Machine Learning Engineer Nanodegree

Capstone Proposal

Matthew Gilson
May 3rd, 2017

Proposal

Domain Background

Space Weather prediction is becoming an increasingly interesting subject of study due to the large numbers of complex systems that are in use on a daily basis that are sensitive to the "space weather" environment. At it's core, the space weather environment is driven by the sun and high-speed plasma that is constantly being ejected from the sun. The vernacular term for this plasma is the "solar wind". In scientific communities the solar wind and it's contained magnetic field is called the "Interplanetary Magnetic Field" (IMF). The IMF interacts with the earth's magnetic field (i.e. the "magnetosphere") and can cause observable and sometimes detrimental effects on earth and in near earth space. Particles in near earth space can be energized during periods of prolonged southward oriented IMF (i.e. when the solar wind plasma's magnetic field is pointing southward) or when the IMF pressure on the magnetosphere suddenly increases. These excited particles tend to drift in the same direction and the drift direction is dependent on their charge (protons drift westward, electrons drift eastward). These drifting particles create the ring current which an electrical current that encircles the earth. Enhanced ring current is the metric which is used define "geomagnetic storms". This is important for a number of reasons:

1. Satellite electronics are sensitive to high energy electrons that can be produced during strong geomagnetic storms.
2. The ring current can create magnetic field perturbations on the earth. Those perturbations can (among other things) destroy transformers that the power grid relies on.

Understanding and predicting geomagnetic storms and their severity would be very useful to people in these and other industries.

There are a few indexes used to measure the severity of a magnetic storm. The most common is the Disturbed Storm Time (DST) index which is calculated from ground based perturbations at low latitudes. The values are computed using hour long averages. This index is not calculated in real-time because it also relies on subtracting off a "quiet" baseline that is determined very

[1] Lei et al, 2009 -- Prediction of SYM-H index by NARX neural network from IMF and solar wind data
[2] O'Brien and McPherron, 1999 -- Forecasting the ring current index Dst in real time

infrequently. A provisional version is available for use where the official DST has not been computed yet. Another version of the DST that is calculated in near real-time and with a much higher time resolution is the Sym-H index. In the scientific literature, the Sym-H index and the DST index are often used interchangeably. They measure the same thing from the same data sources and are reported in the same units. The Sym-H index will be the subject of this report.

For those interested, a list of geomagnetic storms can be found here.

Problem Statement

Since the earth's magnetospheric response is driven primarily by solar wind input, given the solar wind input, the response should be predictable. However, the magnetosphere is a complex system of highly coupled regions that have different dynamics. Up to this point, physics-based simulations and simple empirical models have had a difficult time simulating the response of the SymH (and DST) index based on the solar wind input drivers. The empirical models can do reasonably well during quiet times when the solar wind input drivers are relatively stationary, but when things get active their prediction capabilities drop. One possibility for why they do not do a great job is because they do not take into account the solar wind driver's history which can precondition the magnetosphere. Physics based simulations have a different problem. They suffer from the inability to adequately resolve all of the regions of importance for this particular calculation. My hope is that a new empirical model can be developed that takes the history of the solar wind into account. Fundamentally, this is a multivariate timeseries prediction problem. Ideally, we can simulate both disturbed times as well as quiet times using the same model as that would ease the transition of any such model into operations.

