Machine Learning

RAQUEL SOCORRO LEÓN

Machine Learning

Es una rama de la Inteligencia Artificial, su objetivo es crear un modelo que nos permita resolver una tarea. Una vez elaborado el modelo, lo entrenamos usando gran cantidad de datos. El modelo aprende de estos datos y es capaz de hace predicciones. Según nuestro objetivo, trabajaremos con un algoritmo u otro.

Existen tres tipos de aprendizaje:

- Aprendizaje Supervisado
- Aprendizaje No Supervisado
- Aprendizaje por Refuerzo

Machine Learning (II)

- Aprendizaje Supervisado:

El modelo tiene unos datos de entrada y unos datos de salida conocidos. A partir del entrenamiento elaboramos el modelo que podrá predecir el valor correspondiente a cualquier dato de entrada.

La salida de los datos puede ser un valor numérico o categórico.

Se suele utilizar en problemas de:

- **Regresión**: predecir el precio de una vivienda, expectativas de vida, predicciones meteorológicas.

- **Clasificación**: detección de fraude, detección de correos spam.

Algoritmos de Regresión:

- Simple Linear regression
- Multiple Linear regression
- Polynomial regression
- Support Vector Machine
- Regression
- Regression tree decisión
- Random forest Regression

Algoritmos de Clasificación:

- Neural Networks
- KNN
- Decision Trees
- Random Forest
- Support Vector Machines (SVM)
- Naive Bayes

Machine Learning (III)

- Aprendizaje No Supervisado:

Estos algoritmos discriminan o separan las observaciones (datos de entrada) en diferentes grupos.

Sólo conocemos los datos de entrada, pero no existen datos de salida que correspondan a un determinado *input*. Por tanto, sólo podemos describir la estructura de los datos, para intentar encontrar algún tipo de organización que simplifique el análisis. Por ello, tienen un carácter exploratorio.

Se suele utilizar en problemas de:

- **Clustering**: para el reconocimiento de comunidades en redes sociales, segmentación de datos, sistemas de recomendación, detección de anomalías, etc.

- **Reducción de dimensionalidad**: identificar variables relevantes, etc.

Algoritmos Clustering:

- Hierarchical clustering
- K-means clustering
- K-NN (k nearest neighbors)
- Principal Component Analysis
- Singular Value Decomposition
- Independent Component Analysis
- Dimensionality reduction

Algoritmos Reducción dimensionalidad:

- Principal component analysis (PCA)
- Non-negative matrix factorization (NMF)
- Kernel PCA
- Graph-based kernel PCA
- Linear discriminant análisis
- t-SNE

Machine Learning (IV)

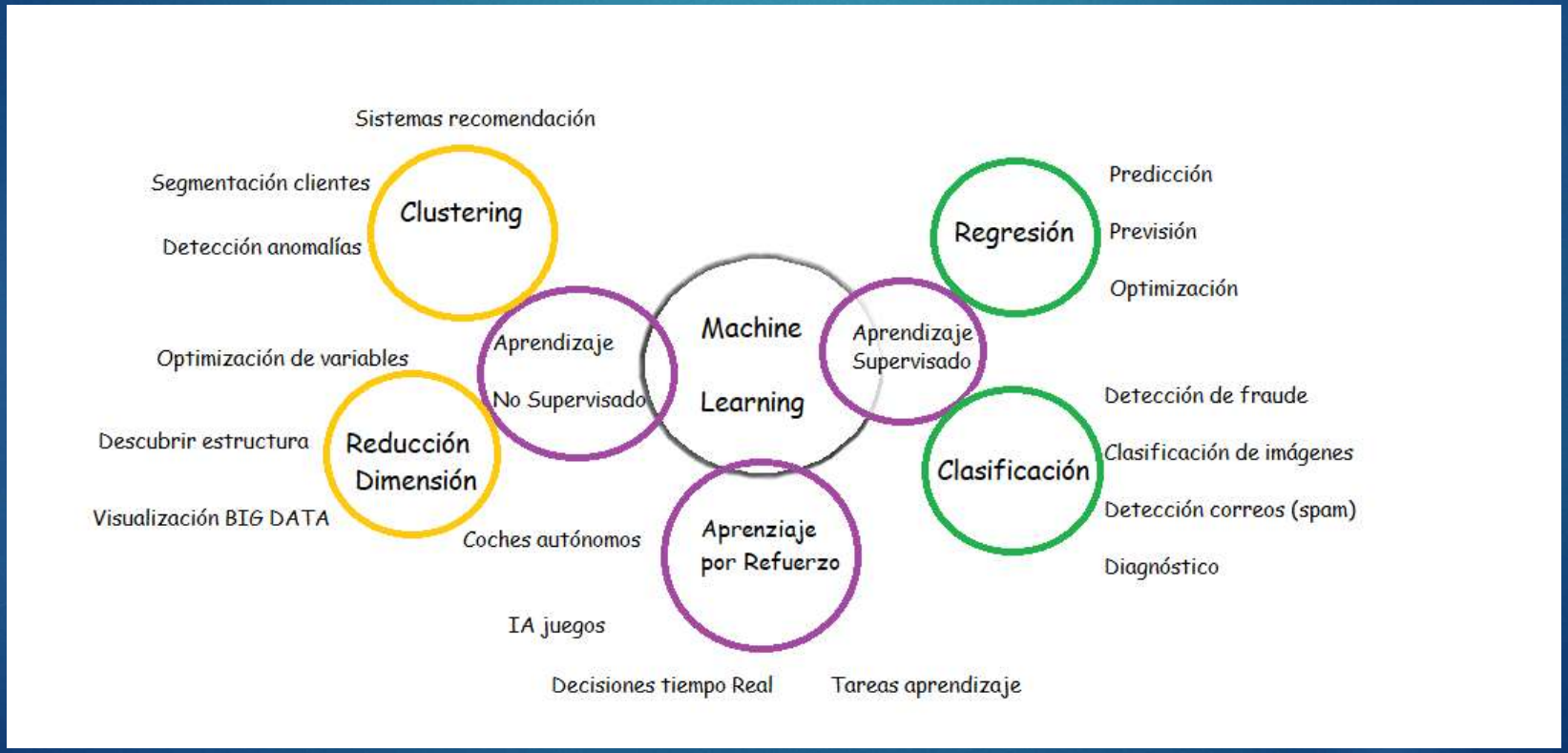
- Aprendizaje de Refuerzo:

