BIG DATA Y MACHINE LEARNING CON PYTHON

Proyecto de fin de ciclo: **Juan Medina**

Parte teórica

¿Qué es el Big Data?

Al principio, los únicos datos que se manejaban eran recolectados por algunas personas y se incluían en pequeñas bases de datos o en simples hojas de cálculo.

Algunos años más tarde, cada usuario empezó a generar sus propios datos a través de, por ejemplo, las redes sociales, y el volumen de los mismos se incrementó de forma sustancial

Y en los últimos años hemos asistido a una verdadera explosión de la cantidad de datos que generamos con la popularización de los smart phones que recopilan datos de cada usuario gracias a los sensores que incorporan, como el GPS. Además, estos sensores se añaden cada día a mas dispositivos, desde relojes hasta frigoríficos, e incluso a la ropa. Y esta imparable generación de datos no hace sino multiplicarse cada día y es previsible que siga haciéndolo con la llegada de nuevos dispositivos (por ejemplo los coches autónomos).

A toda esta cantidad de datos y a las herramientas que usamos para gestionarlos es a lo que llamamos Big Data.

El modo de acceder a todos estos datos también puede ser muy diferente. Algunas veces los recibiremos en tiempo real por Internet y otras se nos proporcionaran directamente un archivo que los contenga.

También es posible acceder a una gran cantidad de datos almacenados a los que nos proporcionan acceso diferentes fuentes.

En este sentido podemos destacar la idea de Open Data que se basa en la premisa de que los datos deben ser accesibles para todo el mundo y todo el mundo debe tener el derecho de usarlos y publicarlos, sin ningún tipo de restricción de copyright, patentes u otros mecanismos de control. Esto es especialmente cierto en el caso de los datos recopilados por los Gobiernos a través de los años. Estos datos no pertenecen al Gobierno, sino que pertenecen a los ciudadanos, así que mucho gobiernos se han adherido a esta idea y han creado sitios web en los que es posible acceder a estos datos.

Algunos ejemplos de estos sitios web son:

- España: http://datos.gob.es/
- Irlanda: https://data.gov.ie/
- EEUU: https://www.data.gov/
- Francia: https://www.data.gouv.fr/

En cualquier caso el acceso a este tipo de datos no siempre es fácil ni se conoce de antemano como los vamos a encontrar, por lo que Tim Berners-Lee propuso un sistema de estrellas para saber cómo de accesible es una fuente:

- **Una estrella:** Datos accesibles en cualquier tipo de formato
- **Dos estrellas:** Datos accesibles organizados en un formato propietario (Datos en Excel)
- **Tres estrellas:** Datos accesibles organizados en un formato no propietario (csv en lugar de Excel)
- **Cuatro estrellas:** Los datos ademas incluyen una URL unica y los datos se estructuran bajo el estandar W3C
- **Cinco estrellas:** Los datos ademas estan enlazados con otros datos son una estructura similar.

¿Qué es el Machine Learning?

El Machine Learning, o aprendizaje automático, se refiere a la ciencia que permite a los ordenadores aprender sin que hayan sido explícitamente programados para algo, sino que son capaces de generalizar comportamientos a partir de una información suministrada en forma de ejemplos, de este modo son capaces de utilizar su experiencia para tomar decisiones futuras más acertadas

Existen varios modos de desarrollar este tipo de aprendizaje, pero las más usuales son:

- **Aprendizaje supervisado:** Se proporciona al programa una serie de datos que incluyen la solución correcta. El programa aprenderá de ellos y será capaz de generalizar una solución para otra serie de datos que no la incluyen.
- **Aprendizaje no supervisado:** Solo se le proporciona una serie de datos al programa que el mismo debe clasificar en grupos encontrando patrones en los mismos. Esto es muy útil por ejemplo en marketing para poder segmentar el mercado en diferentes nichos. Los humanos usamos este tipo de aprendizaje para aprender a hablar. Al principio solo escuchamos sonidos que no somos capaces de entender, pero poco a poco vamos reconociendo patrones.

- **Aprendizaje por refuerzo:** Esta clase de entrenamiento se basa en la técnica de ensayo/error. El programa predice una respuesta y se le entrega la solución, cuantas más veces se ejecute más posibilidad de acierto tendrá. Es lo mismo que nosotros hacemos en ciertos juegos, probamos una estrategia y vemos su resultado y así en el siguiente turno podemos mejorarla.

Encontraremos otros problemas, conceptos y herramientas del Machine Learning en la parte práctica y trataré de explicarlos entonces.

Competiciones de Kaggle

En el año 2006 Netflix hizo público un concurso para mejorar su propio algoritmo de recomendación de películas.

Netflix pensó, con buen criterio, que era mejor y conseguiría mejores resultados con un concurso de premio espectacular (1.000.000 de dólares) que pagando ese dinero a una compañía que hiciese el trabajo.

El concurso se hizo muy popular y se formaron gran cantidad de equipos que lucharon por alzarse con el premio hasta que finalmente uno de ellos lo logró en el año 2009... y bueno, la verdad es que Netflix nunca llego a utilizar ese algoritmo por el que pago un millón de dólares, pero bueno, eso ya es otra historia.

El caso es que aquella competición fue el punto partida de Kaggle (https://www.kaggle.com/), una plataforma que permite a las empresas la posibilidad de organizar competiciones similares y ofrece a los participantes herramientas para enviar y probar sus soluciones.

Existen otras plataformas similares como TopCoder (https://www.topcoder.com/), pero claramente Klagger es la más popular, y más después de que fuera adquirida en marzo de este mismo año por Google.