Lección 2: Los datos y el Machine Learning

Fuente: Universia.

Es innegable la gran utilidad que representa el machine learning en los diferentes campos como economía, salud, transporte, tecnología, por mencionar algunos, pero se debe tener en cuenta que esta tecnología depende de los datos que le son suministrados para que esta lleve a cabo su labor, siendo estos el combustible que potencia la calidad y efectividad de su trabajo.

Los datos son el conjunto de muestras que se utilizan para entrenar el sistema, cada segundo, los seres humanos se convierten en una fuente de datos de manera inintencionada, al compartir sus preferencias de consumo, intereses, valores a través de diferentes plataformas que son de uso cotidiano en la actualidad, como lo son las redes sociales como Twitter, Instagram, e-mail, LinkedIn, Facebook, WhatsApp, entre otras.

La web, las transacciones bancarias, las interacciones de maquina a máquina como el GPS, Wi-Fi o el navegar en internet generan una enorme cantidad de datos , que para el ser humano sería un enorme trabajo el procesamiento de esta cantidad de datos, pero gracias a herramientas como el machine learning (Negocios), esta operación puede ser llevada a cabo en fracciones de segundo , esto tarda el ML en corregir errores además de tener una gran capacidad de almacenamiento gracias a la implementación de cloud computing, este sistema permite el guardado de archivos, programas o sistemas en la nube de manera ilimitada.

Cotidianamente 2.5 exabytes de datos son generados, dicho en otras palabras, 2.5 billones de gigabytes alrededor el mundo (equivalente a 530 millones de canciones) y se cree que para el año 2020, esta gran cantidad de información se habrá multiplicado 40 veces 2, este fenómeno es hoy conocido como Big Data. El Massachusetts Institute of Technology (MIT) la define este término como procesamiento y estudio de grandes volúmenes de información que pueden ser estructurados, semi-estructurados y no estructurados, es decir, el Big data son datos muy variados, que además se exhiben en volúmenes crecientes y a una gran velocidad, lo que se conoce como las tres V’s del Big data.

En otras palabras, el big data se encuentra formada por conjuntos de datos más grandes y de mayor complejidad, los cuales proceden de nuevas fuentes de datos. Estos conjuntos de datos al ser de tal volumen, los softwares de procesamiento más común no cuentan con la capacidad de gestionarlos.

Como se menciona anteriormente, el Big data cuenta con las tres V’s que se presentan a continuación:

Volumen

- Con Big data, se procesan grandes volúmenes de datos que no están estructurados y son de baja densidad. Podrían ser datos de valor desconocido, ejemplo: feeds de datos de Twitter, flujos de clics de una página web, aplicación para móviles, equipo con sensores, etc. Para algunas compañías, esto puede traducirse en decenas de terabytes de datos. Para otras, incluso petabytes.

La velocidad

- se refiere al ritmo al que se reciben los datos. Estos son transmitidos a una gran velocidad y deben ser distribuidos de manera oportuna. Las etiquetas FID, sensores y la medición inteligente crean la necesidad de distribuir corrientes de datos casi en tiempo real.

La variedad

- esta se refiere a los diversos tipos de datos que están disponibles. Anteriormente, los datos convencionales eran estructurados y se podían organizar en una base de datos relacional. Gracias al auge del big data, los datos se presentan en nuevos tipos de datos no estructurados. Los datos no estructurados y semiestructurados, como por ejemplo el texto, audio o vídeo, requieren de un preprocesamiento para poder obtener significado y habilitar los metadatos.

Teniendo en cuenta lo anterior, es inadmisible pensar en Big Data eficiente que no utilice tecnologías de inteligencia artificial como el Machine Learning. En las compañías donde se manejan grandes cantidades de datos no sería posible que un empleado llevara a cabo un análisis de cada conexión entre ellos para así generar conclusiones que vuelvan un servicio rentable. Por ejemplo, compañías como Amazon hacen uso del Machine Learning para investigar entre sus millones de productos y brindar a los usuarios solo aquellos que más se ajustan a su perfil. También utilizan esta herramienta para predecir el número de bajas de clientes que se fueron a utilizar los servicios de la competencia, por medio del análisis de qué

motivos los llevaron a hacerlo o por el contrario, la fidelización de nuevos clientes que se encontraron satisfechos con las solución de sus necesidades.

A partir de esta información se pueden especificar los principales usos del Big data de la siguiente manera:

- Desarrollo de productos: las empresas hacen uso del Big data para predecir la demanda de los clientes, construyen modelos predictivos a través el análisis y clasificación de atributos clave de productos anteriores y actuales, posteriormente realizan la modelación de la relación entre estas características y el éxito comercial de las ofertas.
- Mantenimiento predictivo: por medio del análisis de indicadores de problemas que pueden presentarse antes de que estos se ocurran, las compañías tendrán la capacidad de realizar el mantenimiento de una forma más rentable, además de optimizar el tiempo de servicio de sus activos.
- Experiencia del cliente: la disponibilidad de datos que permitan tener una vista clara de cómo ha sido la experiencia del cliente permite mejorar los servicios que se prestan, así como mejorar la interacción y la fidelización del cliente.
- Fraude: El Big data le permite identificar patrones en los datos, los cuales podrían estar indicando un posible fraude, al tiempo que logra concentrar bastos volúmenes de información para acelerar la creación de informes normativos.
- Eficiencia operativa: al poder evaluar los procesos de producción dentro de las empresas, anticipar la demanda, la opinión de los clientes y la demanda futura, el Big data permite la mejora de los procesos productivos y la toma de decisiones en tiempo real.
- Impulso de la innovación: gracias a la disponibilidad de información que ofrece el Big data, los procesos de innovación serán más constantes y efectivos ya que este se basaría en información real y confiable, mediante el estudio de las dependencias entre seres humanos y las diferentes instituciones.

Como se pude observar, el Big data posee numerosos usos y aporta grandes beneficios a quien haga uso de estas grandes cantidades de información, pero debe tenerse en cuenta que estos grandes volúmenes pueden ser procesados gracias al aprendizaje automático, se establece una relación de reciprocidad dado que, la información que posee el Big data no será potencial hasta no ser utilizada, por otra parte el aprendizaje automático es posible gracias a la disponibilidad de datos que le permite a las maquinas aprender a partir de la experiencia, a mayor cantidad de datos se obtienen resultados de mayor calidad.

Debido a lo anterior, se resalta la importancia de la calidad de datos de los cuales hace uso el machine learning, puesto que estos deben representar toda la diversidad de opciones que se pueden dar, dichos datos deben estar completos y haber sido recolectados de manera adecuada donde no se encuentren datos repetidos o datos atípicos conocidos como “outlayers”. De la calidad y cantidad de los datos dependerán los resultados obtenidos, es por ello que la información recopilada de la cual se hará uso deberá cumplir con los criterios anteriormente mencionados.

Referencias:

- España, U. (12 de septiembre de 2017). Universia España. Recuperado el 02 de octubre de 2018, de http://noticias.universia.es/ciencia-tecnologia/noticia/2017/09/12/1155659/machine-learning-como-usa-big-data.html
- Negocios, U. d. (s.f.). ProMéxico. Recuperado el 02 de octubre de 2018, de http://mim.promexico.gob.mx/work/models/mim/Resource/152/1/images/machine-learning.pdf
- ORACLE. (s.f.). ORACLE. Recuperado el 02 de octubre de 2018, de https://www.oracle.com/co/big-data/guide/what-is-big-data.html
- SAS. (s.f.). SAS. Recuperado el 02 de octubre de 2018, de https://www.sas.com/es_co/insights/big-data/what-is-big-data.html