DeepAR investigation

DeepAR

Summary

DeepAR is an autoregressive model which means only time series data is enough, no other features or observation values are needed (of course, covariate features are also welcome). Then, I am uncertain whether it is still perfect in our case (it always needs some true value in the beginning to predict future values).

Improvements:

Data

1. Do not use the "aging feature" in this way. We do not even need aging features cos I think Ch4 emission is strongly dependent on our features but much less on timing features.

Model

1. Current implementation only has one lstm layer, I suggest using at least two layers as an encoder-decoder structure.
2. Previous embedding layer does not concatenate real covariate features, using them together with time feature to create a linear embedding feature. The paper also mentions "aging" feature (In all experiments we use an “age” feature, i.e., the distance to the first observation in that time series. We also add day-of-the-week and hour-of-the-day for hourly data, week-of-year for weekly data and month-of-year for monthly data).

Training

1. Current training and hyper parameter tuning needs to be done with some log-based libraries such as wandb, MLFlow and etc. The current training scheme does not converge.
2. Missing observations can also be utilized mentioned in the paper (In our model, missing observations can easily be handled in a principled way by replacing each unobserved value zi,t by a sample $\tilde{z}_{i,t} \sim l(\cdot|\theta(h_{i,t}))$ from the conditional predictive distribution when computing (1), and excluding the likelihood term corresponding to the missing observation from (2).)

3. Random sampling may be insufficient, but uncertain which way should be applied.

Loss

1. Actually, our task is a regression task. Instead of predicting real Ch4 emission values, why not predict the offset between real and observed Ch4 emission. That is also a way of normalization.

Libraries

1. Suggest using Pytorch Forecasting for a complete and versatile algorithms and models for time series prediction.