Machine learning: técnicas básicas

El objetivo de esta lectura es conocer las características de los métodos de aprendizaje supervisado y no supervisado asociados a machine learning, profundizando en las fases y en los principales algoritmos utilizados en su desarrollo.

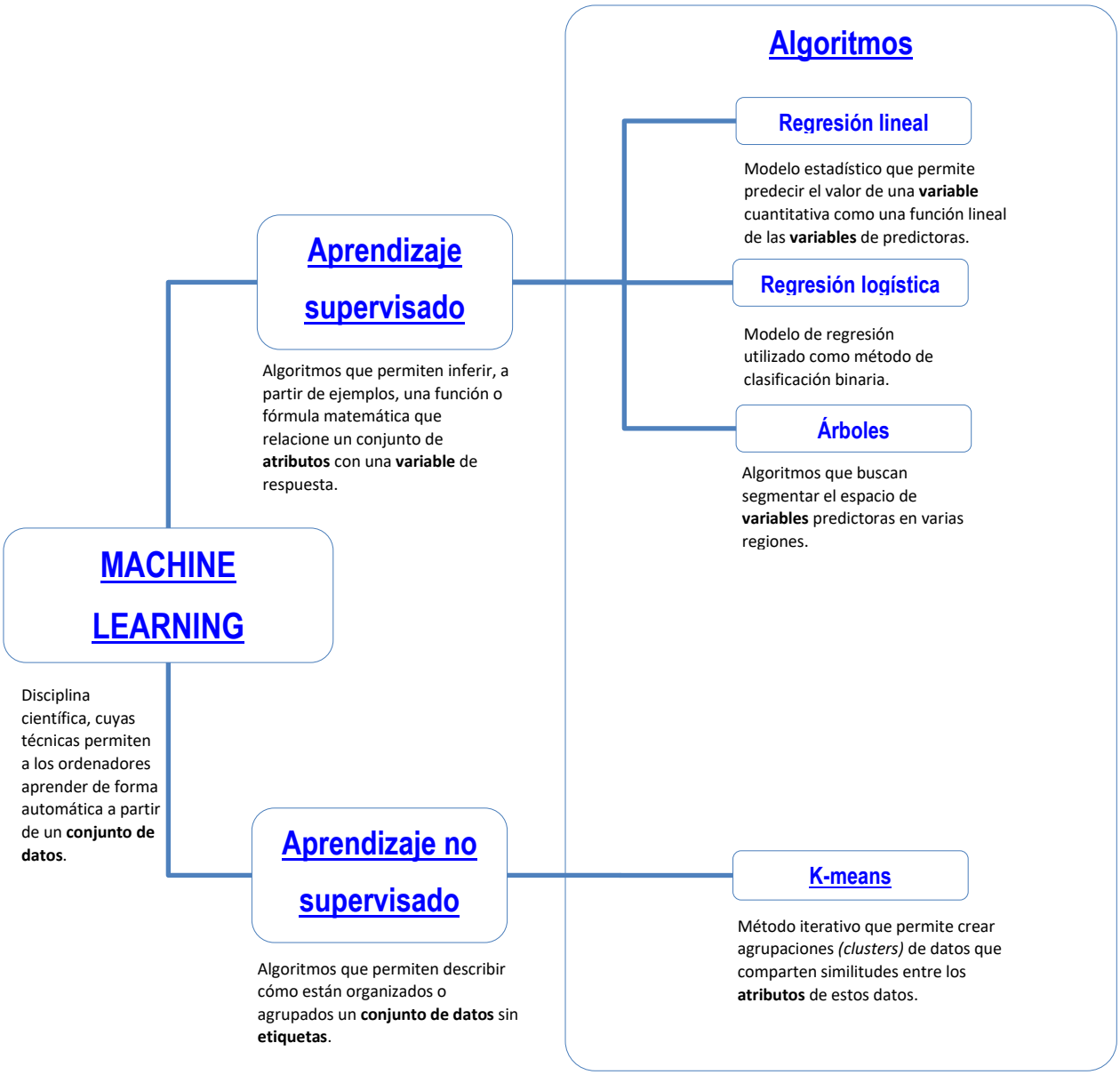
Introducción y métodos de aprendizaje automático

El *machine learning* es una disciplina científica cuyas técnicas permiten a los ordenadores aprender de forma automática a partir de un **conjunto de datos**, de tal forma que seamos capaces de hacer predicciones sobre un proceso o describirlo de forma compacta.

