Capítulo 2
Machine Learning: visión general

1. Un poco de vocabulario

A menudo se utilizan indistintamente varios términos: Inteligencia Artificial (IA) Machine Learning (ML), Deep Learning (DL), Data Science (DS). Sin embargo, no abarcan la misma realidad, por lo que es importante saber diferenciarlos.

Históricamente, la **Inteligencia Artificial** es un campo científico derivado de las matemáticas, la informática y la biología. Dos objetivos estuvieron presentes desde el principio:

– Crear una máquina inteligente capaz de resolver cualquier problema que se le plantee.

– O, más sencillamente, simular un proceso cognitivo, es decir, «dar la impresión» de que la máquina es inteligente.

El primer objetivo corresponde a la llamada inteligencia artificial «**general**» (o IA fuerte). La máquina tendría entonces emociones y capacidades de razonamiento avanzadas, lo que le permitiría «aprender a aprender». Sin embargo, aún no existe (si es que algún día existirá) fuera de los reinos de la ciencia ficción.

Observación

No hay que confundir «IA general» con «IA generalista». En el primer caso, hablamos de IA fuerte. En el segundo caso, hablamos de modelos genéricos que pueden utilizarse en distintas situaciones y no se limitan a un caso de uso concreto.

Por el contrario, el segundo objetivo corresponde a la llamada inteligencia artificial «**estrecha**», o IA débil. Es lo que existe hoy en día. Solo puede resolver problemas de uno en uno: un modelo específico corresponde a un único proceso cognitivo. No hay inteligencia real en el sentido habitual para los humanos, solo una simulación de ella.

Esto es lo que ha llevado a Luc Julia, Director Científico de Renault y cocreador de Siri, a afirmar: «La inteligencia artificial no existe». En su libro homónimo, señala que no hay inteligencia real en los modelos, que no son más que una simulación basada en estadísticas, enormes capacidades de cálculo y memorias muy grandes (en comparación con los humanos).

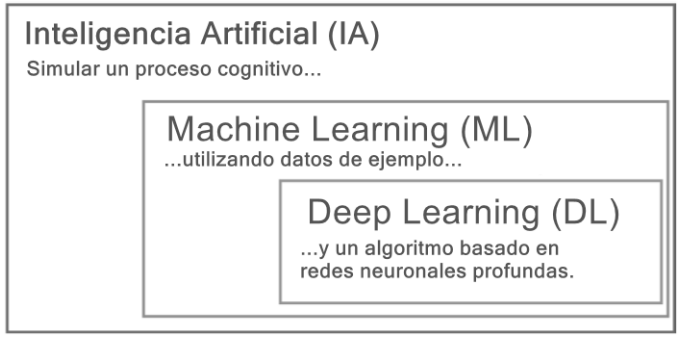
Los programas que utilizan conceptos de inteligencia artificial existen en funcionamiento y nos rodean cada día. A menudo se les denomina erróneamente «IA» o «algoritmos», pero en realidad son modelos. Las técnicas tratadas son muy amplias: desde los sistemas expertos, de moda en los años 80, pasando por la lógica difusa, también de moda en los 80 y 90, hasta los algoritmos genéticos y los algoritmos de búsqueda de rutas (utilizados en los GPS).

El **Data Mining**, por su parte, es un campo que procede principalmente de la estadística y consiste en estudiar grandes cantidades de datos para extraer nuevos conocimientos. No hay noción de «inteligencia» real o simulada, solo patrones por descubrir en los datos. Sin embargo, esta denominación prácticamente ha desaparecido y ha sido sustituida por el campo del **Machine Learning**, que abarca prácticamente las mismas técnicas.

El campo de la inteligencia artificial es vago y está mal delimitado: no existe un acuerdo global sobre su definición. Con el tiempo, se ha ido enriqueciendo con todas las técnicas y algoritmos de Data Mining, y ahora incluye el Machine Learning en su totalidad, aunque ambos eran originalmente distintos.

