

Programa de Machine Learning – Clase a Clase

El cronograma tentativo de clases va a ser:

- Clase 1 : Repaso de Python.Introducción a Machine Learning. Rol del Data Scientist. Tipos de Datos. Análisis de Datos. Tipos de Aprendizaje. Supervisado y No Supervisado. Clasificación de Algoritmos.
- Clase 2 : Teoría de la información. Árboles de decisión. Extracción de features. Validación cruzada. Random Forest.
- Clase 3 : Support Vector Machine. XGBoost. Implementación. Aplicaciones.
- Clase 4 : Aprendizaje supervisado.Conceptos de Regresión, Clasificación y Ranking. Regresión lineal. Regresión logística.Implementación. Aplicaciones.
- Clase 5 : Aprendizaje No Supervisado. K-Means.Implementación. Aplicaciones.
- Clase 6 : Ciclo de vida de proyectos de ML. AutoML. MLOps + Deployment (inicial).
- Clase 7 : Ensamble de modelos.Bagging. Boosting. Stacking.Implementación. Aplicaciones.
- Clase 8 : Exposición de trabajos finales.

Herramientas

Durante la cursada vamos a utilizar las siguientes herramientas:

- Lenguaje de programación:
 - Python 3.8
 - Herramienta pip para instalar librerías de código y dependencias
- Librerías de código:
 - Numpy 1.18
 - SciPy 1.5
- Consola interactiva de Python:
 - iPython
- Herramientas:
 - PyTest para tests, GitHub para repositorios y uWSGI para servidor web.

Mecanismos de Evaluación

El régimen de aprobación de la materia es simple:

- Google form (cuando haya)
- Trabajos prácticos a implementarse y entregar durante las clases (2).
- Exposición en la clase 8.

Definiendo Machine Learning

PRINCIPALES CAMBIOS QUE SE PRODUJERON EN LA TECNOLOGÍA Y EN LOS ÚLTIMOS AÑOS

- MASIFICACIÓN USO DE INTERNET
- SURGIMIENTO DE LAS REDES SOCIALES
- CRECIMIENTO EXPONENCIAL DE DISPOSITIVOS MÓVILES
- INTERFACES DE USUARIO MAS SIMPLES E INTUITIVAS

CADA DÍA CREAMOS 2,5
QUINTILLONES DE BYTES
DE
DATOS. (2,5 Exabytes)

EL 90% DE LOS DATOS DEL
MUNDO DE HOY SE
GENERARON EN LOS
ÚLTIMOS 2 AÑOS

Definiendo Machine Learning

Machine Learning Una Primer Definición

“Es la ciencia que permite que las computadoras aprendan y actúen como lo hacen los humanos, mejorando su aprendizaje a lo largo del tiempo de una forma autónoma, alimentándolas con datos e información en forma de observaciones e interacciones con el mundo real.”

Programación Tradicional vs Machine Learning

Programación Tradicional



Machine Learning



Relación Machine Learning / Big Data

Big Data

Una Primer Definición

“Volumen masivo de datos, tanto estructurados como no-estructurados, los cuales son demasiado grandes y difíciles de procesar con las bases de datos y el software tradicionales.”

Definiendo Big Data

Big Data es el sector de IT que hace referencia a *grandes conjuntos de datos* que por la *velocidad* a la que se generan, la capacidad para tratarlos y los *múltiples formatos y fuentes*, es necesario procesarlos con mecanismos distintos a los tradicionales.

BIG DATA

Volumen

Velocidad

Variedad

Veracidad

Almacenarlos

Recolectarlos

Compartirlos

Buscarlos

DATOS

Analizarlos

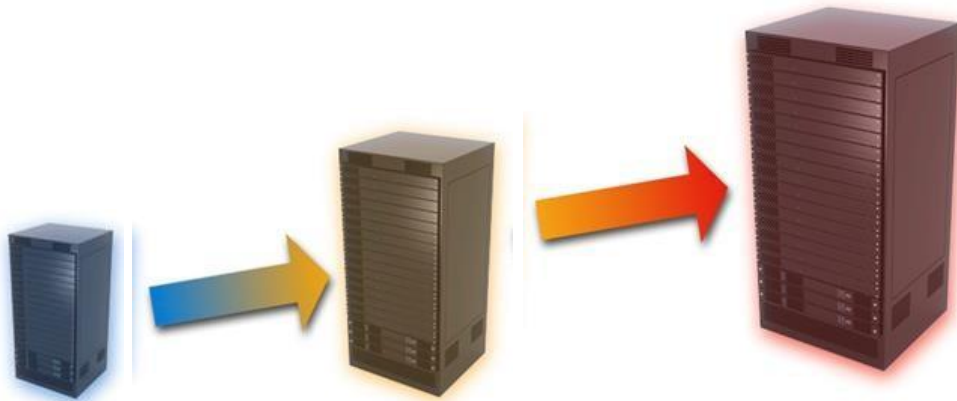
Visualizarlos

Procesarlos

Entenderlos

Implementando Big Data

ESCALAMIENTO



Escalamiento
Vertical

Escalamiento Vertical

- Escalamiento dentro de un mismo servidor.
- Implica incrementar la capacidad de un Servidor agregando más recursos de CPU, memoria y de almacenamiento.