Dentro de esta disciplina científica existen diferentes tipos de aprendizaje automático, que estructuran las diferentes técnicas o algoritmos en función del método de aprendizaje en el que se basan. El siguiente esquema nos aporta una visión a alto nivel de lo que veremos durante esta lectura.

Mapa de la lectura

NOTA: haz clic sobre los conceptos en el mapa para navegar la lectura. Puedes volver a este mapa cuando quieras, haciendo clic sobre el texto “Ir a ‘Mapa de la lectura’”, ubicado al final de cada página del documento.



Aprendizaje supervisado

A pesar de lo que sugiere su nombre, el aprendizaje supervisado no consiste en la intervención humana para la supervisión del proceso, sino en inferir, a partir de **ejemplos**, una función o fórmula matemática que relacione un conjunto de **atributos** con una **variable de respuesta**. El objetivo es predecir (generalizar) la **respuesta** ante futuras observaciones de los **atributos**.

Un ejemplo de aprendizaje supervisado puede ser un algoritmo que aprende a predecir la edad de una persona (**variable respuesta**) a partir de sus gustos, horarios, nivel de estudios, nivel económico, etc. (**variables predictoras**).

Otro ejemplo puede ser un detector de *spam* (correo electrónico no solicitado que se envía de forma masiva con fines publicitarios o comerciales). El detector es un algoritmo, que analiza el historial de mensajes (**ejemplos**) en una bandeja de correo electrónico y las "partes" que tienen estos mensajes (remitente, destinatario, asunto), de acuerdo con parámetros de entrada que uno defina (quién es el remitente, si el destinatario es individual o parte de una lista, si el asunto contiene determinados términos, etc.). Según este análisis, el detector establece una función o fórmula matemática, que le permite **"etiquetar"** (clasificar) cada uno de los mensajes. En este caso, el detector asigna una **etiqueta** a cada uno de los mensajes en el historial; que puede ser "es spam" o "no es spam".

Una vez definida esta función, al introducir un nuevo mensaje no etiquetado en la bandeja de entrada del correo electrónico, el algoritmo es capaz de asignarle la **etiqueta** correcta.

Fases del aprendizaje supervisado

El aprendizaje supervisado se caracteriza por estar dividido en dos fases principales: el "*entrenamiento* y *validación (o test)*", etapa en la que se usan datos históricos, de los que se conoce la **variable respuesta** (por ejemplo, correos electrónicos que ya sabemos si son "spam" o "no spam"); y la "*predicción*", etapa en la que usamos datos nuevos de los que se desconoce la **respuesta** a priori, es decir, **datos no etiquetados**

(por ejemplo, correos electrónicos que aún no han sido clasificados como “spam” o “no spam”).

A continuación, analizaremos cada una de las fases del aprendizaje supervisado:

1. Fase de entrenamiento *y test*.

Se basa en aprender de los datos, de los **ejemplos**, con el objetivo de conseguir un **modelo** que haya captado las relaciones entre las **variables predictoras** y la **variable respuesta**; con el fin de dar un resultado cuando lleguen nuevos datos de los que no se conoce su valor a predecir. Por ejemplo, partiendo de cuantas semanas duran los embarazos y los datos de edad de la madre, peso, alimentación, nivel de actividad física, seguimiento médico, histórico de salud etc., intentar predecir cuantas semanas durará un nuevo embarazo.