Datasets and Inputs

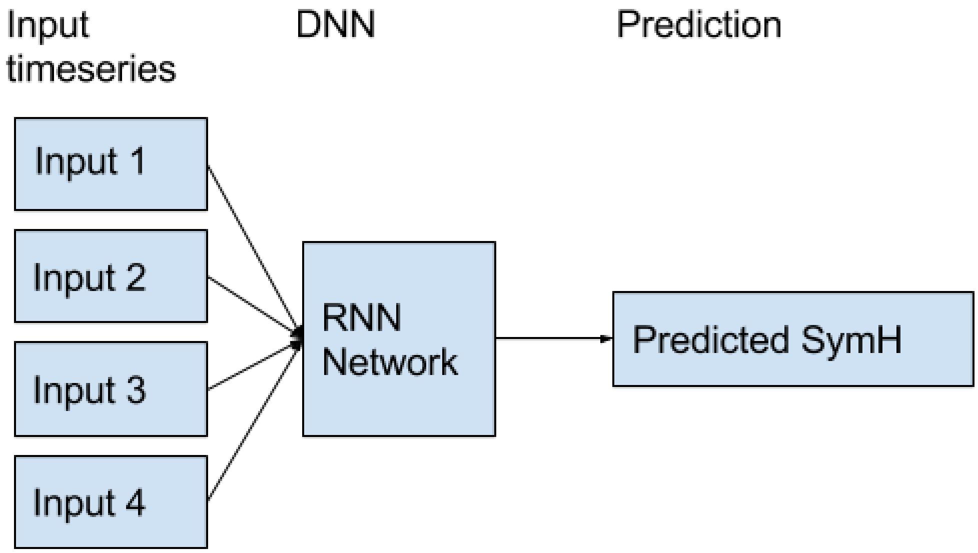
The OMNI dataset is publicly available from NASA and can be downloaded at ftp://spdf.gsfc.nasa.gov/pub/data/omni/high_res_omni/. It contains a multi-variate timeseries with all of the key parameters that we expect to be important for this prediction. Specifically, the dynamic pressure and magnetic field components are reported at any time where there was a functioning solar wind monitoring satellite in the solar wind. After about August 1995, the data coverage is pretty good due to solar wind monitors that could orbit around the Lagrange 1 point. If we use the 5 minute resolution data from August 1, 1995 to January 1, 2017, there are approximately 2.25 million datapoints per timeseries. There are still some gaps of missing data that will need to be dealt with in our solution somehow, but they are a very small percentage of the dataset (~5%). The OMNI dataset also includes the SymH timeseries which is exactly what we are trying to predict.

[1] Lei et al, 2009 -- Prediction of SYM-H index by NARX neural network from IMF and solar wind data

[2] O'Brien and McPherron, 1999 -- Forecasting the ring current index Dst in real time

Solution Statement

There are a number of solutions that we could employ here. ARIMA models have been used in timeseries predictions before so we could try one of those. However, I want to gain experience with Deep Neural Networks (DNN), so I plan on using a Recurrent Neural Network (RNN) in order to effectively track the history dependence. Ideally, a solution like the following should be workable:



Evaluation Metric

To measure our predictions during testing and validation, a simple Mean Square Error (MSE) metric should work well. It is effectively the average euclidean distance between a prediction and the observed value. If we successfully minimize it, then our results will be as close to the actual results as possible which is exactly what we want to achieve.

Benchmark Model

Previous studies[1,2] have used Root Mean Square Error and some have reported simple differences as a percentage of the full dataset. For example, O'Brien and McPherron [1999] shows a histogram of percentage of predictions that are within 5nT, 10nT, ... of the expected value We can benchmark our model against both of these results.

[1] Lei et al, 2009 -- Prediction of SYM-H index by NARX neural network from IMF and solar wind data

[2] O'Brien and McPherron, 1999 -- Forecasting the ring current index Dst in real time

Project Design

To pull this off, we will first work on getting the data in a more reasonable format than is available from the OMNI FTP server. Sqlite3 databases are really handy for this. After that, we'll need to do some preprocessing. Some preprocessing steps that may be necessary are:

- Handling missing values (removal and/or filling)
- Upsampling or downsampling the timeseries values
- Artificially replicating interesting events to "balance" the dataset
- Creating input vectors by using a sliding window approach
- Splitting the dataset into train, validate and test datasets

After that, I'll hypertune the network parameters (including the architecture of the network itself) using the training and validation datasets. Finally, we'll see how well we perform on the testing dataset.

[1] Lei et al, 2009 -- Prediction of SYM-H index by NARX neural network from IMF and solar wind data
[2] O'Brien and McPherron, 1999 -- Forecasting the ring current index Dst in real time