Las redes neuronales simulan el aprendizaje mediante la creación de neuronas simplificadas y conexiones entre ellas (IA); permiten aprender a partir de numerosos ejemplos (ML). En los últimos años, la estructura de estas redes se ha desarrollado enormemente, pasando en menos de quince años de redes de 3 o 4 capas a más de 500 en la actualidad. Estas redes «profundas» (en el sentido del número de capas) constituyen el **Deep Learning**, y forman parte del Machine Learning.

Estas tres definiciones se representan a menudo en forma de muñecas rusas: la Inteligencia Artificial es el campo de estudio que consiste en simular un proceso cognitivo. Una de las formas de hacerlo, pero no la única, es utilizar grandes cantidades de datos y aplicarles estadística: esto es Machine Learning. Por último, el Deep Learning es solo una de las técnicas que se pueden utilizar en Machine Learning.



Por último, la **Data Science** consiste en encontrar hechos interesantes en los datos, representarlos gráficamente y hacerlos utilizables (por ejemplo, para tomar decisiones). El término abarca la preparación de datos, la estadística descriptiva, el análisis y los estudios de tendencias, la visualización de datos y la creación de cuadros de mando.

Aunque no forma parte directamente de la Inteligencia Artificial, la Data Science está próxima a ella, y hay muchos puntos de solapamiento.

2. Las profesiones de los datos

Con el desarrollo del Machine Learning han surgido nuevas profesiones. Ahora hay muchas de ellas, además de empleos más tradicionales como el de director de proyectos o desarrollador. Estas permiten crear un equipo capaz de gestionar un proyecto de principio a fin.

Crear modelos es importante, pero no basta. También hay que ser capaz de desplegar y mantener aplicaciones; para ello se requieren muchas competencias diferentes.

Inicialmente, solo había un Data Scientist, pero la escala de los proyectos, el número de tareas que hay que gestionar y la proliferación de herramientas han llevado a la especialización de estos últimos, lo que a su vez ha llevado a la creación de estas nuevas profesiones, y a medida que el campo siga creciendo, se seguirán creando nuevas profesiones.

Observación

En realidad, no se trata de puestos de trabajo, sino de funciones, por lo que es posible que una misma persona desempeñe varias funciones dentro de un mismo proyecto o en los distintos proyectos en los que participe.

Estas profesiones pueden dividirse en tres ramas:

– Una rama orientada al «modelo»

– Una rama «integración»

– Una rama «soporte»

En la rama «modelo», el primero es el **Data Analyst** (*Analista de datos*). Su papel consiste principalmente en preparar y formatear datos, ya sea para producir KPI (*Key Performance Indicators* o Indicadores Clave de Rendimiento) directamente o para alimentar algoritmos de Machine Learning. También puede proporcionar cuadros de mando y tener conocimientos de visualización de datos.

El **Data Miner** (Minero de Datos) es una especialización del Data Analyst, que buscará tendencias y patrones en los datos, sin que su objetivo sea necesariamente introducir los datos en algoritmos de Machine Learning.

El segundo puesto es el de **Data Scientíst** (Cientifico de Datos). Su función es crear modelos basados en datos preparados de antemano. Este trabajo requiere sólidos conocimientos de matemáticas y estadística, así como habilidades de programación, principalmente en los lenguajes R o Python. A menudo se considera una evolución del puesto de Data Analyst, bien tras una formación especializada o gracias a su experiencia en una empresa, aunque las competencias no se solapan exactamente entre los dos puestos. También se ve como un objetivo en sí mismo, a pesar de que hay otras profesiones que pueden ajustarse mejor a los diferentes perfiles de las personas. Así que es importante entender que no es LA profesión de la Inteligencia Artificial.

El **Data Auditor** (Auditor de Datos), el último puesto de la rama «modelos», consiste en analizar el trabajo de un Data Analyst o Data Scientist para asegurarse de que lo que se ha hecho corresponde a la legislación o a necesidades específicas. Con la normativa europea aprobada el 14 de junio de 2023 (y que sin duda se aplicará a partir de 2026), este puesto se convertirá en esencial para todas las empresas que necesiten certificar sus modelos a nivel europeo.

