Machine Learning y DataOps

DOI: 10.29236/sistemas.n165a2

La automatización de los procesos de Machine Learning facilitan el desarrollo continuo de Analítica.

Holman Diego Bolívar Barón

Desde tiempos inmemoriales, los seres humanos han construido máquinas para simplificar su trabajo y reducir el esfuerzo en completar diferentes tareas. Sin conocer ninguna de las leyes de la mecánica Newtoniana o los teoremas mecánicos de Arquímedes, inventaron palancas, instrumentos y máquinas complejas para realizar procedimientos sofisticados. Mover grandes y pesadas rocas o troncos de madera se hizo más sencillo utilizando un carro. Los pueblos primi-

tivos simplemente observaron cómo la rueda podía mejorar sus vidas (Ethem, 2016).

Una máquina se considera útil si sus usuarios pueden entender fácilmente qué tareas se pueden completar con menos esfuerzo o automáticamente, máquinas automáticas, construidas y programadas para lograr objetivos específicos al transformar la energía en trabajo, los molinos de viento son un ejemplo de herramientas elementales que pueden llevar a cabo tareas completas con un control humano mínimo.

Los seres humanos han tratado de transferir algo de inteligencia a sus herramientas desde los albores de la tecnología. Una rueda puede girar alrededor de un eje fijo millones de veces y el viento debe encontrar una superficie adecuada para empujarla, el camino comenzó con las herramientas, prosiguió con las máquinas simples y luego a las inteligentes (Rahman, 2020).

Las computadoras programables son instrumentos muy extendidos, flexibles y cada vez más poderosos; además, la difusión de Internet permitió compartir aplicaciones de software e información relacionada con un mínimo de esfuerzo. El software de procesamiento de texto, un cliente de correo electrónico, un navegador web y muchas otras herramientas comunes que se ejecutan en la misma máquina, son ejemplos de esa flexibilidad. Es innegable que la revolución de TI mejoró nuestros trabajos diarios, pero sin el aprendizaje automático ML (por Machine Learning), todavía habría muchas tareas que serían ajenas a la informática. El filtrado de spam, el procesamiento de lenguaje natural NLP (por Natural Language Processing), el seguimiento visual con una cámara web o un teléfono inteligente y el análisis predictivo son solo algunas de las aplicaciones que revolucionaron la interacción hombre-máquina, transformando herramientas electrónicas en extensiones cognitivas reales que están cambiando y reduciendo la brecha entre la percepción humana, el lenguaje, el razonamiento y entidades artificiales.

ML no se basa en estructuras estáticas o permanentes, sino en una capacidad continua para adaptar su comportamiento a señales externas correspondientes a conjuntos de datos o entradas en tiempo real y, como un ser humano, estimar un escenario o situación futura con información incierta y fragmentaria (Bonaccorso, 2018).

Por otra parte, la cantidad de datos que se genera es inconmensurable, con base en la publicación de Lori Lewis (2021) en el año 2021 durante 60 segundos se realizaban 1.4 millones de scrolling en Facebook, 9.132 conexiones a Linkedin, 28.000 subscriptores de Netflix estaban viendo alguna serie o película, 1.6 millones de personas estaban haciendo compras online, 69

millones de personas enviaban mensajes por WhatsApp, se generaron 200.000 twits y 500 horas de video eran subidas y reproducidas en Youtube, entre muchas otras.

Actualmente, la cantidad de datos que se producen, se transportan y se transforman es superior a la cantidad de datos generados por la humanidad a lo largo de la historia, por tal razón, no es posible leer, comparar y analizar todos los datos, incluso para supercomputadoras esto puede llegar a tomar mucho tiempo, por tal razón es necesario recurrir a técnicas estadísticas para el análisis de la gran explosión de datos que se genera cada minuto.

Todo modelo de ML requiere mínimo un dataset de entrada, sin embargo, para implementar el modelo, este requiere cierta homogeneidad entre los datos para compararlos e identificar patrones y estructuras que permitan establecer conjuntos o grupos de datos. Un grupo homogéneo de datos es lo que se conoce comúnmente como clúster.

El agrupamiento es el proceso de análisis para clasificar datos en una serie de grupos, en cada grupo los datos son similares entre sí bajo ciertos parámetros o criterios, entre más criterios cumple un grupo de datos se establece que la calidad del grupo es mayor. El agrupamiento es comúnmente denominado clustering (por su denominación en inglés). En clustering también es posible segmentar clusters con propiedades o comportamiento similares y asignar los datos a alguno de ellos. Es posible que no se logre identificar un parámetro o criterio para realizar clustering con los datos, pero es necesario una segmentación, por lo tanto, los datos se pueden agrupar aleatoriamente, esta técnica es conocida como CLARANS (por Clustering Large Applications based upon Randomized Search).

El proceso de calidad de los datos es fundamental dentro de un ejercicio de ML, para ello se suelen desarrollar ejercicios de ETL (por Extract, Transform and Load) que homogenizan los datos para que sean comparables y evaluables entre sí. Una sola proyección puede llegar a requerir la integración de decenas de sistemas de información para estructurar ese Dataset que permite realizar el ejercicio de ML.

Ese proceso de homogenización se puede desarrollar realizando ejercicios de ingeniería de datos apoyándose en nuestro bien apreciado y ponderado SQL (por Structured Query Language) también se puede desarrollar con funciones en lenguaje R o Python. Así mismo, existen herramientas para automatizar este proceso como Databricks, Stratio o Alteryx.

En proyectos robustos de ML se debe asegurar el ciclo de vida del dato y el cumplimiento de los lineamientos generales para la imple-

mentación de la infraestructura de datos con base en la resolución 460 del 15 de febrero de 2022 del Ministerio de Tecnologías de la Información y las Comunicaciones de Colombia.

