Aprendizaje supervisado: Clasificación y regresión

En el aprendizaje supervisado, tenemos un conjunto de datos que consta de características y etiquetas. La tarea consiste en **construir un estimador** que sea capaz de predecir la etiqueta de un objeto dado el conjunto de características.

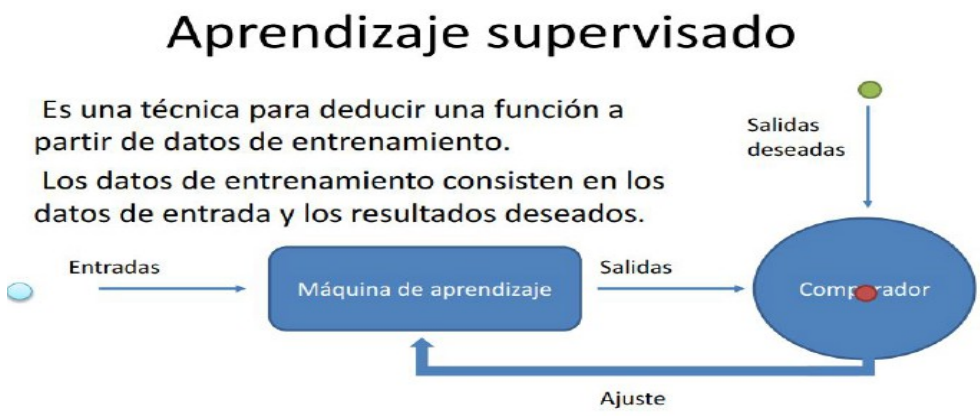
El aprendizaje supervisado se divide en dos categorías: clasificación y regresión. En la clasificación, la etiqueta es discreta, mientras que en la regresión, la etiqueta es continua. Por ejemplo, en astronomía, la tarea de determinar si un objeto es una estrella, una galaxia o un cuásar es un problema de clasificación: la etiqueta es de tres categorías distintas. Por otro lado, podríamos desear estimar la edad de un objeto sobre la base de tales observaciones: este sería un problema de regresión, porque la etiqueta (edad) es una cantidad continua.

Problemas de clasificación

● Un ejemplo relativamente simple es predecir la especie de flor iris dado un sistema de medidas.
● Dada una imagen multicolor de un objeto a través de un telescopio, determine si ese objeto es una estrella, un cuásar o una galaxia.
● Dado una fotografía de una persona, identifique a la persona en la foto.
● Dado una lista de películas que una persona ha visto y su calificación personal de la película, recomendar una lista de películas que les gustaría,también llamados sistemas de recomendación.

Lo que estas tareas tienen en común es que hay una o más cantidades desconocidas asociadas con el objeto que necesitan ser determinadas a partir de otras cantidades observadas.

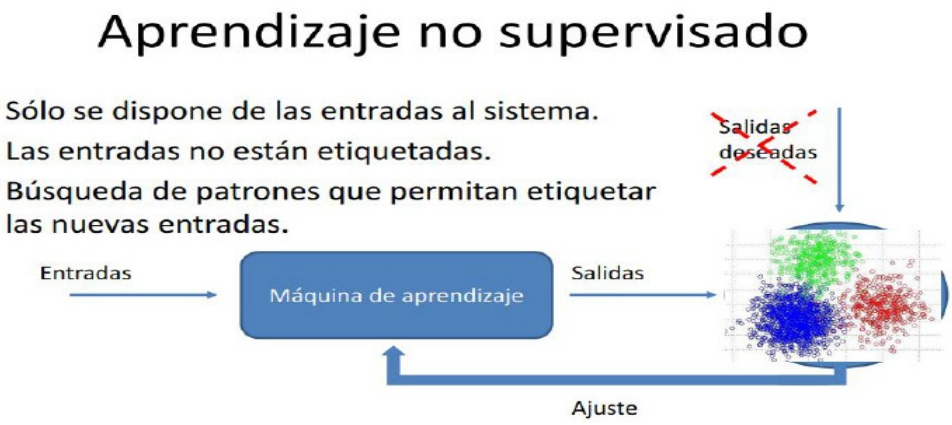
En los problemas de aprendizaje supervisado se enseña o entrena al algoritmo a partir de datos que ya vienen etiquetados con la respuesta correcta. Cuanto mayor es el conjunto de datos más el algoritmo puede aprender sobre el tema. Una vez concluído el entrenamiento, se le brindan nuevos datos, ya sin las etiquetas de las respuestas correctas, y el algoritmo de aprendizaje utiliza la experiencia pasada que adquirió durante la etapa de entrenamiento para predecir un resultado. Esto es similar al método de aprendizaje que se utiliza en las escuelas, donde se nos enseñan problemas y las formas de resolverlos, para que luego podamos aplicar los mismos métodos en situaciones similares.



Aprendizaje no supervisado

En los problemas de aprendizaje no supervisado el algoritmo es entrenado usando un conjunto de datos que no tiene ninguna etiqueta; en este caso, nunca se le dice al algoritmo lo que representan los datos. La idea es que el algoritmo pueda encontrar por sí solo patrones que ayuden a entender el conjunto de datos. El aprendizaje no supervisado es similar al método que utilizamos para aprender a hablar cuando somos pequeños, en un principio escuchamos hablar a nuestros padres y no entendemos nada; pero a medida que vamos escuchando miles de conversaciones, nuestro cerebro comenzará a formar un modelo sobre cómo funciona el lenguaje y comenzaremos a reconocer patrones y a esperar ciertos sonidos.

Por ejemplo tenemos una lista de clientes y cada uno tiene una serie de features.
El algoritmo tratará de agruparlos o clasificarlos en función de la similitud de las features



Aprendizaje supervisado vs no supervisado

Podemos dividir los problemas de aprendizaje automático en dos grandes categorías:

● **Aprendizaje supervisado** , cuando el conjunto de datos viene con los atributos adicionales que queremos predecir. El problema puede clasificarse en dos categorías:

○ Regresión: los valores de salida consisten en una o más variables continuas. Un ejemplo es la predicción del valor de una casa en función de su superficie útil, número de habitaciones, cuartos de baños, etc.

○ Clasificación: las muestras pertenecen a dos o más clases y queremos aprender a partir de lo que ya conocemos cómo clasificar nuevas muestras. Tenemos como ejemplo el Iris dataset que ya mostramos en la entrada anterior

● **Aprendizaje no supervisado** , cuando no hay un conocimiento a priori de las salidas que corresponden al conjunto de datos de entrada. En estos casos el objetivo es encontrar grupos mediante clustering o determinar una distribución de probabilidad sobre un conjunto de entrada.