Comparison of conventional time series forecasting (ARIMA) with LSTM Neural Network

ESC 403 | Introduction to Data Science

Project proposal | 01 April 2022

Contributed by

Nikolai Horozov	nikolai.horozov@uzh.ch
Jerome Sepin	jerome.sepin@uzh.ch
Vithersan Somasundaram	vithersan.somasundaram@uzh.ch
Gabriel Marie Falconer-Stout	gabrielmarie.falconer-stout@uzh.ch,

General topic

The aim of the project is to compare the prediction ability of Long Short-Term Memory (LSTM) model and conventional time series prediction (automatic ARIMA modeling) under the hood of climate data. We start simple: The LSTM model is trained on a simple simulated data set (sinus function) with a structure that is for us humans very intuitive. By doing this task we want to get a feeling how capable the LSTM model is to capture periodically repeating scenarios as it is also the case for real-life climate variables such as for example temperature. As a next step, we want to learn the LSTM model on climate data and do prediction for future values (after a certain period). These predictions are simultaneously done with the ARIMA model and deviations from the true values are compared between the models.

Motivation

Climate has unquestionable an impact in our everyday life. What also finds increasing importance in many areas, is the use of machine learning tools. Naturally comes the question whether these machine learning tools are also applicable to understand the system of climate. Climate data is dependent data and therefore so called recurrent neural networks (RNN) have been developed to incorporate the past observations for future inference. Machine learning methods are sometimes seen as so called black boxes that may have the advantage of getting reasonable results without knowing the system in detail. By feeding enough data and relying on computational power, it has been shown to be very useful to find suitable pattern in also very complex systems. In our project we want the traditional ARIMA times series modeling approach to be compared with Long-Short-Term-Memory (LSTM) neural networks for prediction of future climate observations.

Data

The data comes from the archive of the Center for Machine Learning and Intelligent Systems (UCI) and reports on the weather and level of pollution at the US embassy in Beijing, China. It consists of hourly observations for five years and thus has 43’800 observations.

- Date & Time
- Temperature
- Dew point
- Air pressure
- Cumulated Wind speed
- Wind direction
- Air pollution
- Cumulated hours of snow
- Cumulated hours of rain

Source:
Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H. and Chen, S. X. (2015). Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. Proceedings of the Royal Society A, 471, 20150257.

Data processing

- **Data Simulation**

- **Data Cleaning**

- **Exploratory Data Analysis (EDA)**
  - Sorting, summarization
  - Visualization
  - Plotting

- **ARIMA automatic modeling**

- **Conducting neural network**
  - LSTM
  - Train (70%) and Test (30%)
  - Validation set – To compare our forecast
  - **Creating features and labels:**
    - specific time length: optimal time lag needs to be determined

Research Question

Is there a difference between ARIMA and LSTM for prediction future climate values?

Analysis Techniques / algorithms

Using Time Series Deep Learning to forecast climate data with Keras Stateful LSTM and the forecast library for automatic ARIMA prediction in R to answer the primary research question.