El proceso que se sigue en el aprendizaje supervisado es el siguiente:

a) Se parte de los datos en crudo y, sobre ellos, se realizan las tareas de procesamiento de datos hasta conseguir un formato y estructura adecuada para nuestro análisis. Por ejemplo, nos puede interesar tener datos agregados de semana en semana u obtener nuevas columnas de datos generadas a partir de los datos iniciales, como puede ser el cambio de peso de una semana a otra, o la diferencia entre la actividad física inicial y final. Esto conlleva limpieza de datos, unión de diferentes fuentes de información y generación de **variables** que puedan ser útiles en nuestro problema. Siguiendo con nuestro ejemplo, tendríamos que unir la información médica con la actividad física o con la dieta seguida por la madre.

b) A continuación, se realiza una división de nuestros datos, lo que se conoce como datos de entrenamiento y datos de *test*. El primer **conjunto de datos** de entrenamiento o *training* ***data set*** se compondrá del grueso de los datos, aproximadamente es el 70%. En el ejemplo, se tomarían el 70% de los datos de las mujeres que ya han dado a luz y se utilizarían como conjunto de entrenamiento. El conjunto restante, que es aproximadamente un 30%, se denomina conjunto de test o *test* ***data set*** y servirá para determinar si el

modelo entrenado es suficientemente bueno como para predecir futuros valores. Para ello, en el ejemplo se tomarían los datos del 30% de las mujeres restantes que han dado a luz.

c) Cuando tenemos los **conjuntos de datos** preparados, se inicia un proceso iterativo de **entrenamiento del modelo**, en el que se tiene en cuenta las **variables predictoras** y la **variable respuesta** junto a su **etiqueta**. El algoritmo de aprendizaje estudia automáticamente los datos de entrenamiento y genera un **modelo** que intenta aprender los patrones que siguen las **variables predictoras** en función de la **etiqueta**.

d) Cuando se acaba este paso, se procede a realizar una validación del **modelo** con el conjunto de *test*. Este proceso consiste en el tratar determinar la **etiqueta** a partir de las **variables predictoras** y comparar el resultado obtenido con el **modelo** con el dato real para ver los aciertos y fallos del **modelo**. Es decir, en el ejemplo, nuestro **modelo** nos proporcionaría como resultado el número de semanas que cree que duraría el embarazo y estos resultados se deben comparar con el número de semanas reales que ha durado el embarazo de este 30% de las mujeres que se han seleccionado como conjunto de *test*.

Si consideramos que el **modelo** no ha acertado lo suficiente, podemos volver a realizar este proceso de entrenamiento hasta que consigamos que sea capaz de predecir como lo necesitamos. Supongamos que los resultados del **modelo** de tiempo de gestación, nos proporciona valores muy inferiores o muy superiores a las 40 semanas en todos los casos, entonces debemos desechar dicho **modelo** y considerar otras **variables predictoras** que nos puedan dar mejores resultados. O bien, si nuestro **modelo** se acerca mucho a los resultados de los tiempos de gestación reales (por ejemplo, una mujer ha tenido un embarazo de 38 semanas y se ha predicho 39 semanas), pero consideramos que no es lo suficientemente bueno, también puede que debamos considerar volver a entrenar el **modelo**.

Cuando hemos terminado con la fase de entrenamiento y test, y estamos satisfechos con el **modelo**; se pasa a la segunda fase de predicción.

2. Fase de predicción

En esta etapa, se cuenta con nuevos datos de los que no conocemos la **respuesta**, es decir, solo tenemos las **variables predictoras**. Estos datos se procesarán para darles el mismo formato que al conjunto de entrenamiento y *test*, y se utilizarán como datos de entrada al **modelo** que hemos entrenado. Con ellos, el **modelo** intentará determinar qué patrón de los que ha aprendido para predecir es el que más se ajusta a la **etiqueta** que vamos a dar como respuesta.

