Data-sets Overview

Munich RE

Dataset	Description	Task	Size/Properties	Pre-processing
Traffic Dataset	UCI PEM-SF Traffic Dataset which describes occupancy rate of 440 San Francisco Bay Area Freeways.	Use past week‘s data to forecast over the next 24 hours.	Target Type: [0,1] No. Of columns: 440 No. Of samples: 500k	Aggregate on hourly level.
Electricity Dataset	UCI Electricity Load Diagrams Dataset, containing electricity consumption of 370 customers.	Use past week‘s data to forecast over the next 24 hours.	Target Type: ℝ No. Of columns: 370 No. Of samples: 500k	Aggregate on hourly level.
Retail Dataset	Favorita Grocery Sales Dataset from the Kaggle competition is a metadata for different products and the stores, along with other exogenous time-varying inputs sampled at the daily level.	Use 90 days of past data to forecast product sales of 30 days into the future.	Target Type: ℝ No. Of columns: 130k No. Of samples: 500k	
RBB Data-streams	Time-related sensor data from production machinery.	Use past year‘s data to forecast over the next week.	Target Type: ℝ No. Of columns: 165 No. Of samples: 17k	Aggregate on daily level.

Approach to be used.

- The Electricity and Traffic Datasets are univariate time series containing known inputs alongside the targets.
- The datasets selected are in the perspective of multi-horizon forecasting.
- The aim of my thesis is:

1. Forecast the data.
2. Train the model using the actual and forecasted with DTW(dynamic time warping) as a loss function.
3. DTW will be used as a surrogate loss function implemented using a neural network(GANs).

- Evaluate the model with baselines on the basis of error and epoch-time.

Queries.

Munich RE

1. From the datasets overview, we got to know that Electricity and Traffic datasets have target variable while the other two datasets do not have a target variable. So when I have to decide which model shall I use for prediction and forecasting I think of following problems:

i. Model used for prediction should be same for fair comparison. But, two datasets have target variables(LSTMs or Autoencoders) and other two don‘t have targets(Variational Autoencoder) I‘m not sure which model shall I use?

ii. Alternatively, can I exclude the target and make use of Independent variables? Or make use of some of the features as target?

1. Could you please suggest few papers to grasp the idea of time-series forecasting? I know that there is a fine line between prediction and forecasting in the perspective of time series. But I couldn’t find any papers which explain how the network does forecasting and how do we evaluate those forecasts?

2. In the data-set overview table I’ve mentioned the task to be done which defines what type of multi-horizon forecast I wish to do. Does those tasks seems relevant or perhaps some changes are required.

References.

1. Electricity Dataset: online available at: https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014
2. Traffic Dataset: online available at: https://archive.ics.uci.edu/ml/datasets/PEMS-SF#
3. Favorita Dataset: online available at: https://www.kaggle.com/c/favorita-grocery-sales-forecasting https://www.kaggle.com/siliconx/favoritagrocerysalesforecastingextracted (version 1)