



Machine Learning

Agenda



El Problema del Aprendizaje



Estructura del Aprendizaje



Tipos de Aprendizaje



Machine Learning

El problema del Aprendizaje



El Problema del Aprendizaje

- ¿Qué elementos se encuentran presentes en la imagen?
- ¿Pueden dar una definición para cada elemento identificado?



El Problema del Aprendizaje

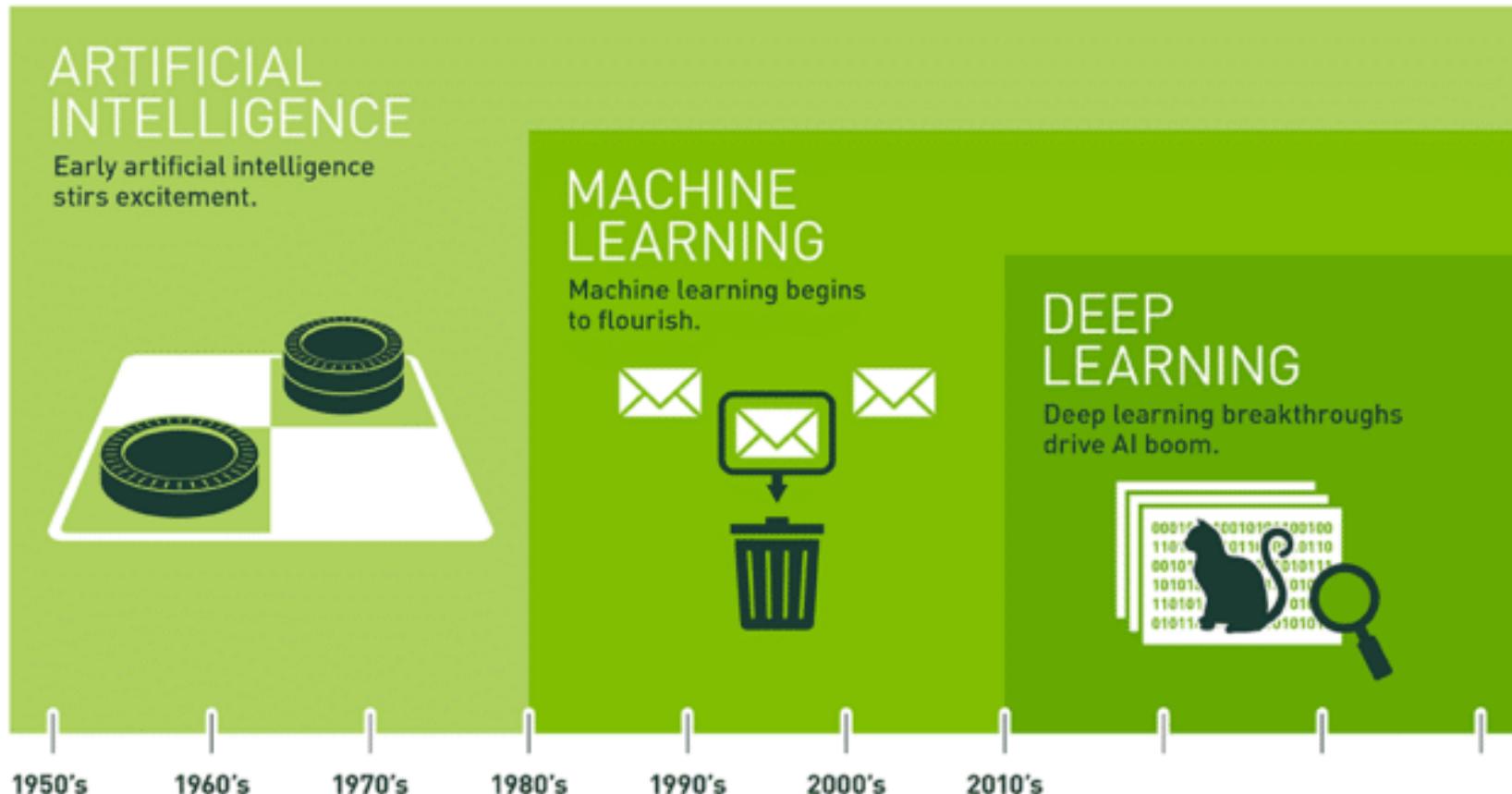
- No aprendemos por medio de **definiciones rigurosas**.
- Aprendemos con **ejemplos**.
- Es decir, se **aprende por medio de datos o ejemplos** (*learn from data*).



