Master Project

Automatic Machine Learning for Time Series Forecasting

Jeroen Swart, Delft Center for Systems and Control, TU Delft
J.Swart-1@student.tudelft.nl

Pieter Verhoeven, Myst AI, Inc., San Francisco
pjverhoeven@myst.ai

Peyman Mohajerin Esfahani, Delft Center for Systems and Control, TU Delft
P.MohajerinEsfahani@tudelft.nl

Context

Time series data occur in countless domains including financial analysis, social activity mining and medical analysis. Forecasting is one of the most sought-after applications of time series, enabling data-driven decision making and therefore automating and optimizing business processes. For example in energy retailing, the estimated future electricity usage guides the amount of energy that is bought. The shared mathematical setting is that there is access to the endogenous series, the predicted series, and often exogenous series, time series that are related to the predicted signal.

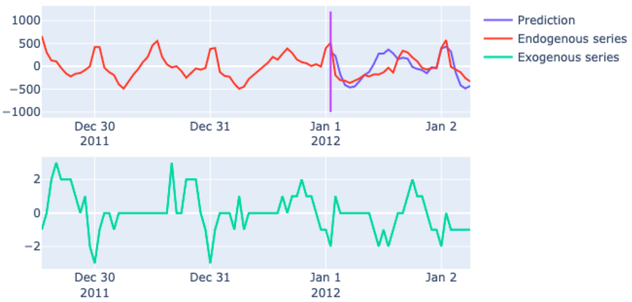

Figure 1: Endogenous and related exogenous time series with prediction

Classical time series forecasting models mostly capture autoregressive patterns in the endogenous frame, leaving information in exogenous data untouched. New regression models like deep learning and tree-based approaches are centered around learning to predict a target from exogenous data in time invariant systems and benefit from large datasets. Since more and more exogenous data becomes available, the combination of the classical and regression paradigm is highly promising.

Taking the regression paradigm to time series forecasting brings some challenges. Time variance needs to be tackled by transforming the data - often by stationarizing - and adding time based features before feeding it into the regression model. For sophisticated regression models a high amount of model hyperparameters need to be tuned to increase performance. It depends on the dataset which data transformations, model choice and hyperparameter configurations are optimal. Due to the extremely high amount of potential configurations we can't see the wood for the trees in all the modeling choices.

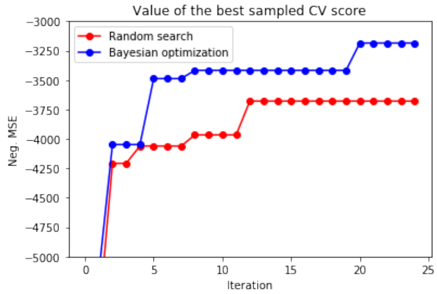

Figure 2: Bayesian optimization versus random search

Recently a lot of research has been done towards automating a subset of the modeling choices: hyperparameter optimization, an optimization with an expensive objective function and no gradient available. Bayesian optimization and genetic algorithms are the most common approaches to this problem. Theoretically also other modeling choices could be included in this optimization. The specific context of this thesis is that it is conducted in a professional environment in which numerous forecasting problems will have to be solved as fast, accurate and cost-efficient as possible.

Project tasks

This master thesis project is aimed at automating the expert effort of modeling choices for forecasting algorithms. At first an oversight of all modeling choices is explored and summarized in a theoretical hyperoptimization framework. Since it is computationally too expensive to optimize over all modeling choices, a relevant subset of choices is selected. A choice is made for optimizing over hyperparameter settings of XGBoost on an endogenous time series with one related exogenous time series, both stationarized, with added time-based features, for one-step-ahead predictions. Finally a learning algorithm is proposed, which learns to initialize Bayesian optimization by looking at similar older forecasting tasks. The computational costs and performance of this approach is compared with regular Bayesian optimization over the defined search space.

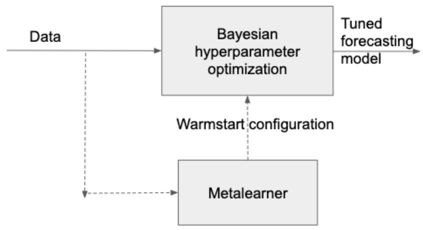

Figure 3: Warmstarted Bayesian hyperparameter optimization