Proyecto De Aprendizaje Máquina

Diseño de un Modelo de Red Neuronal para Predicción de Materia Particulada Antropogénica a partir de datos de aerosoles del sensor MODIS Terra.

Resumen

El trabajo siguiente, proyecto de la materia de Aprendizaje Máquina, trata del desarrollo de un modelo de Red Neuronal para Predicción de Materia Particulada Antropogénica con base en un entrenamiento a partir de un set de datos, de características de los aerosoles, obtenidos del sensor MODIS Terra.

Objetivo General

Diseñar un Modelo de Red Neuronal que permita realizar predicciones de materia particulada antropogénica. El modelo será entrenado con datos de aerosoles, tales como profundidad óptica, concentración de masa, tipo de aerosol, entre otros, todos obtenidos de mediciones del sensor MODIS Terra.

El modelo considerado, hasta el momento, es una red neuronal del tipo Recurrente (*Recurrent Neural Network*), con la posibilidad de utilizar una LSTM (*Long-Short Term Memory*) o incluso una RCNN (*Recurrent Convolutional Neural Network*).

Objetivos Particulares

Para el proyecto de la materia Aprendizaje Máquina, se delimita el alcance de éste, como objetivo final del curso, a los siguientes objetivos:

1. Aportar en desarrollo al proyecto final de la maestría, es decir, el Trabajo de Obtención de Grado (TOG).
2. Dicho aporte, deberá ser el resultad de la ejecución de los pasos 3, 4 y 5 definidos más adelante dentro del capítulo *Metodología*, los pasos 1 y 2 habiendo sido ya atendidos dentro del marco de una materia anterior a la de este curso.
3. Comprender el uso de las librerías Keras y Tensorflow, actualmente consideradas como la base de construcción del modelo de red neuronal.
4. Conocer los diferentes tipos de arquitecturas de redes neuronales, sus usos, ventajas y desventajas, para sí definir correctamente el modelo a utilizar, actualmente definido como del tipo Recurrente con opción a evaluar y/o cambiar a alguna de sus variantes.

Justificación

El Trabajo de Obtención de Grado permite hacer uso del espacio y tiempo necesario para el trabajo de este proyecto, por lo cual, este se considera un aporte al TOG y esto justifica la selección actual de los *datasets*, así como del modelo de red neuronal propuesto y de la metodología a utilizar.

Para la elección del modelo recurrente, de manera independiente se ha estado realizando un estudio en distintas bibliografías*.

*A adjuntar más adelante durante el desarrollo del proyecto de esta materia.

Metodología

Para la realización de este proyecto se utilizará el ciclo de desarrollo CRISP-DM (*Cross-Industry Standard Process for Data Mining*), siguiendo cada una de las etapas.

1. Comprensión del negocio: entender el contexto de los atributos o características de los aerosoles y su relación con la materia particulada antropogénica.
2. Comprensión de los datos: los datos deben ser aquellos proporcionados por el sensor MODIS del satélite Terra de la NASA; la fuente de obtención es a través del sitio web de la NASA. Más detalles pueden ser encontrados en https://github.com/JFrankVC/tog/blob/e373ad75959614746c05fff6e0b195c98aa6c7ac/data_treatment/reports/Reporte_Proyecto_TOG_Recopilaci%C3%B3n_y_Tratamiento_de_Datos.pdf.
3. **Preparación de los datos**: con las técnicas aprendidas en el curso de Aprendizaje Máquina, se realizará un análisis más profundo sobre las características de los datos, de tal manera que estos puedan ser utilizados de manera adecuada dentro del modelo de red neuronal propuesto. Así mismo, se definirá la partición entre *training* y *test*.
4. **Modelado**: en esta fase es en donde se llevará a cabo el diseño y construcción del modelo propuesto, haciendo uso de los datos de *training* tratados anteriormente.
5. **Evaluación**: una vez el modelo construido, se llevarán a cabo evaluaciones con el fin de medir el performance, haciendo uso principalmente de los datos *test*.
6. Despliegue: uso del modelo obtenido a partir de datos no antes vistos.

Cabe destacar que, dentro de este ciclo de desarrollo, se contempla llevara a cabo mini fases cíclicas

Descripción de los datos

Los datos obtenidos, en resumen, tienen las siguientes características:

1. Formato HDF: para manejar este tipo de datos es necesario hacer uso, en Python, de librearías tales como `rioxarray` (en máquina local) o `pyhdf` (en *Google Colab*).
2. Se trata de imágenes: matrices de pixeles con una o más bandas, donde cada banda representa una característica de los aerosoles, como lo puede ser la profundidad óptica, el tipo de aerosol, entre otros.
3. Son datos temporales: se tiene un *dataset* para una región, para un día específico dado en días julianos (*Day of the Year*).

Para comprender el manejo de los datos, además del reporte proporcionado en el paso 2 de la metodología, en el siguiente enlace se encuentra un *Notebook* de *Jupyter* en el cual se muestra qué librerías utilizar y la manera básica de empezar a hacer uso de estas para el manejo de los datos: https://github.com/JFrankVC/tog/blob/e373ad75959614746c05fff6e0b195c98aa6c7ac/data_treatment/modis_data_treatment.ipynb