Implementando Big Data

ESCALAMIENTO



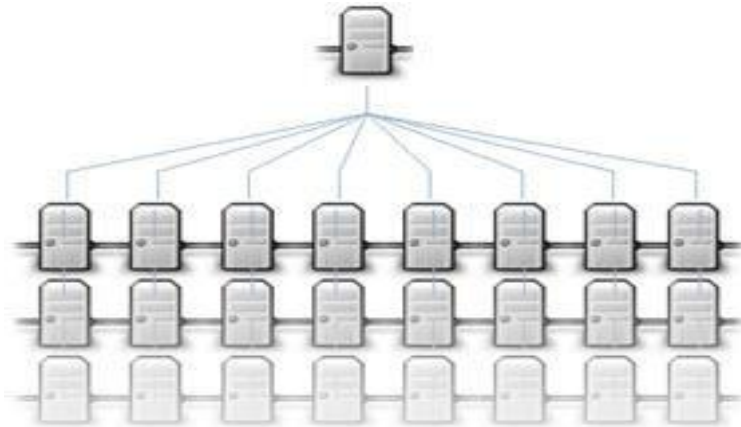
Escalamiento
Horizontal

Escalamiento Horizontal

- Escalamiento en varios servidores.
- Cluster de Servidores.
- Replicación de Datos.
- Particionamiento de Datos.
- Procesamiento Paralelo.

Implementando Big Data

Cluster



Google

amazon

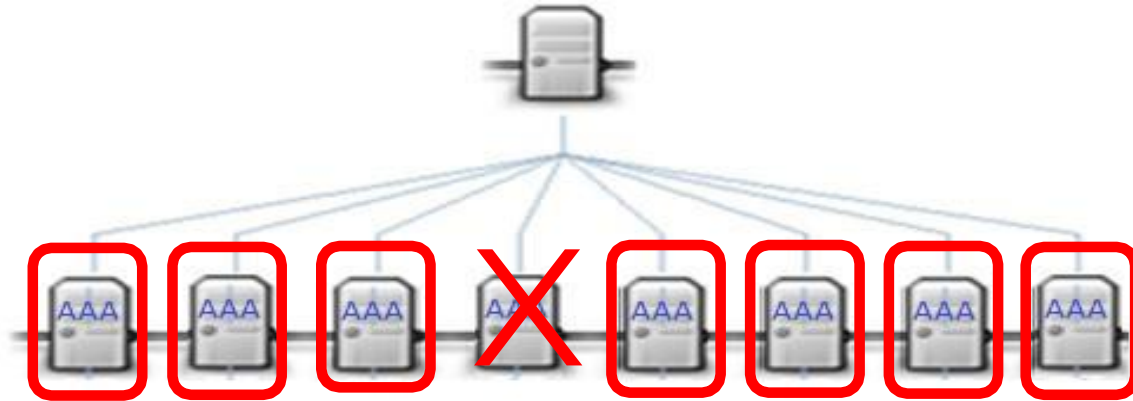
facebook

Grupo de servidores independientes interconectados a través de una red dedicada que trabajan como un único recurso de procesamiento

Implementando Big Data

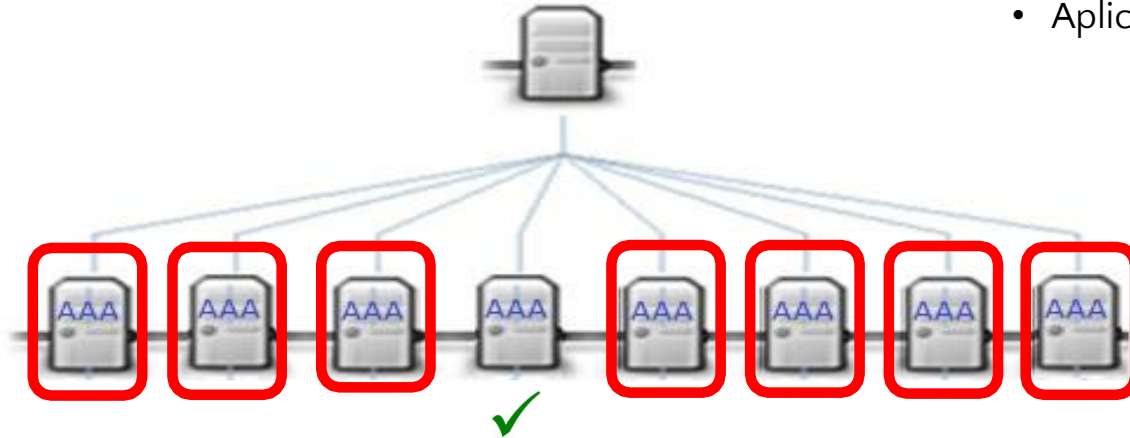
ALTA DISPONIBILIDAD Y TOLERANCIA A FALLOS

- Aplicaciones 7 x 24.
- Aplicaciones de Misión Crítica.