El Problema del Aprendizaje

- Aprender de los datos es viable si no existe una solución analítica.
- Existen datos para aproximar una solución.
- Ciencia, ingeniería, economía, finanzas, etc.

Machine Learning



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

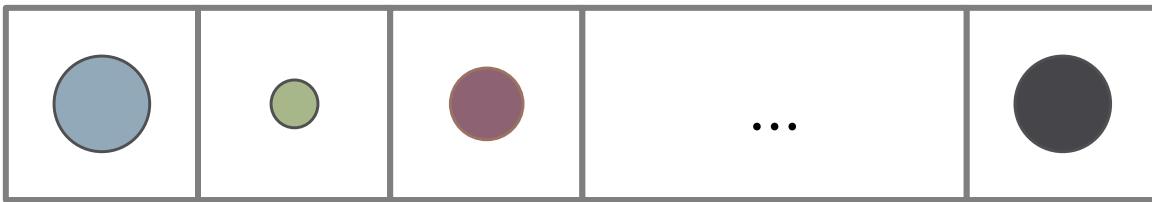
El Problema del Aprendizaje

Problema de sistemas de recomendación para películas

- ¿Cómo puede un sistema recomendar películas a los usuarios?
- Los criterios de cada persona son distintos y muy diversos, complejos.
- Modelarlo suena complicado, desde el punto de vista analítico.
- ¿Existe una solución empírica?

El Problema del Aprendizaje

Preferencias del usuario



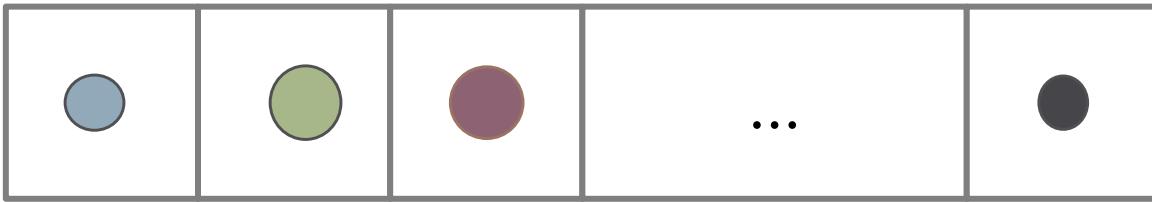
¿Comedia?

¿Acción?

¿Idioma?

¿Tom Cruise?

Modelo de la película



Comparación

Compatibilidad

Componentes del Aprendizaje



Componentes del Aprendizaje

Créditos bancarios

- No hay una fórmula mágica para indicar si un crédito es aprobado o no.
- ¡Es un candidato para aprender de los datos!

Componentes del Aprendizaje

- Cada dato se representa como una variable x (*la información del usuario que solicita el crédito*).
- Cada posible resultado de cada dato x se representa como y .
- La fórmula que nos permite determinar si se aprueba un crédito o no:

$$f: \chi \rightarrow \gamma$$

donde

- χ representa el espacio de los datos de entrada x .
 - $x \in \chi$
- γ es el espacio de los resultados, en este caso sí (*es aprobado*) o no.
 - $y \in \gamma$