Para poder abordar la complejidad, incertidumbre y ambigüedad en el desarrollo de los productos de ML y teniendo en cuenta el artículo 2.2. 9.1.1.3 del Decreto 767 del 16 de mayo de 2022 y el lineamiento del documento Maestro del Modelo de Gestión de Proyectos TI, planteado por el Ministerio de Tecnologías de la Información y Comunicaciones, el cual establece en el numeral 7.2. 10. MGPTI.LI.PLA.10 que: “Se debe considerar el uso de metodologías ágiles y la aplicación del manifiesto Ágil”.

Una metodología ágil podría ser Scrum, la cual permite el trabajo colaborativo mientras se aborda cada uno de los problemas generados durante el desarrollo e implementación, "Con scrum, un producto se basa en una serie de iteraciones llamadas sprints que dividen proyectos grandes y complejos en porciones minúsculas", “Gracias a los sprints, los proyectos son más fáciles de gestionar, permiten a los equipos enviar trabajo de gran calidad más rápido y con más frecuencia, y les ofrecen más flexibilidad para adaptarse al cambio” (Drumond, 2022).

Scrum como metodología de desarrollo de productos de software se enfoca en un desarrollo continuo lo que permite generar innovación, pasando por diferentes MVP (por Minimum Viable Product); sin embargo, la complejidad del despliegue dependerá de los recursos para la correcta funcionalidad de la aplicación, así como la interacción con el o los sistemas operativos sobre los cuales se ejecutan los servidores que soportan la funcionalidad de esta. Cuando la aplicación se mantiene en evolución y desarrollo constante, el despliegue debe automatizarse bajo el paradigma de integración y despliegue continuo CI/CD (por Continuous Integration and Continuous Delivery), lo que actualmente se denomina DevOps (por Development and Operations).

En el marco de los ejercicios de analítica de datos se sigue el esquema propuesto por IBM a través del CRISP-DM (Por Cross Industry Standard Process for Data Mining) que contempla una fase de entendimiento del negocio y de los datos, seguido de fases de preparación, modelamiento, evaluación e implementación.

A la hora de implementar una metodología que permita la automatización del despliegue de ejercicios de ML aparece MLOps (por Machine Learning Operations), ya que proporciona un proceso de desarrollo de ML de extremo a extremo para diseñar, construir y administrar software reproducible, comprobable y evolutivo impulsado por ML.

MLOps tiene como objetivo unificar el ciclo de lanzamiento para el aprendizaje automático y el lanzamiento de aplicaciones de software, además, permite hacer pruebas automatizadas de artefactos de aprendizaje automático (por ejemplo, validación de datos, pruebas de modelos de ML y pruebas de integración de modelos de ML). MLOps facilita la aplicación de principios ágiles a proyectos de aprendizaje automático y reduce la deuda técnica en todos los modelos de aprendizaje automático. MLOps debe ser una práctica independiente del lenguaje, el marco, la plataforma y la infraestructura y permite admitir modelos y conjuntos de datos de aprendizaje automático para construir estos modelos como ciudadanos de primera clase dentro de los sistemas de CI/CD.

Referencias

Bonaccorso. Giuseppe., (2018). Machine Learning Algorithms: Popular Algorithms for Data Science and Machine Learning, 2nd Edition: Vol. 2nd ed. Packt Publishing.

Drumond. Claire, (2022), What is scrum?, Atlassian, disponible en: https://www.atlassian.com/agile/scrum

Ethem Alpaydin. (2016). Machine Learning : The New AI. The MIT Press.

IBM, (2021), CRISP-DM Help Overview, disponible en: https://www.ibm.com/docs/en/spss-modeler/SaaS?topic=dm-crisp-help-overview

Lewis. Lori, (2021), 2021 This is What Happens in An Internet Minute. https://www.allaccess.com/merge/archive/32972/infographic-what-happens-in-an-internet-minute.

MinTIC, Artículo 2.2.9.1.2.1. “Componentes” del Capítulo 1 “ESTRATEGIA DE GOBIERNO EN LÍNEA” del Título 9 del Decreto 1078 de 2015

Rahman. Was., (2020). AI and Machine Learning. Sage Publications Pvt. Ltd.

Holman Diego Bolívar Barón. *Doctor en Informática por la Universidad Pontificia de Salamanca, miembro profesional de la Association for Computing Machinery (ACM), líder del grupo de investigación en Software Inteligente y Convergencia Tecnológica –GISIC-. DEA en Ingeniería del Software, profesor de la Universidad Católica de Colombia e investigador asociado de Minciencias. Su área de investigación se centra en el desarrollo de software para entrenamiento cognitivo y modelos de desarrollo para la evaluación del aprendizaje en entornos interactivos a través de la lógica difusa. Actualmente trabaja en la estructuración de una plataforma de ciencia de datos y aprendizaje de máquina automatizado de código abierto para la investigación de acceso público al servicio de la sociedad.*

Calendario de Eventos 2023

MAR 06 - 10 — GEODATOS

MAR — MARATÓN REGIONAL LATINOAMERICANA DE PROGRAMACIÓN ACIS/REDIS

ABR — JORNADA DE GERENCIA DE PROYECTOS TI

MAY — JORNADA DE CIUDADES + TECNOLOGÍA

JUN — JORNADA INTERNACIONAL DE SEGURIDAD INFORMÁTICA

AGO — ACISTIC

SEP — MARATÓN NACIONAL DE PROGRAMACIÓN ACIS/REDIS

OCT — REDUC@TE

NOV — ENCUENTRO DE REDIS

MAS INFORMACION EN :
WWW.ACIS.ORG.CO
3015530540 - 3043463413

ACIS