Capstone Project 2 – Deep Learning for Weather Forecast

Contents

Introduction

A time series is a sequence of numeric data observations collected over a period of time at regular intervals. Sequential or temporal data observations emerge in many key real-world problems, ranging from biological data, financial markets, weather forecasting, to audio and video processing. Temporal dependency in time series cause two otherwise identical points of time to belong to different classes or predict different behavior. This characteristics generally increase the difficulty of analyzing them.

Weather forecast is among the most popular forecast problems. Weather forecast includes forecasting temperature, pressure, humidity, wind direction and wind speed. Unlike other time series datasets, weather data has unique features. There is season-to-season, year-to-year variability in the trends of weather data. The temperatures and Pressures are correlated. Wind has the direction and the magnitude so should represented by a vector. So all these features could be separately predicted using univariate time series analysis techniques or could be used jointly to predict using multivariate time series techniques.

There has been an extensive research done in machine learning for time series forecasting techniques. Several machine learning algorithms can be used to solve time series forecast problems, such as K- Nearest Neighbors, Support Vector Regression, Multi-Layer Perceptron, and Auto Regressive Integrated Moving Average (ARIMA).

With the advent of Deep Learning algorithms, various supervised and unsupervised models are developed to analyze massive size time series data. Deep Learning is the general term for a series of multi-layer architecture neural networks. Several different Deep Learning approaches are there for time series analysis, such as The Long Short Term Memory (LSTM) neural network, Deep Belief Networks (DBNs), stacked Auto Encoders, Multi-layer Perceptron Regression. The Long Short Term Memory (LSTM) neural network is very popular architecture for time series problems. LSTM neurons keep a context of memory within their pipeline to allow for tackling sequential and temporal problems without the issue of the vanishing gradient affecting their performance.

Forecasting future values of a time series plays an important role in nearly all fields of science and engineering, such as economics, finance, business intelligence and industrial applications. Also in real world applications such as speech recognition, real time sign language translation, finance markets, weather forecast etc. Deep Learning algorithms are known to perform best when there is a massive dataset available for learning. Not every time series problem has massive dataset available. In that case advanced ML algorithms are available for time series applications. The recommendation made in this project can be useful for anyone looking for best ML/DL algorithms for medium size time series problem dataset.

Historical Hourly Weather Dataset

The dataset contains ~5 years of high temporal resolution i.e. hourly measurements of six weather attributes: temperature, humidity, air pressure, wind direction, wind speed and general weather description of 30 US and Canadian cities and 6 Israeli cities. The dataset contains separate file for each weather attribute. Each file contains 36 cities as columns. Rows are time axis and time axis are same in all files for different weather attributes.

Problem Statement

This dataset is not large enough for deep learning algorithms to perform well. So we use various Machine Learning algorithms and Deep Learning algorithms, compare their performance and make a recommendation for best models for time series analysis for medium size data. To increase the performance, a hybrid ensemble model of Machine Learning and Deep Learning can also be used.

Libraries Used

Python, MatplotLib, Seaborn, Numpy, TensorFlow and Keras

The high level steps to solve the problem

1. The dataset contains separate file for each weather attribute. Each file contains 36 cities as columns and time axis as rows. Join these files and reshape to prepare it for time series analysis.
2. This dataset is a real world dataset which means there are missing and incorrect values. Use data wrangling techniques to clean dataset.
3. Perform EDA.
4. Do Normalization and Scaling on data if needed.
5. Split the dataset into training and test datasets.
6. Create Machine Learning models and calculate the accuracy of the predicted weather data.
7. Create Deep Learning models and calculate the accuracy of predicted weather data.
8. Execute the models on GPUs to get the final results.
9. Compare the final results of ML and DL algorithms and recommend models which work best on medium size time series data problems.

Deliverables

1. Project Report
2. Milestone reports
3. Jupyter Notebook for the code
4. Slide Deck

References

1. Link to data set: https://www.kaggle.com/selfishgene/historical-hourly-weather-data
2. Liu, J.N., Hu, Y., You, J.J., Chan, P.W.: Deep neural network based feature representation for weather forecasting. http://worldcomp-proceedings.com/proc/p2014/ICA3405.pdf
3. Aditya Grover∗ IIT Delhi, Ashish Kapoor Microsoft Research, Eric Horvitz Microsoft Research: A Deep Hybrid Model for Weather Forecasting: http://erichorvitz.com/weather_hybrid_representation.pdf
4. John Gamboa: Deep Learning for Time-Series Analysis. https://arxiv.org/abs/1701.01887
5. Dmitry Vengertsev: Deep Learning Architecture for Univariate Time Series Forecasting: http://cs229.stanford.edu/proj2014/Dmitry%20Vengertsev,Deep%20Leraning%20Architecture%20for%20Univariate%20Time%20Series%20Forecasting.pdf
6. How do I Predict Time Series by Farhad Malik : https://medium.com/fintechexplained/forecasting-time-series-explained-5cc773b232b6
7. Acf() and acpf() - https://machinelearningmastery.com/gentle-introduction-autocorrelation-partial-autocorrelation/

8. https://www.datacamp.com/community/tutorials/time-series-analysis-tutorial
9. https://machinelearningmastery.com/time-series-data-visualization-with-python/
10. https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/03.11-Working-with-Time-Series.ipynb#scrollTo=4luYjFSf-Azu
11. https://www.statisticshowto.datasciencecentral.com/lag-plot/
12. Create a weather forecast map: https://developer.ibm.com/clouddataservices/2016/10/06/your-own-weather-forecast-in-a-python-notebook/
13. https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/
14. https://www.analyticsvidhya.com/blog/2016/02/time-series-forecasting-codes-python/
15. Acf and pacf intuition - https://towardsdatascience.com/significance-of-acf-and-pacf-plots-in-time-series-analysis-2fa11a5d10a8
16. Best ARIMA explainantion - https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/
17. https://machinelearningmastery.com/time-series-forecasting-methods-in-python-cheat-sheet/
18. Multivariate model –
    a. SARIMA - http://barnesanalytics.com/sarima-models-using-statsmodels-in-python
    b. GARCH - http://barnesanalytics.com/garch-models-in-python
    c. ARIMAX - http://barnesanalytics.com/analyzing-multivariate-time-series-using-arimax-in-python-with-statsmodels
    d. https://github.com/statsmodels/statsmodels/tree/master/statsmodels/tsa
    e. ARIMA – univariate - https://towardsdatascience.com/time-series-in-python-exponential-smoothing-and-arima-processes-2c67f2a52788
    f.