Machine Learning I

Francisco Javier Mercader Martínez

Índice

Tema 1: Introducción al Machine Learning

¿Qué es el Machine Learning?

- Definición de Machine Learning

"Descubrir regularidades en datos mediante el uso de algoritmos, y mediante el uso de esas regularidades realizar alguna acción" (C. M. Bishop)

- Tareas básicas

Fundamentalmente cuatro:

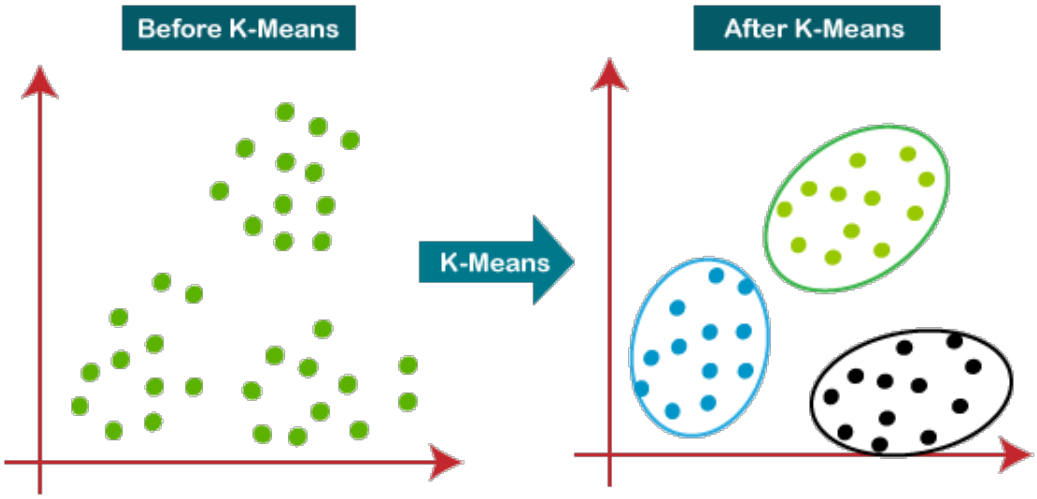
- Clasificación
  - → **Detección de spam:** Se trata de clasificar, mediante identificación de patrones, los correos electrónicos como spam o no spam.
  - → **Detección de fraudes:** Distinción entre transacciones legítimas y sospechosas basándose en patrones y características relevantes.
  - → **Análisis de sentimientos:** Los algoritmos de clasificación pueden utilizarse para determinar el sentimiento expresado en un texto, como positivo, negativo o neutro. Esto es útil para el análisis de opiniones en redes sociales, comentarios de clientes, revisiones de productos, etc.
  - → **Detección de objetos en imágenes:** Especialmente útil en la conducción de coches autónomos.
- Regresión
  - → **Estimación de la demanda de un producto:** Predicción de la demanda de un producto en función de variables como el precio, la publicidad, las tendencias del mercado, entre otras.
  - → **Predicción de la contaminación atmosférica:** Utilizando datos históricos de contaminantes, meteorología y otras variables relevantes, se puede aplicar la regresión para predecir los niveles de contaminación en una ubicación específica.
  - → **Análisis de la relación entre variables económicas:** La regresión puede utilizarse para explorar la relación entre variables económicas, como el crecimiento del PIB y el desempleo, con el fin de entender mejor su interdependencia y tomar decisiones políticas o empresariales informadas.
- Agrupamiento
- Asociación

- Tarea de agrupación en Machine Learning

El **agrupamiento** o **clustering** consiste en detectar agrupaciones en datos no etiquetados empleando alguna media de similitud entre ellas. El objetivo es descubrir patrones y estructuras dentro de los datos.

Algoritmos populares para clustering incluyen el K-Means, el DSCAN, el clustering jerárquico y Mapas Autoorganizados (SOM).

