3era Sesión: Preguntas y Respuestas

Por favor escriban sus preguntas en el cuadro para preguntas. Si tiene preguntas adicionales, por favor comuníquese con cualquiera de los siguientes instructores:

Erika Podest (erika.podest@jpl.nasa.gov)

Pregunta 1: ¿Qué métodos de aprendizaje automático recomiendas para estudios de calidad del agua de ríos urbanos?
[Translation] Which machine learning methods do you recommend for water quality studies of urban rivers?
Respuesta 1: El algoritmo a utilizar dependerá de los datos que tenga disponibles y de cómo desee medir la calidad del agua (valores discretos o continuos). Aquí hay algunas referencias sobre algunos algoritmos de aprendizaje automático utilizados para la calidad del agua (por ejemplo, redes neuronales, bosque aleatorio, etc.): https://doi.org/10.1016/j.jhydrol.2019.124084, https://doi.org/10.1016/j.jwpe.2022.102920

Pregunta 2: Si quisiera aprender más sobre cómo adaptar estos algoritmos vistos en el seminario a un estudio de caso personal, ¿qué sitios web o cursos me recomendarías?
[Translation] If I wanted to learn more about how to adapt these algorithms seen in the seminar to a personal case study, what websites or courses would you recommend?
Respuesta 2: La primera recomendación es simplemente tomar el código, mirar los datos disponibles y generar un conjunto de datos en el mismo formato. Esto le permitirá reutilizar la mayor parte del código proporcionado en esta capacitación. También he adjuntado algunos ejemplos a continuación de capacitaciones adicionales que podrían proporcionar algunas técnicas más avanzadas.

Curso gratuito: https://www.coursera.org/specializations/machine-learning-introduction

https://developers.google.com/machine-learning

Libro(s):

Machine Learning con PyTorch y Scikit-Learn (https://www.oreilly.com/library/view/machine-learning-with/9781801819312/)

Deep Learning (https://www.deeplearningbook.org/)

Materiales que producimos para introducir conceptos de ML:

https://github.com/astg606/py_courses/tree/master/modules/machine_learning

https://www.coursera.org/learn/machine-learning-models-in-science

Libro interactivo con muestras de código (Deep learning específicamente): https://d2l.ai/

Pregunta 3: Al principio de la presentación se comentó que hypertuning era benéfico cuando hay datos "outliers". En el ejemplo, ¿cómo se podrían identificar estos outliers?
[Translation] At the beginning of the presentation it was commented that hypertuning was beneficial when there is "outlier" data. In the example, how could those outliers be identified?
Respuesta 3: En general, el ajuste de hiper parámetros se usa para el conjunto de datos general y no solo es beneficioso para los valores atípicos. Sin embargo, nos ayuda a generalizar a través de muestras no vistas que podrían no ser tan similares al conjunto de datos de entrenamiento. Puede identificar valores atípicos a través de diagramas de caja o histogramas. La Sesión 2 cubre estos en la sección Análisis exploratorio de datos (EDA).

Pregunta 4: Con la función de pred_prob, ¿podría crear un mapa de esas probabilidades a manera de acompañar la clasificación con un mapa de límites de confianza o posibles lugares con error en mi clasificación?

[Translation] With the pred_prob function, could I create a map of those probabilities in order to accompany the classification with a map of confidence limits or possible places with errors in my classification?

Respuesta 4: pred_prob podría ser un primer paso para evaluar la certeza del modelo al realizar la predicción; sin embargo, esto solo le indica la perspectiva del modelo y no los valores de incertidumbre de la decisión real. A veces, cuando los modelos se sobreajustan, tienden a confiar demasiado en las observaciones que no son correctas y esto distorsiona sus límites de confianza. Puede utilizar técnicas de explicabilidad para poder visualizar esos valores de confianza.

Pregunta 5: ¿ML y autoML se complementan, o se pueden hacer individualmente?

[Translation] Do ML and autoML complement each other or can be done individually?

Respuesta 5: AutoML es una herramienta debajo del concepto de ML. Por consiguiente, cuando estás entrenando modelos utilizando AutoML, en definitiva estás practicando ML.