Implementando Big Data

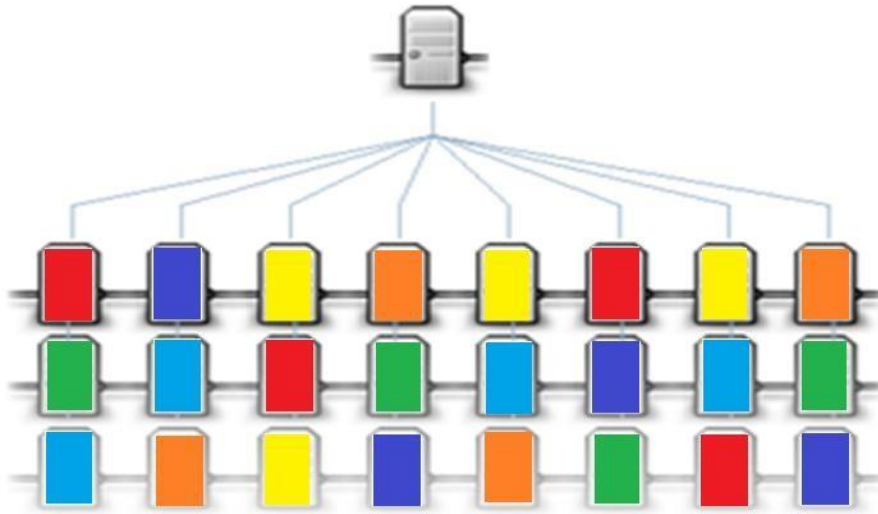
ALTA DISPONIBILIDAD Y TOLERANCIA A FALLOS



- Aplicaciones 7 x 24.
- Aplicaciones de Misión Crítica.

Implementando Big Data

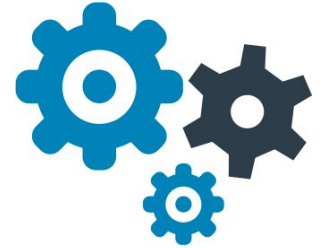
PARTICIONAMIENTO DE DATOS



- Un solo Servidor no soporta almacenar la totalidad de los datos.
- Se deben particionar los datos en múltiples Servidores del Cluster.
- Además los datos se encuentran replicados.

Implementando Big Data

PROCESAMIENTO PARALELO




- Varios servidores procesan un mismo programa de forma simultánea para resolver un determinado problema.

Terminología Básica y Notaciones

En Machine Learning generalmente se utilizan matrices y notaciones vectoriales para referirnos a los datos, de la siguiente forma:

- Cada fila de la matriz es una muestra, observación o dato puntual.
- Cada columna es una característica (o atributo), de la observación mencionada en el punto anterior (“feature” en la imagen inferior).
- En el caso más general habrá una columna, que llamaremos objetivo, etiqueta o respuesta, y que será el valor que se pretende predecir. (“label” en la imagen inferior).

Terminología Básica y Notaciones

 Features					Label
Position	Experience	Skill	Country	City	Salary (\$)
Developer	0	1	USA	New York	103100
Developer	1	1	USA	New York	104900
Developer	2	1	USA	New York	106800
Developer	3	1	USA	New York	108700
Developer	4	1	USA	New York	110400
Developer	5	1	USA	New York	112300
Developer	6	1	USA	New York	114200
Developer	7	1	USA	New York	116100
Developer	8	1	USA	New York	117800
Developer	9	1	USA	New York	119700
Developer	10	1	USA	New York	121600

Terminología Básica y Notaciones

- Con respecto a los algoritmos de Machine Learning, normalmente tienen determinados parámetros “internos”.
- Por ejemplo en los árboles de decisión, hay parámetros como profundidad máxima del árbol, número de nodos, número de hojas,...a estos parámetros se les llama “*hiperparámetros*”.
- Llamamos “*generalización*” a la capacidad del modelo para hacer predicciones utilizando nuevos datos.

