted_sst.csv`.

Thought Process

Data Handling

- **Handling Missing Data**: Used forward fill to handle missing values to maintain the temporal sequence.
- **Feature Engineering**: Initially considered lag features but decided against them for simplicity and consistency.

Model Development

- **Model Choice**: Chose LSTM due to its capability to handle time series data effectively.
- **Hyperparameters**: Used a single LSTM layer with 50 units and a Dense output layer. Opted for the 'adam' optimizer and 'mse' loss function.

Challenges

- **Data Consistency**: Ensuring the input features during prediction matched those used during training.
- **Model Performance**: Achieving a balance between training accuracy and generalization on validation data.

Conclusion

I enjoyed doing this project. This project successfully developed an LSTM-based model to predict Sea Surface Temperature, crucial for understanding and forecasting ENSO events. The approach highlighted the importance of proper data preprocessing, model selection, and evaluation techniques. Future work can focus on enhancing model accuracy and robustness through advanced feature engineering and model tuning. Through this project I learned many new concepts of ML and data handling.

I would like to show my gratitude to all those involved in creating this project for us, looking forward to work on interesting projects in AI, ML, DL and NN with both the clubs .

Thanking You

Yours humbly

Utkarsh Raj