Se basa en mejorar la respuesta del modelo usando un proceso de retroalimentación. El algoritmo aprende observando el mundo que le rodea. Su información de entrada es el *feedback* o retroalimentación que obtiene del mundo exterior como respuesta a sus acciones. Por lo tanto, el sistema aprende a base de ensayo-error.

Por ejemplo: el aprendizaje de los coches autónomos, chatbot, etc

Machine Learning (V)

- Resumen de ML



Análisis de componentes principales (PCA)

- El análisis de componentes principales (*principal component analysis*) o **PCA** es una de las técnicas de aprendizaje no supervisado.

- Una de las aplicaciones de PCA es la reducción de dimensionalidad (variables), perdiendo la menor cantidad de información posible, es decir, menor cantidad de varianza.

- Se basa en permitir reducir el número de variables transformándolas dando lugar Componentes Principales (CP) que expliquen gran parte de la variabilidad en los datos. Cada dimensión o CP generada por PCA será una combinación lineal de las variables originales, y serán además independientes o no correlacionadas entre sí.

- Por ejemplo: si tenemos un gran número de variables cuantitativas posiblemente correlacionadas es un indicativo de existencia de información redúndate.

Análisis de componentes principales (PCA) (II)

- Vectores propios (eigen vector) y valores propios (*eigen values)*

Los vectores propios y los valores propios corresponden a números y vectores asociados a matrices cuadrada.

- Dada una matriz cuadrada M, $n \times n$, diremos que v es un vector propio de M y que λ es un valor propio de M si se cumple que $M \cdot v = \lambda \cdot v$.

Tomemos como matriz $M = \begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix}$, si la multiplicamos por el vector $\begin{pmatrix} 3 \\ 2 \end{pmatrix}$

tendremos que $\begin{pmatrix} 2 & 3 \\ 2 & 1 \end{pmatrix} \cdot \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} 12 \\ 8 \end{pmatrix} = 4 \cdot \begin{pmatrix} 3 \\ 2 \end{pmatrix}$

$M \cdot v$ $\lambda \cdot v$

Análisis de componentes principales (PCA) (III)

- Propiedades de los vectores propios:

- Solo las matrices cuadradas n x n tienen vectores propios, pero no todas las matrices cuadradas lo tienen. Si una matriz n × n tiene vectores propios, entonces el número de vectores propios que tendrá será exactamente n.

- *Un* vector propio escalada, es decir, si se multiplica por cierto valor antes de multiplicarlo por un matriz, el vector propio continuará manteniendo su propiedad, ya que solo se cambia su longitud , no su dirección.

- Los valores propios son los valores con los que se multiplica el vector propio y que da lugar al vector original.

- Todos los vectores propios de una matriz son perpendiculares entre sí. Esto significa que podemos representar los datos en función de las nuevas coordenadas formadas por los vectores propios.

Análisis de componentes principales (PCA) (IV)

- Estandarización de las variables

El cálculo de los componentes principales depende de las unidades de medida empleadas en las variables.

Por tanto, antes de aplicar el PCA es necesario estandarizar las variables para que tengan **media 0 y desviación estándar 1**, ya que, de lo contrario, las variables con mayor varianza dominarían al resto, aunque en el caso en que las variables estén medidas en las mismas unidades, podemos optar por no estandarizarlas.

Análisis de componentes principales (PCA) (V)

- Componentes Principales (PCA)

Cada componente principal se obtiene por combinación lineal de las variables originales.

El proceso a seguir para calcular la primera componente principal es:

1. Estandarización de las variables: se resta cada valor la media de la variable a la que pertenece.
2. Se calcula los vectores y valores propios (eigenvector – eigenvalue)

Análisis de componentes principales (PCA) (VI)

- Proceso del PCA utilizando la función prcomp() en R:

1. Estandarización de los datos, es decir, media 0 y desviación típica 1, utilizamos prcomp() e incluimos el parámetro scale que hace referencia a dt 1

2. Elementos de prcomp:
   - sdev : desviaciones estándar de los componentes principales.
   - rotation: matriz cuyas columnas contienen los vectores propios, es decir, la combinación lineal de las variables originales. Es decir, contiene el valor de los loadings
   - center: media de las variables estandarizadas
   - scale: desviación típica de las variables estandarizadas
   - x : calculo automático del valor de las componentes principales para cada observación multiplicando los datos por los vectores propios, es decir, nos proyecta los datos sobre el nuevo eje. También se puede obtener mediante scores

Análisis de componentes principales (PCA) (VII)

- Summary del prcomp()

Realizar un summary del prcomp() nos devuelve la siguiente información:

Standard deviation: desviación estándar de cada componente

Proportion of Variance: proporción de varianza explicada

Cumalative Proportion: proporción de varianza explicada acumulada.