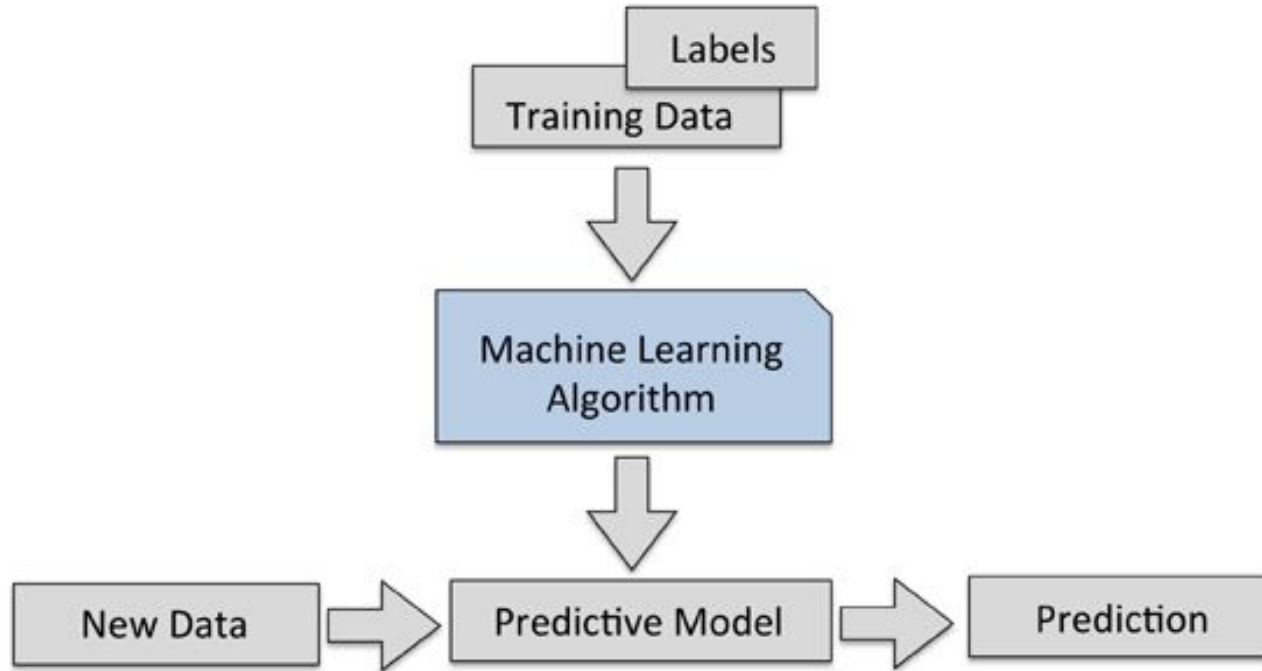
Tipos de Machine Learning

- Aprendizaje supervisado : Se refiere a un tipo de modelos de Machine Learning que se entrenan con un conjunto de ejemplos en los que los resultados de salida son conocidos.
- Aprendizaje no supervisado: El objetivo será la extracción de información significativa, sin la referencia de variables de salida conocidas, y mediante la exploración de la estructura de dichos datos sin etiquetar.
- Aprendizaje profundo: Es un subcampo de Machine Learning, que usa una estructura jerárquica de redes neuronales artificiales, que se construyen de una forma similar a la estructura neuronal del cerebro humano, con los nodos de neuronas conectadas.

Aprendizaje Supervisado

- Los modelos aprenden de los resultados conocidos y realizan ajustes en sus parámetros interiores para adaptarse a los datos de entrada.
- Una vez que el modelo es entrenado adecuadamente, y los parámetros internos son coherentes con los datos de entrada y los resultados de los datos de entrenamiento, el modelo podrá realizar predicciones adecuadas ante nuevos datos no procesados previamente.

Aprendizaje Supervisado



Aprendizaje Supervisado

Hay dos aplicaciones principales de aprendizaje supervisado: clasificación y regresión:

- **Clasificación:**

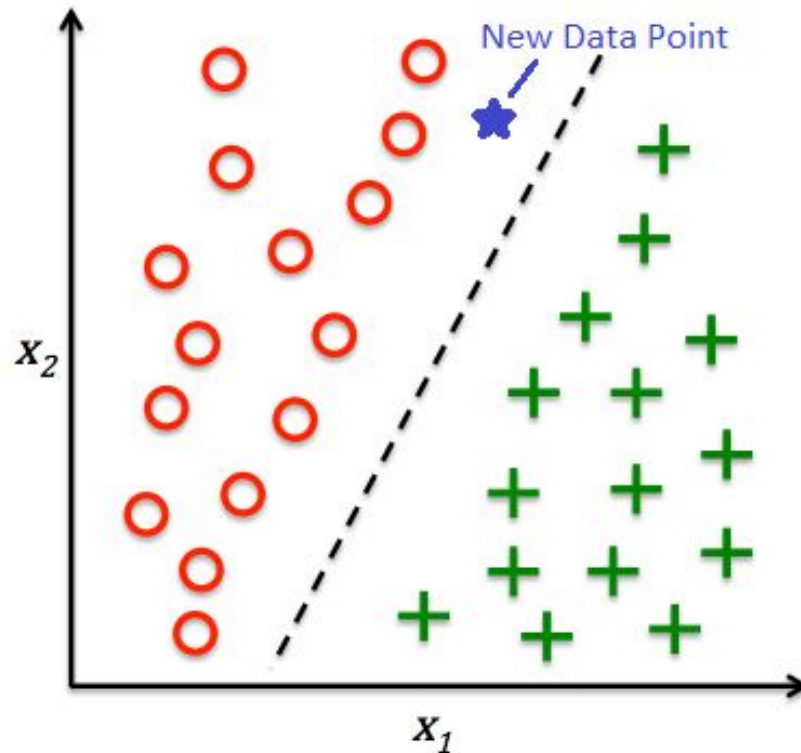
Clasificación es una subcategoría de aprendizaje supervisado en la que el objetivo es predecir las clases categóricas (valores discretos, no ordenados, pertenencia a grupos).

- El ejemplo típico es la detección de correo spam, que es una clasificación binaria (un email es spam — valor “1” — o no lo es — valor “0” —). También hay clasificación multi-clase, como el reconocimiento de caracteres escritos a mano (donde las clases van de 0 a 9).

Aprendizaje Supervisado

- Un ejemplo de clasificación binaria: hay dos clases de objetos, círculos y cruces, y dos características de los objetos, X_1 y X_2 .
- El modelo puede encontrar las relaciones entre las características de cada punto de datos y su clase, y establecer la línea divisoria entre ellos.
- Así, al ser alimentado con nuevos datos, el modelo será capaz de determinar la clase a la que pertenecen, de acuerdo con sus características.

Aprendizaje Supervisado



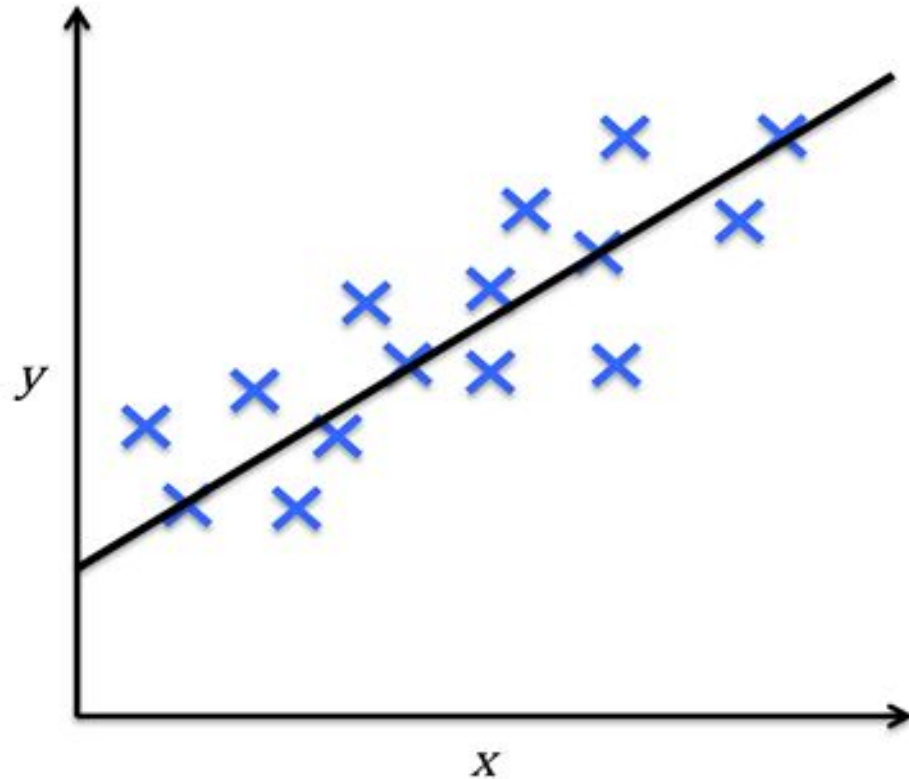
Aprendizaje Supervisado

- **Regresión:**

La regresión se utiliza para asignar categorías a datos sin etiquetar. En este tipo de aprendizaje tenemos un número de variables predictoras (explicativas) y una variable de respuesta continua (resultado), y se tratará de encontrar una relación entre dichas variables que nos proporcione un resultado continuo.