Por ejemplo, cuando una nueva mujer se queda embarazada, se han de recoger los datos correspondientes a las **variables** de peso, edad, actividad física, historial médico etc., para que el **modelo** nos prediga el número de semanas de gestación.

Dentro de este método de aprendizaje, podemos encontrar algoritmos o **modelos** analíticos como la **regresión lineal**, la **regresión logística** o los **árboles de decisión**, que se explicaran más adelante en la lectura.

Aprendizaje no supervisado

Son algoritmos que permiten describir cómo están organizados o agrupados un **conjunto de datos no etiquetados**. Por lo tanto, el aprendizaje no supervisado tiene lugar cuando no se dispone de **datos "etiquetados"** para el entrenamiento, es decir, sólo conocemos los datos de entrada, pero no existen datos de salida que corresponden a esos datos de entrada. El objetivo de este método es describir la estructura de los datos, para intentar encontrar algún tipo de organización que simplifique el análisis.

Los algoritmos y estrategias para abordar problemas no supervisados se centran en describir cómo están organizados o cómo se podrían agrupar ese **conjunto de datos no etiquetados**. Por lo tanto, el objetivo del aprendizaje no supervisado es conseguir agrupaciones de datos no detectables a simple vista, en base a las características, a las **variables**, que describen cada uno de los **ejemplos** de la muestra. Por ejemplo, si queremos segmentar un conjunto de clientes en base a un conjunto de **variables** (edad, sexo, gustos, etc.) para identificar grupos similares que serán tratados de maneras distintas por una nueva campaña de *marketing*. En este caso, como **variables predictoras**, se conocen los datos de edad, sexo y gustos, y no se conoce el grupo de clasificación en el que se encuadra el cliente (esta **variable de respuesta** es la que se intenta determinar), por lo que los datos de cliente son **datos no etiquetados**.

Otro aspecto que caracteriza a los problemas no supervisados es que la forma de evaluar o validar un **modelo** no se puede realizar de forma tan exhaustiva como para el caso del aprendizaje supervisado, ya que no se dispone de la **etiqueta** con la que resultaría sencillo comparar los resultados del **modelo** con las **etiquetas** reales. Entonces, la forma de evaluar será más dependiente de la interpretación que se haga de los resultados y su utilidad. En cualquier caso, también existen **métricas** que miden, desde un punto de vista matemático, cómo de bien o mal ha sido capaz el **modelo** de separar los datos en base a sus propiedades (por ejemplo, midiendo la semejanza entre los elementos de un mismo grupo de clientes).

Fases del aprendizaje no supervisado

En el aprendizaje no supervisado se distinguen dos etapas principales: por un lado, la etapa de entrenamiento o training y, por otro, la de transformación o aplicación del **modelo** entrenado aprendido.

1. Fase de entrenamiento *y test*

En la primera etapa, el proceso de entrenamiento comienza con los datos en crudo disponibles para resolver el problema. Lo primero que se lleva a cabo es la fase preprocesamiento de esos datos, con lo que quedaría preparada la matriz de **atributos** con todos los **ejemplos** limpios para la fase de aprendizaje, en la que se realizará la construcción del **modelo** utilizando el algoritmo que mejor se adapte a nuestro objetivo. Una vez aprendido el **modelo**, este será evaluado y validado.

Por ejemplo, se disponen de datos de la antigüedad de construcción de edificaciones en una ciudad, el tipo de construcción (residencial, oficinas, local comercial, sin edificar, etc.), la densidad de población y coste por unidad de superficie. A partir de estos datos, se intentarán crear agrupaciones de zonas urbanas con características similares para ayudar a las autoridades locales en la planificación de la ciudad. A diferencia del aprendizaje supervisado, no se conoce con antelación cuáles serán esas agrupaciones, ni cuáles son las características comunes que las caracterizan.

2. Fase de transformación

La fase de transformación tiene el objetivo de aplicar el **modelo** aprendido en el momento en el que llegan nuevos datos. El preprocesamiento de los datos debe ser el mismo que en la fase de *training*, de tal modo que se realicen las transformaciones necesarias para aplicar el **modelo** aprendido en la fase anterior y obtener el resultado final.

