Time Series Forecasting on Sparse Data

Machine Learning Process

Prepare
- Load large collection of data for desired task
- Clean data set by filling in missing values

Preprocess
- Manipulate shape, center, and size of data to prepare data for training
- Split data into a training and testing set
- For TSFs, autoregression

Train
- Start with randomly guessing model, then iterate over training data to improve performance
- Model is ready for use after training

Postprocess
- Apply trained model to testing data to obtain prediction
- Undo manipulations from preprocess step

Assess
- Compare predicted and observed responses to assess model performance
- Tweak training options and preprocessing to create better models

What are Time Series Forecasters?

Time Series Forecasters (TSFs) are a type of machine learning algorithm designed to predict quantities that vary with time. They utilize autoregression to forecast the near future based on the most recent trends.

	Original	-1	-2	-3	-4	-5	-6
t=0	*a*	N/A	N/A	N/A	N/A	N/A	N/A
t=1	*b*	*a*	N/A	N/A	N/A	N/A	N/A
t=2	*c*	*b*	*a*	N/A	N/A	N/A	N/A
t=3	*d*	*c*	*b*	*a*	N/A	N/A	N/A
t=4	*e*	*d*	*c*	*b*	*a*	N/A	N/A
t=5	*f*	*e*	*d*	*c*	*b*	*a*	N/A
t=6	*g*	*f*	*e*	*d*	*c*	*b*	*a*

Original events *a, b, c, d, e, f, & g* occur at the listed time steps. Autoregression generates the rightmost columns by pushing the events down by the above value. The model reads the table horizontally, using the pattern of the previous events to predict the next event. The looked-back previous events are known as *lags*.

Problem

The problem with TSFs is that their training requires many data samples related to the desired function. For smaller applications, such quantities of data are scarce.

Solution

The solution is in the preprocessing step (figure below).

Data Folding (1) shortens and widens the set, extracting a repeated pattern.

Autoregression (2) applies several lags as shown above.

Response (5) copies the folded data to use as observed responses.

PCA (3) & (6) reduces the width of the data set to make the training process easier.

Splitting (4) & (7) breaks the data into a training and testing set.