Pregunta 6: ¿Qué tanto puede impactar en una clasificación automática las anomalías de reflectancia presentes de forma inherente en las imágenes satelitales (insumos en general)? ¿Qué tan importante considera que sea eliminar/enmascarar estos píxeles antes del entrenamiento del modelo?

[Translation] How much can reflectance anomalies inherently present in satellite images (inputs in general) impact an automatic classification? How important do you think it is to remove/mask these pixels before training the model?

Respuesta 6: Las anomalías de reflectancia pueden definitivamente afectar el rendimiento de un modelo de aprendizaje automático. Los valores que no forman parte de la distribución natural de los datos pueden hacer que el modelo de ML no tenga un rendimiento adecuado. Sin embargo, las anomalías que son de la naturaleza, por ejemplo, cuando el NDVI es muy alto debido a los bajos valores espectrales del infrarrojo cercano y rojo, es importante incluirlos en los datos de entrenamiento para que el modelo pueda aprender esto. Se recomienda limpiar y remover cualquier anomalía que no sea natural (errores de instrumentos, errores de preprocesamiento) durante el preprocesamiento y la limpieza de datos que se realiza durante la exploración de datos.

Pregunta 7: ¿El costo se relaciona al costo computacional?
[Translation] Is the cost related to the computational cost?
Respuesta 7: Sí, note que computacionalmente se tarda mucho más porque se está utilizando un gran número de modelos.

Pregunta 8: Quiero preguntar, ¿cuánto tiempo estará disponible el material del curso en github?
[Translation] I want to ask how long will the course material be available on github?
Respuesta 8: No hay tiempo límite. Puede accederlo luego de finalizada la sesión.

Pregunta 9: ¿Existe un límite en el número de clases a usar en estos métodos de machine learning?
[Translation] Is there a limit to the number of classes to use in these machine learning methods?
Respuesta 9: No existe un límite estricto para el número de clases en la clasificación multiclase. Existen límites prácticos que dependen de varios factores, como la complejidad de los datos, la cantidad de datos de entrenamiento y la complejidad del algoritmo ML que se utiliza. El rendimiento del algoritmo ML puede degradarse a medida que aumenta el número de clases. De tener pocas observaciones de algunas clases, puede que sea mejor agregarlas en una sola.

Pregunta 10: Como consulta en apego a su experiencia, ¿es posible que el aprendizaje automático podría ampliarse al entrenamiento inferencial de estimación de biomasa en áreas con cobertura boscosa con base a imágenes ráster Sentinel 1 o ALOS PALSAR?
[Translation] As a query based on your experience, is it possible that machine learning could be extended to inferential training of biomass estimation in areas with forest cover based on Sentinel 1 or ALOS PALSAR raster images?
Respuesta 10: Definitivamente. Algunos ejemplos de referencia adjuntos.

Pham, T. D., Yoshino, K., Le, N. N. y Bui, D. T. (2018). Estimating aboveground biomass of a mangrove plantation on the Northern coast of Vietnam using machine learning techniques with an integration of ALOS-2 PALSAR-2 and Sentinel-2A data. *International Journal of Remote Sensing*, *39*(22), 7761-7788.

Singh, A., Kushwaha, S. K. P., Nandy, S., Padalia, H., Ghosh, S., Srivastava, A. y Kumari, N. (2023). Aboveground Forest Biomass Estimation by the Integration of TLS and ALOS PALSAR Data Using Machine Learning. *Remote Sensing*, *15*(4), 1143.

Pregunta 11:¿ Cuál es el método más adecuado para validar la precisión de un algoritmo? ¿Es importante contar con datos de presencia y no presencia, o solo bastaría con presencia?

Respuesta 11: Scikit-learn tiene una función para calcular precisión, incluyendo muchas otras métricas que puede evaluar desde su documentación. Existen un sin número de métricas para evaluar su modelo que dependerá de qué métricas serán de importancia para su investigación. Adjunto un artículo que utilizamos en nuestro grupo para formalizar nuestra práctica de validación.
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E. y Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote sensing of Environment*, *148*, 42-57.