Un ejemplo de regresión lineal: dados X e Y , establecemos una línea recta que minimice la distancia (con el método de mínimos cuadrados) entre los puntos de muestra y la línea ajustada. Después, utilizaremos las desviaciones obtenidas en la formación de la línea para predecir nuevos datos de salida.

Aprendizaje Supervisado



Aprendizaje No Supervisado

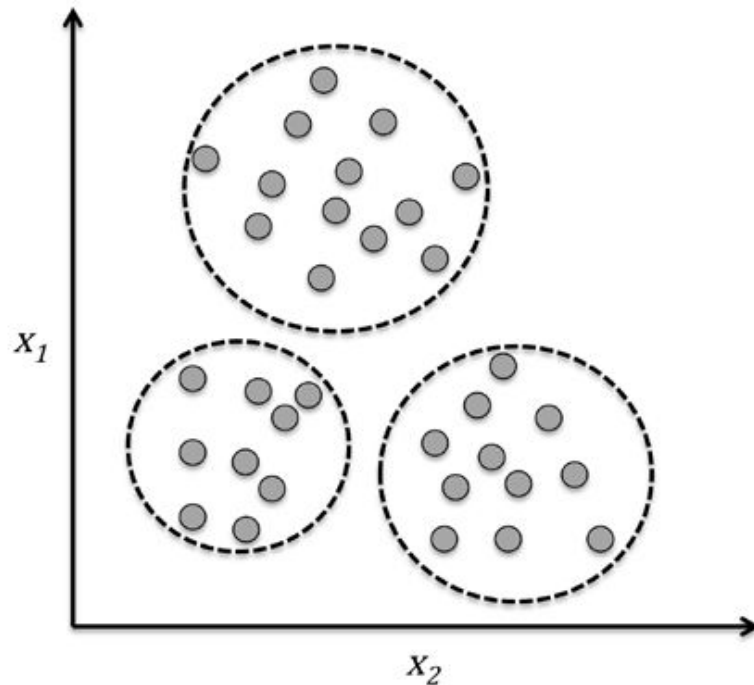
Hay dos categorías principales: agrupamiento y reducción dimensional.

- **Agrupamiento ó Clustering:**

- El agrupamiento es una técnica exploratoria de análisis de datos, que se usa para organizar información en grupos con significado sin tener conocimiento previo de su estructura.
- Cada grupo es un conjunto de objetos similares que se diferencia de los objetos de otros grupos.
- El objetivo es obtener un número de grupos de características similares.

Aprendizaje No Supervisado

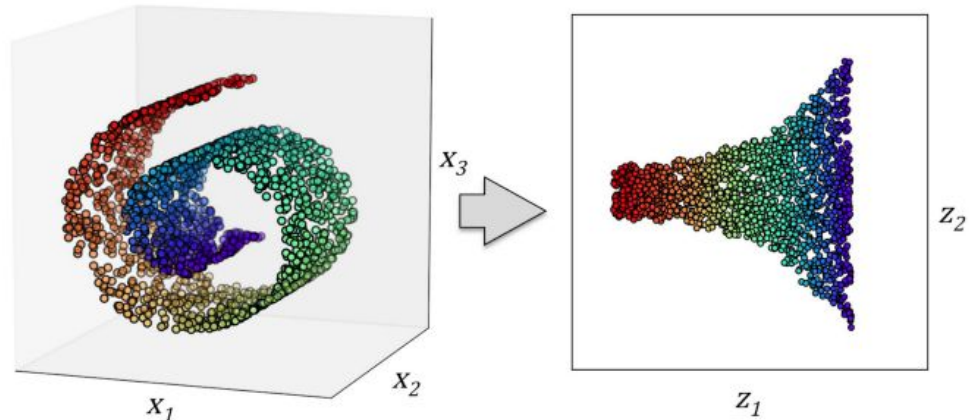
Un ejemplo de aplicación de este tipo de algoritmos puede ser para establecer tipos de consumidores en función de sus hábitos de compra, para poder realizar técnicas de marketing efectivas y “personalizadas”.



Aprendizaje No Supervisado

La **reducción dimensional** funciona encontrando correlaciones entre las características, lo que implica que existe información redundante, ya que alguna característica puede explicarse parcialmente con otras (por ejemplo, puede existir dependencia lineal).

Estas técnicas eliminan “ruido” de los datos (que puede también empeorar el comportamiento del modelo), y comprimen los datos en un sub-espacio más reducido, al tiempo que retienen la mayoría de la información relevante.



Aprendizaje Profundo Deep Learning

Esta arquitectura permite abordar el análisis de datos de forma no lineal.