En la rama de «integración» hay tres profesiones. Su función es permitir que los modelos interactúen con sistemas informáticos más complejos. Sus tareas van desde la recuperación de datos de entrada hasta la devolución de resultados o el reentrenamiento de modelos.

El primer puesto es el de **Data Architect** (Arquitecto de Datos). Al igual que el Solution Architect (Arquitecto de Soluciones) en las TI más tradicionales, su función es crear una arquitectura que permita que todos los elementos interactúen entre sí, pero centrándose en los flujos de datos. Dados los volúmenes implicados, a menudo se requieren conocimientos de Big Data (Spark y Hadoop se encuentran entre las habilidades más solicitadas).

El segundo puesto es el de **Data Engineer** (Ingeniero de Datos). El trabajo del ingeniero consiste en implementar la arquitectura definida por el Data Architect. Requiere buenos conocimientos de programación y automatización de procesos. El Data Architect también necesita conocer esta vertiente técnica para crear arquitecturas de alta calidad.

Observación

*El **Big Data Engineer** (Ingeniero de Macrodatos) es un ingeniero de datos que maneja grandes cantidades de datos. Los conocimientos necesarios son un poco más especializados, ya que la presencia de clústeres Spark/Hadoop complica los flujos de datos.*

El tercero es el **Data Integrator** (Integrador de datos). Su función es garantizar que los datos puedan transferirse de un sistema a otro, en el formato y con la sintaxis adecuados. Esta función requiere, por tanto, conocimientos de buses de datos, middleware y transformación de datos (ETL).

En la literatura y en los artículos se menciona cada vez con más frecuencia un puesto de trabajo: el de **Machine Learning Engineer** (Ingeniero de Machine Learning). Se trata de un ingeniero de datos que se ha formado en Data Science y Machine Learning, o un Data Scientist que se ha iniciado en la programación. Por tanto, tienen una doble cualificación y están muy solicitados.

Por último, en la rama de «apoyo», asistimos a la aparición de dos nuevas profesiones, así como de una soporte a menudo «oculta». Su función no es realizar proyectos, sino apoyar a las demás profesiones y al proyecto una vez que está en producción.

El primero es el **Data Support** (Soporte de Datos). Se trata de un servicio de asistencia con competencias adicionales en proyectos de datos, lo que le permite supervisar los modelos y actuar con rapidez cuando surge un problema. Cuanto mayor sea el número de modelos en producción, más crucial será su papel para garantizar la continuidad del servicio.

El **Data Steward** (Supervisor de Datos), por su parte, es un gestor de proyectos, también con conocimientos de datos. Dado que los proyectos de Machine Learning no pueden gestionarse del mismo modo que los proyectos de desarrollo más tradicionales, requieren una gestión de proyectos adecuada. Por tanto, su papel consistirá en discutir el proyecto con el cliente, estimar el trabajo restante, garantizar una buena comunicación entre todos los miembros del equipo y, en proyectos grandes, supervisar el progreso paralelo de las distintas disciplinas. También puede tener la función de mantener la plataforma de datos de la empresa una vez creada, incorporando los distintos requisitos y transmitiéndolos después a los Data Engineers.

Es probable que este puesto se divida en dos en un futuro próximo, dada la importancia de las tareas cubiertas.

El **Data Labeler** (Etiquetador de Datos) es una profesión muy extendida pero de la que se habla poco, que también puede incluirse en la rama de «soporte». Su objetivo es etiquetar los datos de entrenamiento para futuros modelos. Muy a menudo, como esta tarea requiere pocas competencias y aporta poco valor añadido, se delega en países donde la mano de obra es más barata. Varios escándalos han salido a la luz, porque las grandes empresas de IA recurren a menudo a Data Labelers a los que se les paga apenas unos céntimos por cada dato etiquetado. Es el caso de OpenAI, que utilizó mano de obra keniana para etiquetar datos para ChatGPT. Debido a la presión social, esta profesión seguramente tendrá que evolucionar, o al menos regularse.

Las líneas de negocio actuales son las siguientes:

La mayoría de estas profesiones no existían hace apenas unos años, y seguramente surgirán muchas más próximamente, lo que modificará automáticamente el ámbito de cada una de ellas.