Componentes del Aprendizaje

El conjunto de datos \mathcal{D} recopila todos los datos x que tenemos a la mano, de la forma

$$(x_1, y_1), \dots, (x_n, y_n)$$

donde

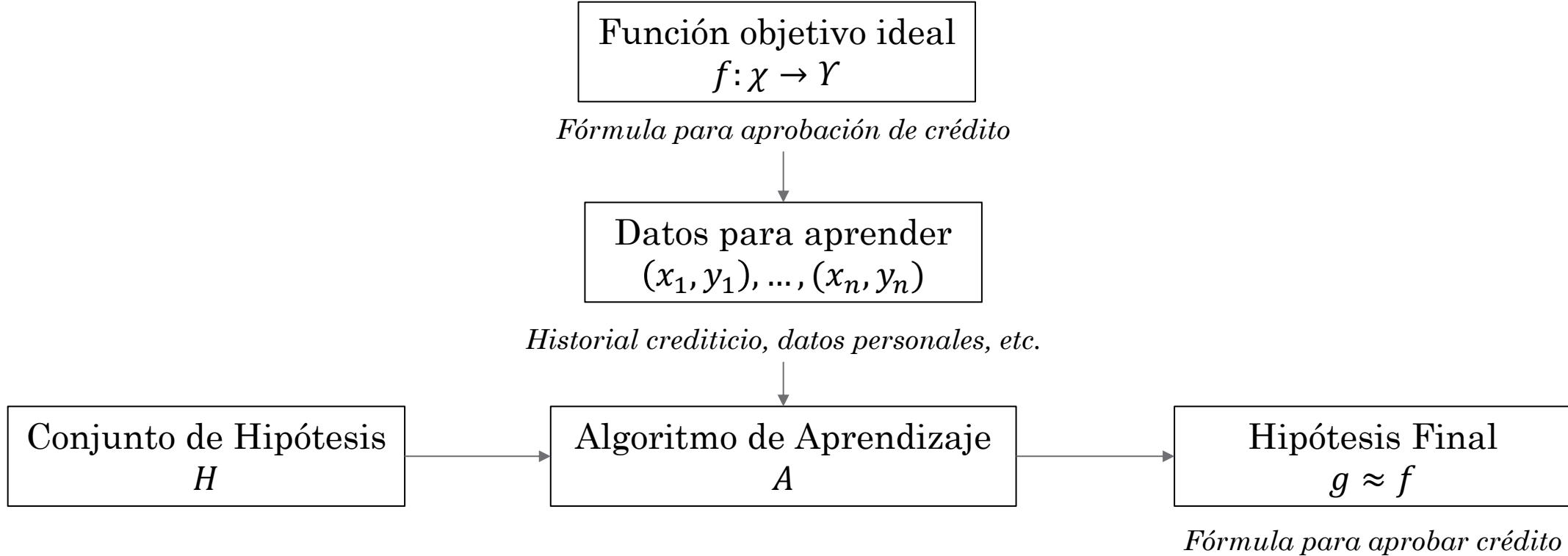
$$y_n = f(x_n)$$

para $n = 1, \dots, N$

Componentes del Aprendizaje

- En práctica, es imposible determinar f , por lo que la única opción es acercarnos a ella.
- H es el espacio de todas las posibles funciones o reglas que se acercan a f . Unas se pueden acercar más que otras.
- Para encontrar $g \approx f$, utilizamos un algoritmo o método de aprendizaje que nos permite utilizar los datos para aprender esa regla de clasificación.

Componentes del Aprendizaje





Componentes del Aprendizaje

Ejercicio #1:

Consideren el problema para determinar si un correo es spam o no.

1. ¿Cuáles son los datos de entrada? (X)
2. ¿Cuáles son las posibles salidas? (Y)
3. ¿Qué características debe tener el conjunto de datos?



Componentes del Aprendizaje

Ejercicio #2:

Consideren el problema para determinar un diagnóstico medico.

1. ¿Cuáles son los datos de entrada? (X)
2. ¿Cuáles son las posibles salidas? (Y)
3. ¿Qué características debe tener el conjunto de datos?