La alta precisión para los datos sin presencia es más importante porque significa que el algoritmo realmente aprende el patrón y la relación del conjunto de datos de entrenamiento y puede solicitar nuevos datos, no solo memorizar el conjunto de entrenamiento.

Pregunta 12: ¿Qué material de referencia recomiendan para continuar investigando sobre explicabilidad e interpretabilidad (XAI)?

Respuesta 12: Adjunto algunas referencias:
https://christophm.github.io/interpretable-ml-book/, https://arxiv.org/abs/1910.10045.

Pregunta 13: ¿Cuáles algoritmos de Machine Learning son mejores para el análisis de crecimiento urbano?

Respuesta 13: Definitivamente dependerá de los datos de entrenamiento disponibles y las necesidades de investigación que discutimos en la Sesión 1. Adjunto algunas referencias relacionadas con el tema y sus algoritmos.

Gómez, J. A., Patiño, J. E., Duque, J. C. y Passos, S. (2019). Spatiotemporal modeling of urban growth using machine learning. *Remote Sensing*, *12*(1), 109.

Kim, Y., Safikhani, A. y Tepe, E. (2022). Machine learning application to spatio-temporal modeling of urban growth. *Computers, Environment and Urban Systems*, *94*, 101801.

Pregunta 14: En mi caso, clasifico imágenes directamente desde Google Earth Engine. Como no tienen todas estas bibliotecas disponibles, ¿es posible utilizar estos algoritmos en un cuaderno del Google Colab, tomando las imágenes desde el repositorio de Google Engine sin descargarlas localmente? Es decir, pasarlo a formato numpy desde la nube de Engine directamente, sin tener que descargarlas y tener que importarlas como .tiff
[Translation] In my case I classify images directly from Google Earth Engine. Since you don't have all these libraries available, is it possible to use these algorithms in a Google Colab notebook, pulling the images from the Google Engine repository without downloading them locally? That is to say, pass it to numpy format from the Engine cloud directly, without having to download them and having to import them as .tiff
Respuesta 14: Sí, definitivamente es una posibilidad. Adjuntamos un enlace de ejemplo que podría ayudar. Necesitará saber de qué colección de datos le gustaría descargar los datos.
https://developers.google.com/earth-engine/tutorials/community/intro-to-python-api

Pregunta 15: ¿Cómo se integran al modelo de machine learning datos in situ y datos satelitales como el NDVI sabiendo que representan diferentes valores?
Respuesta 15: Puede usar el NDVI como guía para crear sus datos de entrenamiento. Luego puede crear su conjunto de datos combinando las imágenes satelitales y sus datos in situ.

Pregunta 16: ¿Las clases poco balanceadas afectan el resultado de la clasificación? Lo pregunto porque en el caso de la vegetación, en muchas ocasiones es difícil tener la misma cantidad de información de entrenamiento por clase. Cuando no se puede solucionar este problema, ¿qué se recomienda?
Respuesta 16: Sí, puede afectar el rendimiento del modelo. Según el algoritmo, puede usar técnicas de muestreo descendente y ascendente para mejorar el equilibrio del conjunto de datos (suponiendo que se trate de datos de entrenamiento basados en puntos). También podría usar funciones de pérdida que penalizarían su modelo para aprender de las clases subrepresentadas (esto es común cuando se usan redes neuronales) en casos donde se tengan datos de entrenamiento continuos.

Pregunta 17: Existen alertas de probable pérdida de bosque (deforestación). ¿Podría haber un algoritmo de probable tala? ¿Se podría distinguir ese tipo de perturbación en el pixel?

Respuesta 17: Definitivamente. Hay diferentes formas de ver este problema. Podría usar modelos conscientes temporalmente para identificar regiones de cambio relacionadas con la deforestación. También puede crear mapas individuales por día para mapear la cobertura de árboles y luego crear mapas de diferencia para identificar esta deforestación. Incluso podría entrenar su modelo para detectar árboles deforestados en función de sus datos de entrenamiento.