Sean Toon

IT-460-18500-M01 Machine Learning 2024 C-5

9/19/2024

Milestone One – Draft of Model and Dataset Identification and Defense

The case and data that I am choosing to analyze is the Sunspot data to help predict future sunspots. The CSV file containing sunspot data offers records of observed sunspot data. The data seems to represent the observations with the variables; year, month, date in decimal form, number of sunspots observed, and a potential mark to represent a significant day of sunspot occurrence. The goal of the model I will create is to help recognize any patterns and correlations between earth dates, sunspots, and solar cycles.

The model that I plan to use to represent the data and to help predict future sunspots is multiple linear regression. Multiple linear regression seems to be the most suitable algorithmic method to predict future sunspots that was provided in our textbook, so far. There may be some disadvantages such as assuming linearity relationships, sensitivity to outliers and while it may capture the events of solar spots in correlation to solar cycles, it may not capture how the spots correlate with each other. Overall, if the model is trained correctly and enhanced carefully, I might be able to do a good job. I suspect that as I continue reading about more models in the textbook, there may be a future way to enhance the model my MLR model or make me change the model completely. Another way that I may have to enhance the model is by adding more variables.

A model that can represent past sunspot data and help predict future sunspots can provide data to astrophysicists and others who are concerned with solar flares, or any other solar activity. The data can be used to analyze how solar activity and space weather affects technology, infrastructure and the health of humans on Earth. For example, the last solar storm that happened in May of 2024, was so great that it caused auroras, disturbed drones, suspended agricultural activities and more. (SIDAC, 2024)

The tool that I will be using to create my model is R. R is a great tool to use for manipulating, modeling, and visualizing data as it offers many different libraries and methods to do so. The step-by-step process that I will use to create my model looks like:

1. Collect Data
2. Explore and prepare the data
3. Search for any correlations or relationships using graphs, plots and matrices.
4. Train model on the data
5. Evaluate model performance
6. Improve model performance

References

Solar Influences Data Analysis Center.*Extremely severe geomagnetic storm!*[https://www.sidc.be.](https://www.sidc.be./) Retrieved Sep 19, 2024, from <https://www.sidc.be/article/extremely-severe-geomagnetic-storm>