Por ejemplo, si para la elaboración del **modelo** para la clasificación de zonas urbanas que se indicaba en la fase de *training*, se han realizado labores de limpieza y normalización de datos y se han generado **variables** derivadas (como, por ejemplo,

el coste por habitante de la zona, a partir de los datos de coste por unidad de superficie y densidad de población); estas mismas transformaciones deberán aplicarse en el caso de que la ciudad se expanda y se configuren nuevas zonas urbanas a clasificar.

Los algoritomos mas conocidos dentro de este metodo de aprendizaje son los de *clustering*, como es el caso de *K means*, que se explicará más adelante en la lectura.

Algoritmos

Hemos visto en el apartado anterior los principales métodos de aprendizaje supervisado y no supervisado utilizados en la disciplina de *machine learning*. En este apartado, vamos a profundizar en la explicación de las principales técnicas o algoritmos que podemos encontrar dentro de cada uno de estos métodos.

> Un algoritmo es un conjunto de reglas que, aplicadas sistemáticamente a unos datos de entrada, resuelven un problema en un número finito de pasos.

Partiendo de esta definición y siguiendo la diferenciación entre algoritmos de aprendizaje supervisado o no supervisado pasamos a conocer en detalle algunos de los algoritmos más utilizados en *machine learning*.

Regresión lineal

La **regresión lineal** se encuentra dentro de los algoritmos de aprendizaje supervisado y se trata de un **modelo** estadístico**,** que permite predecir el valor de una **variable** cuantitativa (es decir, una **variable numérica**), como una función lineal de las **variables de entrada**, también conocidas como **variables predictoras**.

Un ejemplo de **regresión lineal** podría ser el análisis de las ventas de un nuevo modelo de teléfono que acaba de lanzar un fabricante de dispositivos. Para poder realizar esta predicción, se requiere de un **modelo** analítico, que permita identificar la relación en el tiempo entre la **variable de respuesta** (volumen de ventas) y las diferentes **variables predictoras**, es decir, las diferentes **variables** que pueden explicar este volumen de ventas (cuota de mercado, número de unidades producidas,

gasto en campañas de comunicación o volumen de ventas de la competencia entre otras).

Es importante destacar que para que el método de **regresión lineal** sea válido, tanto la **variable** a predecir, como las **variables predictoras** deben ser cuantitativas.

Una vez identificada la función que incluye los parámetros que definen las ventas, y que permiten establecer una fórmula matemática para las ventas del modelo de teléfono, se puede llevar a cabo **modelos** que permitan predecir las ventas en un futuro próximo.

Regresión logística

La **regresión logística**, como la **regresión lineal**, se encuentra dentro de los algoritmos de aprendizaje supervisado y se trata de un **modelo** de regresión utilizado como método de clasificación binaria (nos ofrece una respuesta con dos valores únicamente: sí/no, hombre/mujer, 0/1, etc.), puesto que, en lugar de valores numéricos, éste permite estimar la probabilidad de que ocurra (o no) un evento como función de otro tipo de **variables**.

Por ejemplo, puede utilizarse para determinar el sexo de una persona a partir de otras **variables** explicativas.

En este ejemplo, el algoritmo estudia y aprende sobre un histórico de datos en el que se tiene el peso y la edad (**variables** explicativas) de muchas personas y si son hombre o mujer (**variable de respuesta** binaria). Una vez el **modelo** ha encontrado la relación entre las **variables predictoras** y la **variable de respuesta**, puede para nuevas personas en las que se conoce el peso y la edad predecir si son hombre o mujer.

En lugar de utilizar la función de **regresión lineal**, este **modelo** se basa en una función sigmoide o logística (con forma de S) que nos permite establecer un umbral de decisión que separe las dos clases.