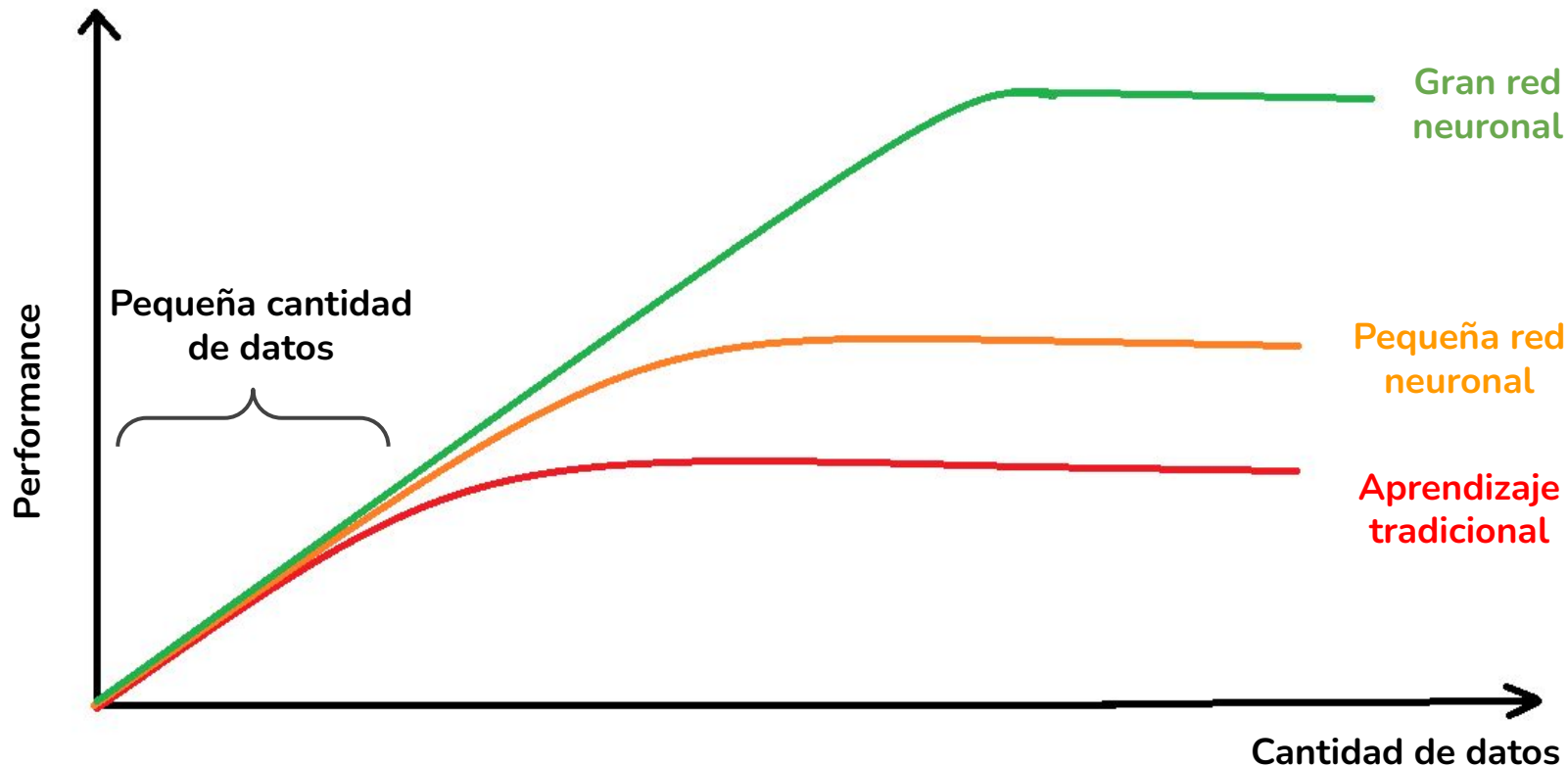
La primera capa de la red neuronal toma datos en bruto como entrada, los procesa, extrae información y la transfiere a la siguiente capa como salida.

Este proceso se repite en las siguientes capas, cada capa procesa la información proporcionada por la capa anterior, y así sucesivamente hasta que los datos llegan a la capa final, que es donde se obtiene la predicción.

