Resumen

Las manchas solares han sido un tema que siempre ha llamado la atención, desde los astrónomos chinos hasta culturas de Latinoamérica como los Mayas, los cuales podían observar el sol en determinadas fechas del año sin afectar el ojo humano considerablemente. Hoy en día existe la tecnología suficiente para poder hacer estas observaciones y analizar con mayor profundidad las afectaciones que dichas manchas ocasionan al sol y en consecuencia a la Tierra. Creando un modelo de Machine Learning llamado Regresión Lineal, se plantea un algoritmo que pueda predecir la aparición de estas manchas solares usando un Dataset proporcionado a la comunidad científica por instituciones astronómicas y poder realizar pruebas y diseños propios.

Introducción

**¿Qué son las manchas solares?**

Una mancha solar es una región del Sol que tiene una temperatura más baja que sus alrededores, y con una intensa actividad magnética. Una sola mancha puede llegar a medir hasta 12,000 km (casi tan grande como el diámetro de la Tierra), pero un grupo de manchas puede alcanzar 120,000 km de extensión e incluso algunas veces más.



**Estado del arte**

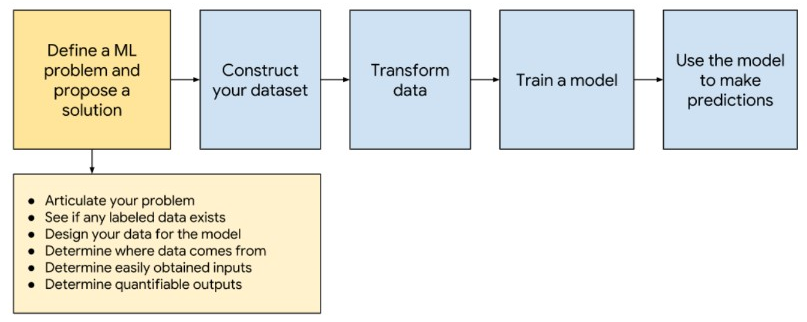
Se tiene documentado que las primeras observaciones fueron hechas por astrónomos chinos en el 2800 A.C., derivado de que las manchas solares pueden distinguirse a simple vista, con ciertos factores a considerar (la hora, el día del año, el clima, etc.). En Latinoamérica se ha investigado que los Mayas también pudieron ver estas manchas solares anteriormente en la ciudad de Mayapan, considerada la ciudad astronómica y militar más importante de la Península de Yucatán durante el período de 1519-1000 A.C.



Los Números Internacionales de Manchas Solares son cantidades que miden el número de manchas solares en el Sol (Wolf o Zürich que es R = k (10g + s), donde R es el número individual de manchas, g es el número de grupos de manchas solares y k es el factor que varía con la ubicación y la instrumentación, también conocido como factor de observación). En épocas recientes, han servido durante mucho tiempo como la principal serie temporal que define la actividad solar desde 1700. En 1998, los científicos Hoyt y Schatten realizaron una nueva reconstrucción de la actividad solar (GSN) el cual se recomienda hoy en día para analizar la actividad solar (R = Ns + 10 \* NG. Donde Ns es el número de manchas, y NG el número de grupos contabilizados en todo el disco solar visible), esta base de datos se pudo mejorar incorporando datos del tiempo histórico no considerados por Wolf.

En la actualidad, varios observatorios alrededor del mundo, además de la NASA, investigan la actividad solar ya que se ha comprobado que las manchas solares son indicador del comportamiento del Sol y su actividad magnética, así como la radiación que emana. Estas pueden influenciar todos los planetas, aún más a los que están cercanos al mismo, como la Tierra (variaciones del clima, períodos de lluvia y sequía, variación en la longitud del día, perturbaciones magnéticas que podrían afectar a la parte superior de nuestra atmósfera y las telecomunicaciones). Las instituciones astronómicas proveen de Bases de datos con acceso público para que astrónomos y aficionados alrededor del mundo puedan hacer sus análisis y contribuir a la comunidad científica. Dichas instituciones tienen ya realizadas sus predicciones de manchas solares para los siguientes 7 días, **este trabajo propone usar la base de datos que emiten las instituciones astronómicas para poder comprobar e igualar las estimaciones de manchas solares realizadas por ellos mismos, usando aprendizaje automático y ciencia de datos para poder obtener el mejor resultado.**

Desarrollo



**Definiendo el problema de Machine Learning**

La forma más sencilla de poder predecir las manchas solares es usando aprendizaje automático para la predicción de una variable numérica, en este caso **regresión lineal,** ya que estamos usando solamente datos planos, y no imágenes, video etc. (en este caso se usarían CNN, RNN o incluso RandomForest). Regresión es un modelo o método estadístico que trata de modelar la relación entre una variable continua y una o más variables independientes mediante el ajuste de una ecuación lineal.

