Modelos de aprendizaje automático mediante la selección de características

El término aprendizaje automático (Machine Learning en inglés) generalmente se refiere a un sistema que aprende de los datos en lugar de aprender de un programa. Nuestro objetivo es entrenar un modelo o programa, para producir las salidas correctas para las entradas dadas, sin programarlas explícitamente.

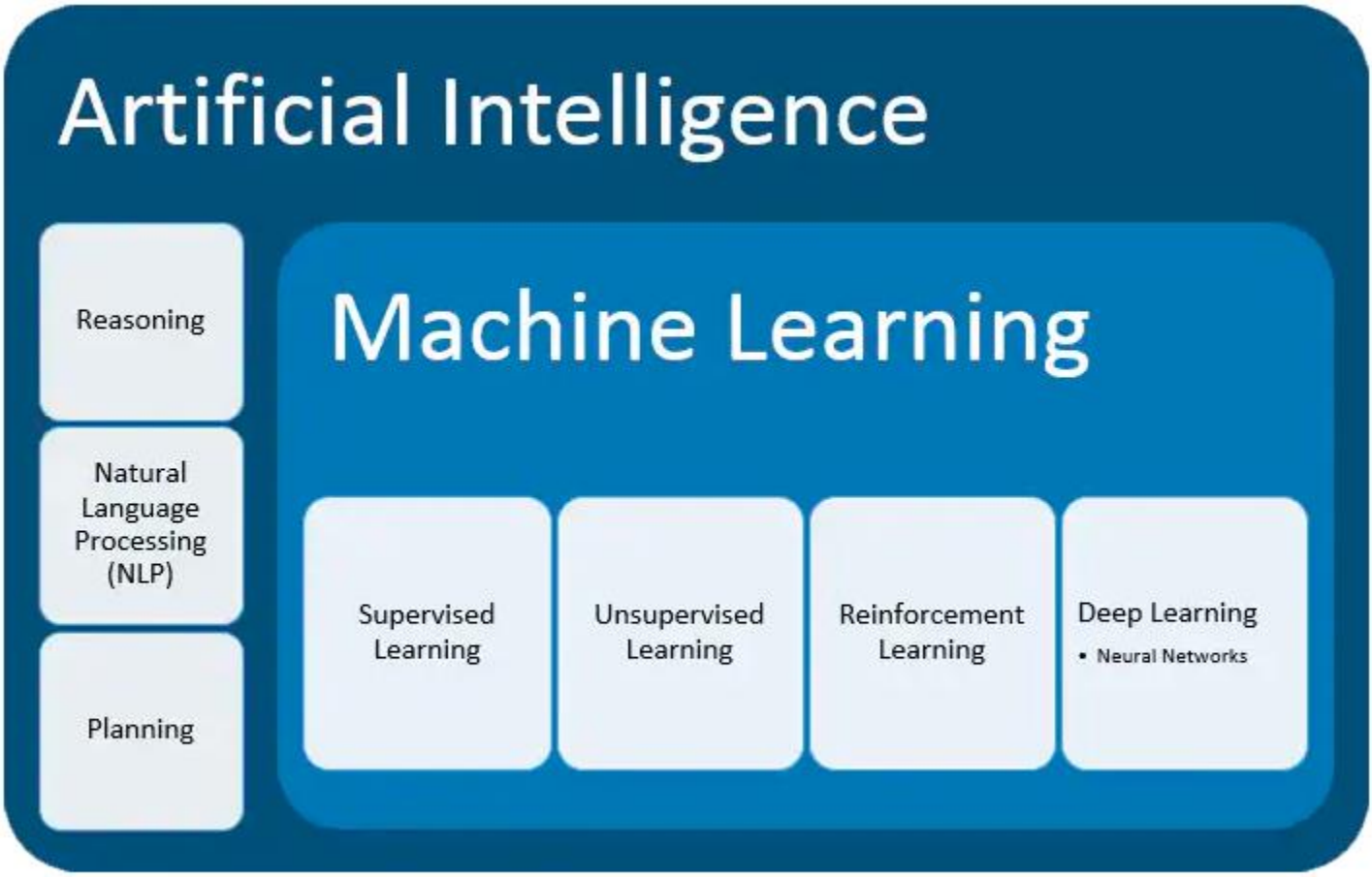


Figura 6. ML https://www.ibm.com/mx-es/analytics/machine-learning

Durante este proceso, el modelo aprende el mapeo entre las entradas y las salidas ajustando sus parámetros internos. Una forma común de entrenar el modelo es proporcionándole un conjunto de datos que tienen que ver con la solución de un problema, para el cual se conoce la salida correcta. Para cada una de estas entradas, le decimos al modelo cuál es la salida correcta para que pueda ajustarse a sí mismo, con el objetivo de producir eventualmente la salida deseada para cada una de las entradas dadas. Dependiendo de la naturaleza del problema que se está atendiendo, existen diferentes enfoques basados en el tipo y volumen de los datos:

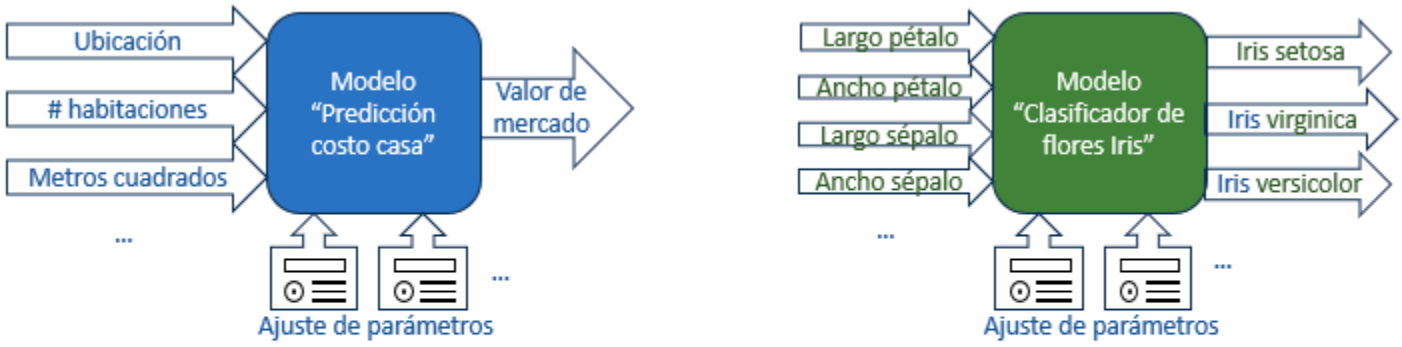
- El **aprendizaje supervisado** típicamente típicamente procesa un conjunto establecido de datos y una cierta comprensión de cómo se clasifican estos datos. La intención es encontrar patrones en los datos que se pueden aplicar a un proceso de analítica. Por ejemplo, se puede crear una aplicación con base en imágenes que describan el lenguaje de señas para sordo mudos.

- El **aprendizaje no supervisado** se utiliza cuando el problema contiene una cantidad masiva de datos sin etiquetar, por lo que, se requiere un proceso iterativo que analice los datos sin intervención humana. Se puede utilizar por ejemplo para detectar spam usando clasificadores de machine learning, basados en clustering y asociación a la actividad del usuario.
- El **aprendizaje de refuerzo** es un modelo donde el sistema aprende a través de la prueba y el error. Por lo tanto, una secuencia de decisiones exitosas conduce a mejorar el proceso. Este tipo de aprendizaje difiere del aprendizaje supervisado, porque el sistema no está entrenado con el conjunto de datos de ejemplo.
- El **aprendizaje profundo** (deep learning) es un método específico de aprendizaje automático (machine learning) donde se incorporan las redes neuronales en capas sucesivas para aprender de los datos de manera iterativa, sus principales usos son el reconocimiento de imágenes, voz y aplicaciones de visión por computadora. Las redes neuronales complejas del aprendizaje automático están diseñadas para emular el funcionamiento del cerebro humano, así que las computadoras pueden ser entrenadas para lidiar con abstracciones y problemas mal definidos.

Algoritmos de aprendizaje automático supervisado

Los dos tipos principales de aprendizaje automático supervisado son la regresión y clasificación:

Los modelos para tareas de **regresión** asignan los valores de entrada en una sola salida para proporcionar un valor continuo, como se ilustra en la figura 6a, mientras que en la **clasificación** el modelo debe decidir a qué categoría pertenece una determinada entrada y cada categoría está representada por una sola salida llamada etiqueta, mientras que las entradas se denominan características (ver figura 6b).



a) Aprendizaje automático supervisado para regresión

b) Aprendizaje automático supervisado para regresión

Figura 6. Modelos a) Regresor de precios de la vivienda. b) Clasificador de flores

Algunos modelos/algoritmos comunes de aprendizaje supervisado incluyen los siguientes:

- Árboles de decisión: Familia de algoritmos que utiliza un gráfico en forma de árbol, donde los puntos de ramificación representan decisiones y las ramas representan sus consecuencias.
- Bosques aleatorios: Algoritmos que crean un gran número de árboles de decisión durante la fase de entrenamiento y utilizan una combinación de sus resultados.
- Máquinas de vectores de soporte: algoritmos que mapean las entradas dadas como puntos en el espacio para que las entradas que pertenecen a categorías separadas se dividan por el mayor espacio posible.

- Redes Neuronales Artificiales: Modelos que consisten en múltiples nodos simples, o neuronas, que pueden interconectarse de varias maneras. Cada conexión puede tener un peso que controla el nivel de la señal que se transporta de una neurona a la siguiente.

A primera vista, puede parecer que cuanta más información podamos usar como entrada, mayores serán nuestras posibilidades de predecir la salida (s) correctamente. Sin embargo, en muchos casos, ocurre lo contrario; Si algunas de las características que utilizamos son irrelevantes o redundantes, la consecuencia podría ser una disminución (a veces significativa) en la precisión de los modelos.

La selección de características es el proceso de seleccionar el conjunto de características más beneficioso y esencial de todo el conjunto dado de características. Además de aumentar la precisión del modelo, una selección exitosa de características puede proporcionar las siguientes ventajas:

- Los tiempos de entrenamiento de los modelos son más cortos.
- Los modelos entrenados resultantes son más simples y fáciles de interpretar.
- Es probable que los modelos resultantes proporcionen una mejor generalización, es decir, funcionan mejor con nuevos datos de entrada que son diferentes a los datos que se utilizaron para el entrenamiento.

Al observar métodos para llevar a cabo la selección de características, los algoritmos genéticos son un candidato natural. En la siguiente actividad mostraremos un conjunto de datos real y utilizaremos el GA para seleccionar las mejores características para un problema de clasificación.