Ejemplo K-Means



- Tarea de asociación en Machine Learning

La tarea de **asociación** se centra en descubrir reglas de asociación entre eventos en un conjunto de datos, lo que significa identificar qué elementos tienden a aparecer juntos en dichos eventos. El objetivo es revelar después del afeitado, hay un 80% de posibilidades de que el cliente compre también crema de afeitado.

La asociación es una tarea no supervisada, los datos a menudo provienen de transacciones o eventos, y no se requieren etiquetas previas.

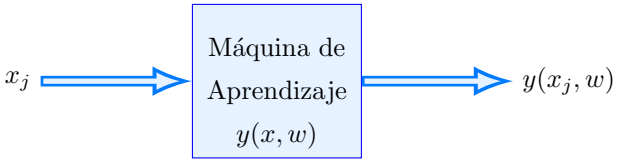
Algoritmos como Apriori se utilizan comúnmente para generar reglas de asociación en los datos, reglas como "Si A, entonces B". Estas reglas se utilizan en análisis de mercado y sistemas de recomendación.

1.1) Planteamiento del problema

En el contexto del Machine Learning, el **conjunto de hipótesis** se refiere a un conjunto de funciones o modelos matemáticos que se utilzian para aproximar una relación desconocida entre las **entradas** (x) y las **salidas deseadas o targets** (t) de un conjunto de datos.

Cada hipótesis representa una posible aproximación de la relación subyacente en los datos.

El objetivo del **aprendizaje supervisado** es encontrar la hipótesis que mejor se ajuste a los datos de entrenamiento manteniendo la capacidad de hacer predicciones precisas para datos nuevos **(capacidad de generalización)**.



- Necesario: Conjunto de entrenamiento

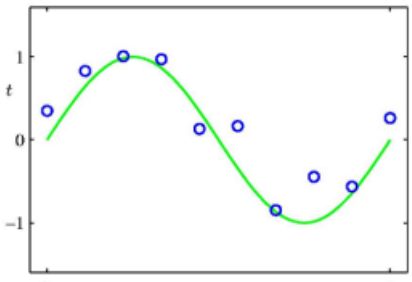
  Pares: $\{x_j, t_j\}$ con $j = 1, 2, \ldots, N$.
  $x_j = \{x_{j1}, x_{j2}, \ldots, x_{jD}\}$ entrada j-ésima; vector con D **componentes o características**.
  $t_j = \{x_{j1}, x_{j2}, \ldots, x_{jT}\}$ target j-ésimo; vector con T componentes.

- Objetivo: Aprendizaje supervisado

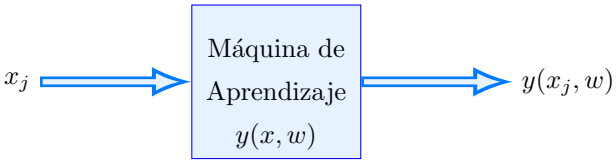
  Encontrar las **variables o pesos** del modelo $(\mathbf{w})$ que resuelvan el problema: $y(x_j, \mathbf{w}) \approx t_j$ para $j = 1, 2, \ldots, N$. A esta tarea se la denomina **entrenamiento**.

Ejemplo: Problema de **regresión**

$y = \sin(2\pi x) + n(x)$, donde $n(x)$ es un ruido gausiano pequeño.



Conjunto de entrenamiento: $\{x_j, t_j\}_{j=1}^{N=10}$



Aproximador polinómico: $y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + x_M x^M = \sum_{j=0}^{M} x_j x^j$

- M es un parámetro que determina la complejidad del modelo (orden del polinomio).
- Los parámetros no entrenables que determinan el modelo o el entrenamiento se denominan en Machine Learning **hiperparámetros**.

1.2) Planteamiento de la solución

Se quiere encontrar las variables del modelos (coeficientes del polinomio) para que éste minimice una función de coste o error, por ejemplo, la función de error SSE ("Sum of Square Error") dada por

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

Existe una solución analítica única mediante álgebra lineal.

$$E(w) = \frac{1}{2} \sum_{n=1}^{N} (y(x_n, w) - t_n)^2$$