Modeling the number of sunspots using machine learning

Nikolaos Paraskakis, Dionissios Hristopulos
Technical University of Crete, Chania, Crete, Greece

Solar activity has a significant impact on various aspects of human life, including satellite communications, power distribution systems, and climate change. Sunspots are the fundamental indicators of solar activity. Accurate prediction of sunspot activity can help mitigate potential hazards and improve the understanding of solar dynamics. In this study, we present a comparative analysis of three supervised machine learning models for predicting sunspots: Gaussian Process Regression, Long Short-Term Memory neural network, and LightGBM. We used a time series of the yearly mean total number of sunspots to train and evaluate the performance of each model. The sunspot data exhibit a long-term periodicity (solar cycle) of 11 years. The dataset was split into training, validation, and testing sets, and various well-known performance metrics were employed to assess the prediction accuracy. Gaussian Process Regression (GPR) is a non-parametric, probabilistic approach that is particularly useful for data with complex patterns. A Gaussian Process (GP), $f(x)$, is completely specified by its mean function $m(x)$ and covariance (kernel) function $k(x,x')$. We experimented with two different kernels: (i) the product of an exponential and a periodic kernel, and (ii) a linear stochastic oscillator kernel. We also tested warped GPR by applying the κ-logarithmic transformation. We optimized the model hyperparameters using Maximum Likelihood Estimation (MLE) and RMSE minimization on a validation set. A Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) employs a deep learning architecture to learn patterns. It consists of cells, each of which is connected to three gates (input, forget, and output) responsible for information flow. We implemented an LSTM model with multiple layers and optimized its architecture and hyperparameters using grid search and validation loss minimization. We also employed LightGBM, a gradient-boosting framework of regression trees, that is well-known for its efficiency and accuracy in regression tasks. Our results show that all three models exhibit strengths and weaknesses. GPR delivers uncertainty estimates and can capture complex patterns using different kernels, but it requires the computationally intensive inversion of large covariance matrices. LSTM performs well in capturing long-term dependencies, but it needs large amounts of data, time, and resources for tuning and training, and it suffers from error accumulation on long-term predictions. LightGBM delivers the same characteristics, except that it is more computationally efficient and its training is faster. In conclusion, this study provides insights into the performance and characteristics of three powerful machine learning methods, GPR, LSTM-RNN, and LightGBM, for sunspot number prediction.

References

[1] D.T. Hristopulos, Random Fields for Spatial Data Modeling". Springer, Netherlands (2020).
[2] C.E. Rasmussen, Williams C. K. I., "Gaussian processes for machine learning". MIT Press (2005).
[3] E. Snelson, Z. Ghahramani, C. Rasmussen, Advances in Neural Information Processing Systems 16 (2003).
[4] D. T. Hristopulos, A. Baxevan, Entropy, 24(10), 1362 (2022).
[5] V. D. Agou, A. Pavlides, D. T. Hristopulos, Entropy,24(3), 321 (2022).
[6] I. G. Gonçalves, E. Echer, E. Frigo, Advances in Space Research, 65(1), 677–683 (2020).