Telefónica | EDUCACIÓN DIGITAL

Módulo 1:

Introducción al Machine Learning

Módulo 1

Introducción al Machine Learning

1.1. ¿Qué es el Machine Learning?

El aprendizaje automático[1] o Machine Learning se engloba dentro de las disciplinas de la Inteligencia Artificial[2]. Es un método científico que nos permite usar los ordenadores y otros dispositivos con capacidad computacional para que aprendan a extraer los patrones y relaciones que hay en nuestros datos por sí solos. Esos patrones se pueden usar luego para predecir comportamientos y en la toma de decisiones.

Hasta la llegada del Machine Learning, la automatización del análisis en la toma de decisiones requería siempre de un experto humano capaz de descubrir algunas reglas más o menos ajustadas. El experto, basándose en los datos, intentaba descubrir cuáles eran los distintos patrones que nos permitían resolver el problema. Formulaba un conjunto de reglas más o menos exactas y dichas reglas eran específicamente programadas por ingenieros de software. El conjunto resultaba en un modelo más o menos próximo a la realidad, que se usaba para estimar lo que iba a ocurrir en un nuevo caso o en el futuro y en eso se basaba la toma de decisiones. Este sistema de toma de decisiones presenta varios problemas:

- **escasez de expertos**: A menudo nos puede ser difícil encontrar expertos formados en el domínio de nuestro problema.
- **desarrollo automatizado**: Una vez los expertos han detectado las reglas, éstas deben ser automatizadas e integradas en nuestros sistemas para poder aplicarlas a grandes volúmenes de datos automáticamente. Dicho trabajo involucra a ingenieros de software, que deben entender y programar dichas reglas. En ocasiones, la transmisión de información entre el experto y el ingeniero de software da lugar a errores y requiere un tiempo considerable de implementación.
- **renovación de los modelos**: La producción de una solución de este tipo acostumbra a ser lenta y costosa, y los modelos son difíciles de adaptar cuando los datos evolucionan y sus reglas cambian.
- **escalabilidad**: El volumen y la complejidad de los problemas que podemos solucionar con este método es limitado.

A medida que el volumen y la complejidad de los datos han ido creciendo, se ha hecho absolutamente necesario el uso del Machine Learning. Incluso en ficheros con pocos datos, cuando el número de propiedades que se manejan empieza a crecer nos es difícil detectar los posibles patrones que los relacionan. En la Figura 1.1 vemos uno de los ejemplos canónicos que se utilizan para describir cómo hacer aprendizaje automático. El fichero describe algunas propiedades de un tipo de flores, el *iris*, y el objetivo del problema es predecir la especie a la que pertenece cada ejemplar según sus características. Como se puede ver, a pesar de que el número de campos no es muy elevado, a simple vista es muy difícil detectar las reglas que determinan dicha clasificación.

[1] https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico
[2] https://es.wikipedia.org/wiki/Inteligencia_artificial

longitud sépalo	anchura sépalo	longitud pétalo	anchura pétalo	especie
5.1	3.5	1.4	0.2	Iris-setosa
6.5	3.0	5.5	1.8	Iris-virginica
4.7	3.2	1.3	0.2	Iris-setosa
5.8	2.8	5.1	2.4	Iris-virginica
5.2	2.7	3.9	1.4	Iris-versicolor
5.4	3.9	1.7	0.4	Iris-setosa
5.7	2.8	4.5	1.3	Iris-versicolor
6.3	3.3	4.7	1.6	Iris-versicolor
4.9	2.4	3.3	1.0	Iris-versicolor
6.0	2.2	5.0	1.5	Iris-virginica
6.9	3.2	5.7	2.3	Iris-virginica
5.0	2.0	3.5	1.0	Iris-versicolor

Figura 1.1: Estructura de datos correspondiente a un ejemplo canónico de aprendizaje automatizado. Cada fila corresponde a un ejemplar de **iris** y sus las columnas contienen las medidas de sépalo y pétalo y la especie a la que pertenece el ejemplar. Existen reglas muy claras que nos permiten catalogar dichas flores en su especie a partir de sus medidas, pero para los humanos es difícil detectar dichas relaciones, aún cuando el número de propiedades involucradas no es muy alto, como en este caso.

Cuando la complejidad aumenta, los ordenadores superan en capacidad de análisis a los expertos humanos y evitan la necesidad de programar explícitamente soluciones a medida. De esta forma, hemos pasado de la toma de decisiones dirigida por expertos a la toma de decisiones dirigida por datos.

Los ingredientes imprescindibles para cualquier solución de Machine Learning son dos: los datos y los modelos de Machine Learning. Por eso, en el *Módulo 2* hablaremos de cómo debemos preparar los datos para poderlos usar en el Machine Learning. En el *Módulo 3* presentaremos un ejemplo de modelo de Machine Learning para la clasificación y regresión. En el *Módulo 4* trataremos otros tipos de problemas no supervisados y discutiremos los algoritmos[3] que usan para aprender. Veremos casos prácticos de cada tipo de aprendizaje y descubriremos qué modelo se ajusta mejor a cada problema y qué nuevas informaciones nos proporciona.

Por supuesto, el contenido de este curso no abarca el proceso completo necesario para integrar un sistema de aprendizaje automático en nuestra empresa. Aún así, intentaremos dar las herramientas suficientes para que podáis entender cómo funcionan algunas soluciones de Machine Learning y en qué problemas nos pueden ser de gran ayuda.

Dado que el aprendizaje automático persigue lograr que las máquinas consigan extraer patrones de los datos, el primer paso será proporcionar dichos datos en un formato que las máquinas puedan usar para aprender de ellos. En el siguiente módulo explicaremos cómo conseguirlo.

[3] https://es.wikipedia.org/wiki/Algoritmo#Tipos_de_algoritmos_seg.C3.BAn_su_funci.C3.B3n