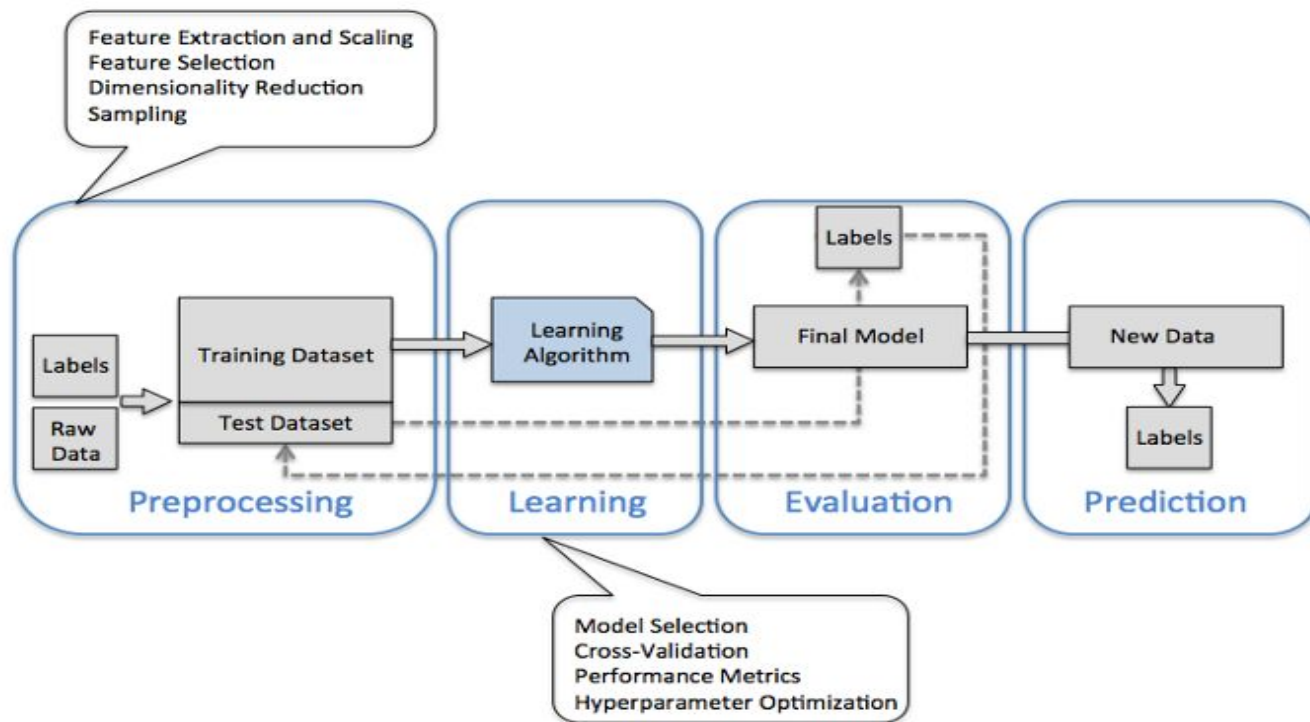
Esta predicción se compara con el resultado conocido, y así por análisis inverso el modelo es capaz de aprender los factores que conducen a salidas adecuadas.

Es uno de los principales algoritmos utilizados en la creación de aplicaciones y programas para reconocimiento de imágenes.

Cuándo conviene utilizar Deep Learning ?



Metodología para construir algoritmos de ML



Metodología para construir algoritmos de ML

Preprocesamiento:

- Usualmente los datos se presentan en formatos no óptimos (o incluso inadecuados) para ser procesados por el modelo.
- Muchos algoritmos requieren que las características estén en la misma escala (por ejemplo, en el rango $[0,1]$) para optimizar su rendimiento, lo que se realiza frecuentemente aplicando técnicas de normalización o estandarización en los datos.
- Podemos también encontrar en algunos casos que las características seleccionadas están correlacionadas, y por tanto son redundantes para extraer información con significado correcto de ellas.
- En este caso tendremos que usar técnicas de reducción dimensional para comprimir las características en subespacios con menores dimensiones.

Metodología para construir algoritmos de ML

Entrenando y seleccionando un modelo:

- Es esencial comparar los diferentes algoritmos de un grupo para entrenar y seleccionar el de mejor rendimiento. Para realizar esto, es necesario seleccionar una métrica para medir el rendimiento del modelo.
- Para asegurarnos de que nuestro modelo funcionará adecuadamente con datos reales, utilizaremos la técnica denominada **validación cruzada (Cross Validation)** antes de utilizar el conjunto de datos de prueba para la evaluación final del modelo.
- En general, los parámetros por defecto de los algoritmos de Machine Learning proporcionados por las librerías no son los mejores para utilizar con nuestros datos, por lo que usaremos técnicas de optimización de “**hiperparámetros**”

Metodología para construir algoritmos de ML

Evaluando Modelos y Prediciendo con Datos Nuevos:

- Una vez que hemos seleccionado y ajustado un modelo a nuestro conjunto de datos de entrenamiento:

Podemos usar los datos de prueba para estimar el rendimiento del modelo en los datos nuevos, por lo que podemos hacer una estimación del error de generalización del modelo, o evaluarlo utilizando alguna otra métrica.

Bibliografía

- The Elements of Statistical Learning | Trevor Hastie | Springer
- An Introduction to Statistical Learning | Gareth James | Springer
- Deep Learning | Ian Goodfellow | <https://www.deeplearningbook.org/>