Introducción © EDICIONES ROBLE, S.L.

Indice

Introducción

Introducción. Daniel Rodríguez Pérez

El aprendizaje automático —o "aprendizaje máquina", del inglés *machine learning*— es una rama de la Inteligencia Artificial en la que se recogen las diferentes técnicas para dotar a un ordenador de la capacidad de "aprender" patrones a partir de conjuntos de datos de ejemplo. Para ello, es necesario crear modelos que sean una abstracción de los datos. Estos modelos pueden ser de diferentes tipos: fórmulas matemáticas, conjuntos de reglas o estructuras de conexiones. Una vez obtenidos, los modelos pueden ser utilizados para aumentar la eficacia con la que se resuelven múltiples problemas ya conocidos o para resolver nuevos. Se enumeran, a continuación, algunas de las principales aplicaciones de los modelos de aprendizaje automático.

Realizar predicciones

Por ejemplo, se pueden identificar los clientes que pueden abandonar próximamente una compañía o aquellos que podrían estar interesados en adquirir un producto o servicio concreto, así como anticipar variaciones de la demanda, tanto los incrementos como las disminuciones, para poder adaptar con antelación los procesos productivos a las necesidades del mercado en cada momento.

Reconocer patrones de interés

Por ejemplo, identificar los productos o servicios que son adquiridos de forma conjunta por los clientes para optimizar la presentación de los mismos, tanto en el canal físico como en el on-line.

Buscar grupos en los datos

Obtener las diferentes tipologías de clientes que existen para poder focalizar campañas de marketing, fidelización o para usar como guía en la creación de nuevos productos.

Actualmente, el aprendizaje automático es una de las principales tendencias que existen dentro de las tecnologías de la información. Con el continuo aumento de los sistemas Big Data, y gracias al abaratamiento de los costes de almacenamiento y captura de datos, se hace necesario disponer de algoritmos de aprendizaje automático que permitan descubrir la información oculta en estos sistemas para poder ponerla en valor. Así, tanto para las empresas tecnológicas como para muchas otras, cuyas principales líneas de negocio se encuentran completamente alejadas de la tecnología, estos algoritmos resultan ser sus activos más valiosos, incluso por delante de activos físicos, que son más fácilmente reemplazables o imitables por los competidores, ya que permiten profundizar en el conocimiento de los clientes, lo que posibilita ofrecerles los productos o servicios más adecuados en cada momento, y mejoran los procesos internos, dotándolos de mayor eficiencia. Entre las aplicaciones de las técnicas de aprendizaje automático, se pueden enumerar:

- Motores de búsqueda.
- Detección del fraude.
- Análisis del mercado de valores.

- Detección de intrusiones.
- Reconocimiento del habla y de textos escritos a mano, etc.

El principal objetivo de este módulo es ofrecer una visión general de las principales técnicas y algoritmos utilizados en la actualidad dentro del campo del aprendizaje automático. Para la parte práctica del módulo se utilizará la librería de Python Scikit-Learn, una de las más populares en la actualidad, debido a la cantidad de técnicas que implementa y la facilidad de su manejo. Por esto, se asume que el alumno parte con experiencia en el entorno IPython y que cuenta con conocimientos acerca del uso de herramientas de análisis estadístico y técnicas de limpieza y transformación de datos.

El módulo comienza con una introducción al aprendizaje automático en el que se revisará su importancia en el procesado de minería de datos. Al mismo tiempo, se realizará un repaso de los principales tipos de aprendizajes existentes y se pondrá en marcha un primer modelo sencillo. El propósito de esta primera unidad es dar a conocer al alumno la importancia de las técnicas estudiadas en los entornos de datos para obtener valor e identificar los principales estudios que se pueden realizar. Al mismo tiempo, el alumno podrá ver la implementación de un primer modelo básico.

La segunda unidad estudia los modelos en los que el conjunto de datos utilizados contiene una variable que ha de ser posteriormente reproducida, la cual no estará disponible cuando el modelo sea utilizado. Este tipo de modelos se conocen con el nombre de modelos supervisados. Se hará hincapié en la regresión lineal para la predicción de valores continuos y la regresión logística y los árboles de decisión para la identificación de categorías.

La tercera unidad analiza la otra gran familia de modelos: los modelos no supervisados. En esta ocasión, el conjunto de datos utilizado para el entrenamiento no contiene ninguna variable que deba ser reproducida posteriormente. Las herramientas utilizadas serán el algoritmo k-means, para la identificación de objetos similares, los clústeres jerárquicos, para la organización de clústeres, y DBSCAN, para la identificación de regiones en los conjuntos de datos.

La cuarta unidad presentará las principales técnicas utilizadas para la selección de las características de los datos más adecuadas para la creación de los modelos estudiados en las secciones anteriores. Además, se estudiará la forma de seleccionar los modelos más adecuados en cada situación. Las técnicas descritas en esta unidad son necesarias debido a que, en los conjuntos de datos disponibles en entornos reales, junto a la información relevante para la creación de los modelos, existe otra que no lo es o se encuentra duplicada. Saber identificar la información relevante permite mejorar la calidad de los modelos construidos. La selección del mejor modelo tampoco suele ser una tarea trivial, ya que algunos modelos pueden memorizar los datos utilizados para el entrenamiento, generando lo que se conoce como “sobreajuste”, lo que provoca que el rendimiento observado durante el entrenamiento no se dé en conjuntos de datos diferentes al original.

La quinta unidad analizará los modelos conexionistas, en donde el conocimiento emerge de redes formadas por unidades sencillas interconectadas. La idea que hay detrás de estos modelos es reproducir la forma en la que funciona el cerebro, basándose en el principio de que los fenómenos mentales pueden ser descritos mediante redes de unidades sencillas, como son las neuronas. Los modelos más conocidos de conexionismo son las redes neuronales, que serán estudiados en esta unidad junto a algunas de sus principales aplicaciones.

Finalmente, en la sexta unidad se analizarán las reglas de asociación y el *market basket analysis*. Estas técnicas son utilizadas para identificar los ítems que aparecen de forma conjunta en muestras de datos. Una de las principales aplicaciones es la identificación de los productos que se agregan de forma conjunta en las cestas de la compra de las tiendas.

Los objetivos generales del módulo que los alumnos alcanzarán tras su estudio se pueden resumir en los siguientes:

1. Entender el papel del aprendizaje automático dentro de los procesos de minería de datos y cómo se utiliza para descubrir patrones en bases de datos y en la construcción de modelos predictivos.
2. Entender la diferencia entre aprendizaje supervisado y no supervisado y saber en qué situaciones se aplican.
3. Entender el proceso de entrenamiento y de evaluación de los algoritmos y el papel de las técnicas de validación cruzada.
4. Ser capaces de seleccionar familias de algoritmos y algoritmos concretos de acuerdo con las características del problema.
5. Saber utilizar bibliotecas de algoritmos de aprendizaje automático en entornos de *Data Science*.
6. Saber medir la capacidad predictiva de un modelo e identificar aquellos que han de ser reentrenados o sustituidos por otros.