Machine Learning Engineer Nanodegree

Capstone Proposal

Oscar Daniel Garcia Lino
April 27th, 2019

Wind Power Forecasting using LSTM RNN models

Domain Background

As a renewable energy source, wind power generation will play a bigger role in the future [1]. However, one of the main problems of wind energy is its dependency on environmental conditions. In the electricity grid, at any moment, balance must be maintained between electricity consumption and generation. The nature intermittency of wind adds complexity to the management of wind energy but this intermittency can be partly mitigated through forecasting techniques, which aim at reducing the uncertainty of future wind power generation of a wind farm or portfolio [2].

Wind power forecasting models can be classified into two categories depending on their methodology: Physical approach and Statistical approach. The physical approach is based on lower atmosphere or numerical weather predictions, in the other hand, the statistical approach is based on vast amount of historical data without considering meteorological conditions. There is a third approach, the hybrid approach, which combines physical methods with statistical methods, particularly uses weather predictions and time series analysis. [3]

Problem Statement

For this capstone project I'm going to use a hybrid approach to forecast wind power production for a short term (8 hours ahead) using time series data. My intention is to use deep learning to solve the forecasting problem, specifically a Long Shor-Term Memory (LSTM) neural network. If time permits, I would like to compare a neural network forecasting against a XGBoost forecasting model.

Datasets and Inputs

The dataset to be used was obtained from the website for the Sotavento Experimental Wind Farm project. The Sotavento project was promoted by the Government of Galicia for the promotion, training and research of renewable energy sources [4]. The website offers real time data visualization of the wind farm SCADA data. Historical time series data is also available and accessible for public domain.

The available historical data goes from July 2004 until now (keeps updating daily). There are only three variables available in the historical data these are the wind speed (m/s), wind direction (º) which are weather related features, and power energy produced (kWh) which is going to be the output feature and what I intend to forecast. The hourly wind speed and wind direction is the average value during the hour, and the hourly energy value is power produced during the hour. I am also planning to use date features as inputs to my model (ex. Month, week of the year) this may help to catch seasonality during the year. The table below shows a snapshot of these variables:

timestamp	wind_speed	wind_direction	energy	week
2004-01-07 01:00:00	2.11	7.0	0.00000	2
2004-01-07 02:00:00	3.27	348.0	0.06231	2
2004-01-07 03:00:00	4.58	340.0	0.41000	2
2004-01-07 04:00:00	4.08	343.0	0.02765	2
2004-01-07 05:00:00	3.93	343.0	0.21769	2

The historical data can be obtained in 10 min, hourly or daily time step. For the purpose of this project, I am going to use the hourly historical data for the three available variables. The dataset was manually copied into a comma-separated values file. There are a couple of months of data gaps between 2004 and 2010 which may cause some problems for the neural network model, therefore I will potentially only use the data from 2010 until now to build the forecasting model.

Solution Statement

There are several time-series forecasting techniques like auto regression (AR) models, moving average (MA) models, Holt-winters, ARIMA, to name a few [5]. These techniques are very popular and traditional for time series forecasting but they have a bottleneck. This is because is not easy to incorporate new signal in the forecasting model like events or weather data for example [6]. More powerful methods need to be used to be able to incorporate multiple signals in forecasting models. This is where deep learning comes into play. Long short-term memory (LSTM) recurrent neural network (RNN) is becoming popular for time series forecasting. Unlike other machine learning algorithms, LSTM are capable of automatically learning features from sequence data, support multi-variate data, and can output a variable length sequence that can be used for multi-step forecasting [7].

The dataset that I am going to use for this project represents a multivariate time series of power-related variables that can be used to model and forecast power production for the Sotavento wind farm. For purposes of this project, I am planning to develop a LSTM model to forecast hourly power production eight hours ahead. Wind farm hourly time series data from 2010 up to 2018 is going to be used for training the model. The test data set is going to contain data from 2019.

Benchmark Model

I am going to use a persistence model as benchmark model to stablish a baseline to compare my forecasting model. A persistence model is the simplest way of producing a forecast [8]. The persistence model uses the value at the previous time step (t-1) to predict the expected outcome at the next time step (t+1). In other words, it assumes that for example the energy produced by the wind farm for the next hour is going to be same as the energy produced during the past hour. The persistence model is going to be treated as univariate time series forecasting problem and only the wind farm energy generated is going to be considered for the forecast.