Del mismo modo, como las tecnologías evolucionan muy rápidamente, las personas que trabajan en estas distintas profesiones deben mantenerse al día y recibir formación constante, pues de lo contrario podrían encontrarse rápidamente con competencias obsoletas.

3. El crecimiento del Machine Learning

El Machine Learning es un campo en pleno crecimiento desde hace varios años, sin duda gracias a una serie de éxitos notables. Sin embargo, los primeros trabajos sobre el tema se remontan a varias décadas atrás. Se pueden esgrimir varios argumentos para explicar esta explosión actual.

En primer lugar, la **potencia de los ordenadores** ha aumentado exponencialmente. Con la llegada de la computación en nube, ya no es necesario comprar hardware caro. Basta con alquilarlo durante el tiempo que se utilice, lo que significa que todo el mundo puede tener acceso a máquinas impresionantes. También hay que tener en cuenta que, para determinados tipos de cálculo, en particular el Deep Learning, las GPU (en la tarjeta gráfica) han superado a las CPU (en el procesador), multiplicando la potencia disponible.

Los cálculos son más rápidos, pero el Machine Learning funciona con datos. Cuanto más complejo es el problema, mayor es el volumen de datos necesario. Esta es la razón por la que a menudo se vincula el Big Data con el Machine Learning: las fuentes de datos se multiplican (por ejemplo, con la llegada de los objetos conectados) y se pueden almacenar, gestionar y utilizar. En resumen, los grandes **volúmenes de datos** necesarios están ahora disponibles y son más fáciles de gestionar.

Más potencia, más datos, pero eso no lo explica todo. Los resultados de los **algoritmos** no han dejado de mejorar en los últimos años, gracias a la investigación alentada por el éxito. También hay que señalar que la investigación ha salido del ámbito de la investigación universitaria, ya que GAFAM y sus equivalentes chinos BATX cuentan con laboratorios de investigación privados, cuya importante financiación se destina a la innovación en Inteligencia Artificial.

Los algoritmos son más potentes y rápidos, por lo que pueden procesar aún más datos.

Al principio, cada uno tenía que codificar su propio algoritmo, y a menudo resultaba difícil compartir código entre investigadores. Pero con la llegada del Machine Learning, empezaron a aparecer **frameworks** (Bibliotecas o Librerías en castellano). Éstos ponen los algoritmos a disposición de quien los quiera, sin tener que leer complejos artículos científicos ni entender las ecuaciones subyacentes. La mayoría de estos frameworks son gratuitos y de código abierto, lo que facilita la cooperación entre investigadores y la mejora de las implementaciones propuestas.

Observación

Sin embargo, el crecimiento actual del Machine Learning no debe hacernos creer que se trata de una herramienta fácil de usar que garantiza el éxito. Muchos proyectos nunca pasan de la fase PoC (Proof of Concept; Prueba de concepto) y nunca llegan a producción porque no ha sido posible obtener modelos con un rendimiento suficiente. Cualquiera que trabaje en Machine Learning debe ser consciente de ello.

El último aspecto que explica el auge actual es la importancia de los medios de **comunicación**. De hecho, el primer boom mediático tuvo lugar después de 2012, cuando los algoritmos de clasificación de imágenes mostraron resultados que permitieron ponerlos en producción (por ejemplo, para la detección de defectos o el procesamiento de imágenes en tiempo real para coches autónomos). Un segundo boom mediático, aún mayor, tuvo lugar a finales de 2022 con la llegada de la IA generativa (liderada por ChatGPT de OpenAI), que permitió al gran público descubrir los avances en este campo. Gracias a ello, las empresas disponen de más recursos para desarrollar nuevos algoritmos o utilizarlos.

4. Formas de aprendizaje y tareas de ML

El Machine Learning consiste en crear un modelo que es el resultado de un aprendizaje (o entrenamiento). Un «modelo» es, por tanto, un programa informático no ha sido creado por un desarrollador, sino mediante un algoritmo de aprendizaje.

Existen varias formas de aprendizaje definidas por los datos de entrada (o variables explicativas) y los datos de salida (o variables deseadas).