**Construir el Dataset**

La base de datos que las instituciones astronómicas y científicas proporcionan incluye:

\*Datos gregorianos como son Año, Día y Mes.

\*Número de Manchas solares diarias, -1 indica que ese día no hay datos reales (clima, visibilidad, etc..).

\*Desviación Estándar diaria de las manchas solares visibles y las reales según el estándar GSN.

\*Número de observaciones diarias.

\*Indicador definitivo o provisional, 1 indica que es el valor definitivo y 0 indica que el valor es provisional.

**Transformar los datos**

El análisis de los datos se realizará usando diferentes librerías, como son:

\*Numpy. Operaciones entre matrices.

\*Pandas. Procesamiento de datos.

\*Matplotlib. Creación de gráficas.

\*Seaborn. Creación de gráficas.

\*Sklearn. Creación de modelos para regresión lineal.

**Entrenando el modelo de Machine Learning y sus predicciones.**

El primer paso sería la limpieza de datos, los cuales usaríamos Numpy y Pandas para este proceso. Seaborn es de gran ayuda para identificar la correlación entre las variables (indispensable para regresión lineal). Para la regresión lineal, se crearán 2 conjuntos principales (test y training) para el modelo, test tendrá un porcentaje del 20% del total de la muestra para su entrenamiento. Una vez creado el modelo, se realizarán las siguientes métricas de error para verificar la veracidad del modelo y sus predicciones:

\*Medium Square Error (MSE)

\*Medium Average Error (MAE)

\*Root Medium Square Error (RMSE)

\*Root mean logarithmic (RMLSE)

El coeficiente de relación también puede ayudarnos ya que es una métrica para medir la dispersión en los diagramas de **Regresión Lineal**, el cual cuantifica la relación lineal entre dos variables. Así podríamos identificar si nuestro análisis inicial para crear el modelo fue el correcto. Una de las mejores formas de visualización del resultado de un modelo de regresión es el obtenido de un scatter plot, donde en el eje X aparezca la predicción y en el eje Y el valor real.

Conclusiones

La regresión lineal es la solución más confiable para este problema hasta ahora, derivado de que la correlación de las manchas solares con las demás variables en el Dataset es consistente. Las métricas de error son esenciales para la confiabilidad del modelo creado. En un futuro se podría crear un modelo más avanzado, con redes neuronales o incluso se podría avanzar aún más y tener un modelo en el cual con base en imágenes del sol, nos pueda identificar manchas solares significativas e incluso proporcionar información acerca de su afectación a la Tierra y las Telecomunicaciones.

Referencias

* D Royal Observatory of Belgium, Brussels. (2022, August 10). *SIDC -- Solar Influences Data Analysis Center*. Sidc; Royal Observatory of Belgium, Brussels. https://www.sidc.be/
* SPACE WEATHER PREDICTION CENTER. (2022, August 10). *Homepage*. NOAA / NWS Space Weather Prediction Center; NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION. https://www.swpc.noaa.gov/
* Abhinand. (2022, August 10). *Daily Sun Spot Data (1818 to 2019)*. Kaggle. https://www.kaggle.com/datasets
* Vaquero, J. M., Mr, & Moreno-Corral, M. A., Mr. (2007). Historical sunspot records from Mexico. *Geofísica Internacional , Departamento de Física Aplicada, Universidad de Extremadura, Cáceres, Spain . Instituto de Astronomía, Universidad Nacional Autónoma de México, Campus Ensenada, Mexico*, *1*(47 (3)), 1–2.
* Echer, E., Mr, Rigozo, N. R., Mr, Nodermann, D. J. R., Mr, Vieira, L. E. A., Mr, Prestes, A., Mr, & De Faria, H. H., Ms. (2003). El numero de manchas solares, indicacion de actividad del Sol. *Revista Brasileira de Ensino de Fisica*, *25*(2).