Métodos basados en árboles

Los **métodos basados en árboles**, también encuadrados bajo el tipo de aprendizaje supervisado, consisten en algoritmos que buscan segmentar el espacio de **variables predictoras** en varias regiones, por lo que se establecen tramos diferentes donde la fórmula matemática que predice el comportamiento de los datos varía para adecuarse a los datos pertenecientes a ese tramo. Dentro de cada región, se utiliza la media (valor medio o promedio) o la moda (valor más repetido) de las observaciones de entrenamiento (valores medidos) en esa región para hacer la predicción.

El método más sencillo es el árbol de decisión básico, aunque existen otros métodos, más avanzados, que mejoran la precisión de este **modelo** (por ejemplo, *bagging, random forest* y *boosting*).

Para entender cómo funciona un árbol de decisión, observemos el siguiente ejemplo. Se trata de predecir el salario (**variable** no conocida) de un jugador de fútbol en función de la cantidad de años que lleva jugando en la liga y la cantidad de goles que ha marcado en las temporadas anteriores (**variables** conocidas).

Para ello contamos con un ***data set*** de entrenamiento en el que conocemos el salario, los años jugados y los goles marcados. Utilizando un árbol de decisión, obtenemos un conjunto de reglas que nos permite predecir el salario de un futbolista para el que únicamente tenemos datos de los años jugados y los goles marcados.

Es decir, siguiendo este ejemplo, el FC Barcelona podría estimar el salario que cobra Cristiano Ronaldo (jugador del Juventus FC) en función de los goles y años jugados, para ello partiría de la información del salario, años jugados y goles marcados por los jugadores de su plantilla.

K-means

Finalmente, como algoritmo encuadrado dentro del aprendizaje no supervisado y clasificado como una técnica de *clustering*, el ***K-means*** es un método iterativo que permite crear agrupaciones (*clusters*) de datos que comparten similitudes entre los **atributos** de estos datos.

En términos generales, el algoritmo de ***K-means*** asigna una agrupación concreta a una observación (o dato de entrada) de forma aleatoria y, a continuación, realiza iteraciones sobre la forma de las agrupaciones hasta que se determina la posición real de esa observación en el mapa gráfico de agrupaciones.

Un ejemplo en el que se podría utilizar este tipo de algoritmo sería, el caso en el que una empresa de distribución de contenidos televisivos desea establecer diferentes paquetes de canales en función de los intereses de sus clientes. En este caso la utilización de ***K-means*** ayudaría a identificar los grupos de personas que tienen intereses parecidos.

Conclusiones

Tras haber leído esta lectura, obtenemos una visión general de los métodos de aprendizaje de *machine learning,* así como un esquema teórico de las distintas técnicas o algoritmos comúnmente utilizados para cada tipo de aprendizaje automático.

A modo de resumen de lo que es el aprendizaje supervisado, tenemos que tener en cuenta que, para realizar este tipo de **modelos**, hemos de contar con un histórico que nos permitirá aprender los patrones de las **variables predictoras** para predecir una **variable respuesta** que en el histórico es conocida. Se realiza una comparativa de como predice el **modelo** frente a la realidad y, cuando se considera que el **modelo** es suficientemente bueno, se pasa a la fase de predicción en la cual solo se obtienen las **variables predictoras** y se da una **respuesta** de la que no se conoce a priori si es correcto o no. Como algoritmos principales, se consideran **regresión lineal, regresión logística** y los **métodos basados en árboles**.

Desde el punto de vista del aprendizaje no supervisado es importante recordar que el objetivo principal de este tipo de técnicas es “organizar objetos en grupos según su distribución o similitud”. Cada individuo o **ejemplo** de nuestro ***data set*** estará caracterizado por un conjunto de **atributos** que lo definen y, como salida del **modelo**, obtendremos una agrupación de dichos **ejemplos**. Como algoritmo representativo de este tipo de aprendizaje, se cita el algoritmo de *clustering* ***K-means***.