Evaluation Metrics

There are many different performance measures to choose from when it comes to time series predictions. For this project I am proposing to use scale-dependent errors as evaluation metrics. Scale-dependent errors are accuracy measures that are based only on e_t and cannot be used to make comparisons between series that involve different units [9]. The forecast that I am planning to execute in this project involves only one output variable, therefore scale-dependent errors seems suitable in this case.

The two most commonly used scale-dependent measures are based on the absolute errors or squared errors [9]:

$$Mean\ absolute\ error: MAE = mean(|e_t|)$$

$$Root\ mean\ squared\ error: RMSE = \sqrt{mean(e_t^2)}$$

Therefore, I am planning to use both MAE and/or RMSE metrics and compared the results between them.

Project Design

The Sotavento Wind Farm dataset has four columns: timestamp, wind speed, wind direction and energy produced. Week of the year may be helpful for the forecasting that I am trying to accomplish in this project, therefore a new column will be added to the dataset with the week of the year. With the week of the year we may be able to catch any variation in power generation due to weather seasons in the area where the wind farm is located.

The timestamp columns will have to be analyzed as well to identify if daylight saving time (DST) changes is presented in the data. If DST is presented then it may be a good idea to change the timestamps to UTC time instead as this may cause some issues as well when training the LSTML models.

Time series SCADA can present different anomalies over time due to sensor failures or equipment maintenance. Failures can result in data gaps and this gaps in the data can potentially impact training of the machine learning models. Reason why it is important to identify any gaps in the data before applying machine learning on it. If data gaps are present then interpolation may be used to fill those gaps.

Time series decomposition could be helpful to understand more the time series data. In decomposition, the time series data is split into several components: trend, seasonality and remainder [9]. A library in python that can be used for decomposition is the statsmodel library, which provides an implementation of the classical decomposition method in a function called seasonal_decompose() [7].

A recurrent neural network will be implemented to resolve the multi-step forecasting problem proposed in this project. RNNs are specifically designed to work, learn, and predict sequence data like time series data. One of the most successful and widely used networks is the long short-term memory (LSTM) network. Further, specialized architectures have been developed that are specifically designed to make multi step

sequence predictions, generally referred to as sequence-to-sequence prediction or seq2seq for short. An example of a RNN architecture designed for seq2seq problems is the encoder-decoder LSTM [7]. An encoder-decoder LSTM will be implemented as proposed architecture to resolve the forecasting problem.

Lastly, root mean squared error (RMSE) will be used as evaluation metric.

Jason Brownlee has a good tutorial about developing LSTM models for multi-step time series forecasting [7]. His tutorial will be taken as a reference to build the LSTM architecture suggested to resolve the wind power forecasting problem proposed in this project.

References

[1] AESO, "aeso," [Online].

[2] A. C.-T. O. L.-G. Cristobal Gallego-Castillo, "CORE - A review on the recent history of wind power ramp forecasting," [Online]. Available: https://core.ac.uk/download/pdf/78495744.pdf. [Accessed 23 April 2019].

[3] P. G. X. H. Xiaochen Wang, "ScienceDirect - A Review of Wind Power Forecasting Models," [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1876610211019291. [Accessed 22 April 2019].

[4] S. G. Foundation, "sotaventogalicia," Sotavento Galicia, S.A., 2005. [Online]. Available: http://www.sotaventogalicia.com/en. [Accessed 19 April 2019].

[5] "Towards Data Science," Towards Data Science Inc, 17 January 2018. [Online]. Available: https://towardsdatascience.com/using-lstms-to-forecast-time-series-4ab688386b1f. [Accessed 27 April 2019].

[6] D. Yuan, Director, *Two Effective Algorithms for Time Series Forecasting.* [Film]. InfoQ, 2018.

[7] P. Jason Brownlee, "How to Develop LSTM Models for Multi-Step Time Series Forecasting of Household Power Consumption," Machine Learning Mastery, 10 October 2018. [Online]. Available: https://machinelearningmastery.com/how-to-develop-lstm-models-for-multi-step-time-series-forecasting-of-household-power-consumption/. [Accessed 27 April 2019].

[8] "Persistence Method," University of Illinois, [Online]. Available: http://ww2010.atmos.uiuc.edu/(Gh)/guides/mtr/fcst/mth/prst.rxml. [Accessed 1 May 2019].

[9] R. J. H. a. G. Athanasopoulos, "Forecasting: Principles and Practice," OTEXTS, [Online]. Available: https://otexts.com/fpp2/accuracy.html. [Accessed 27 April 2019].

[10] "Meteologica," Meteologica S.A., [Online]. Available: http://meteologica.com/.