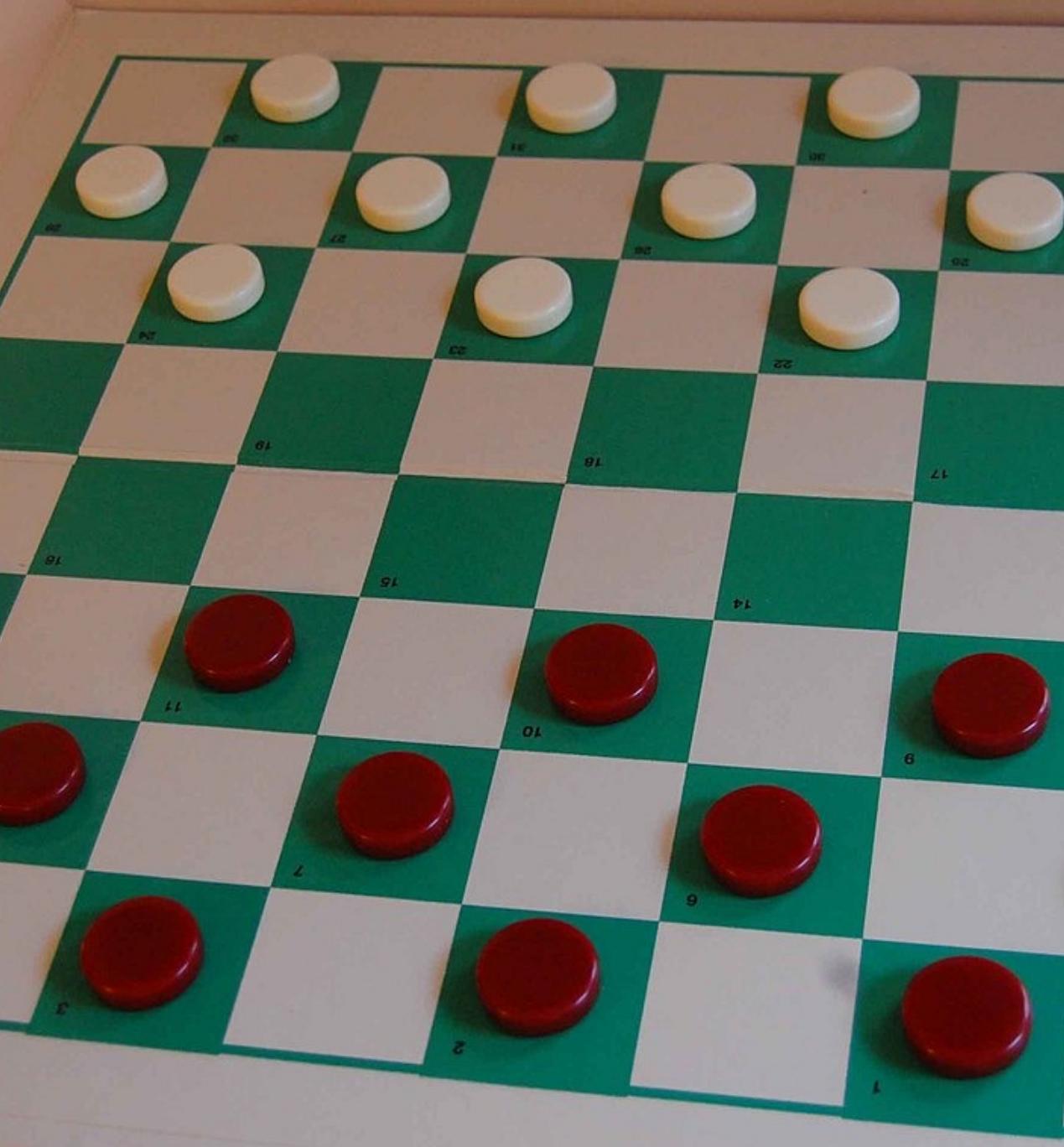


Componentes del Aprendizaje

Ejercicio #3:

Consideren el problema para determinar la polaridad de opinión en un mensaje:

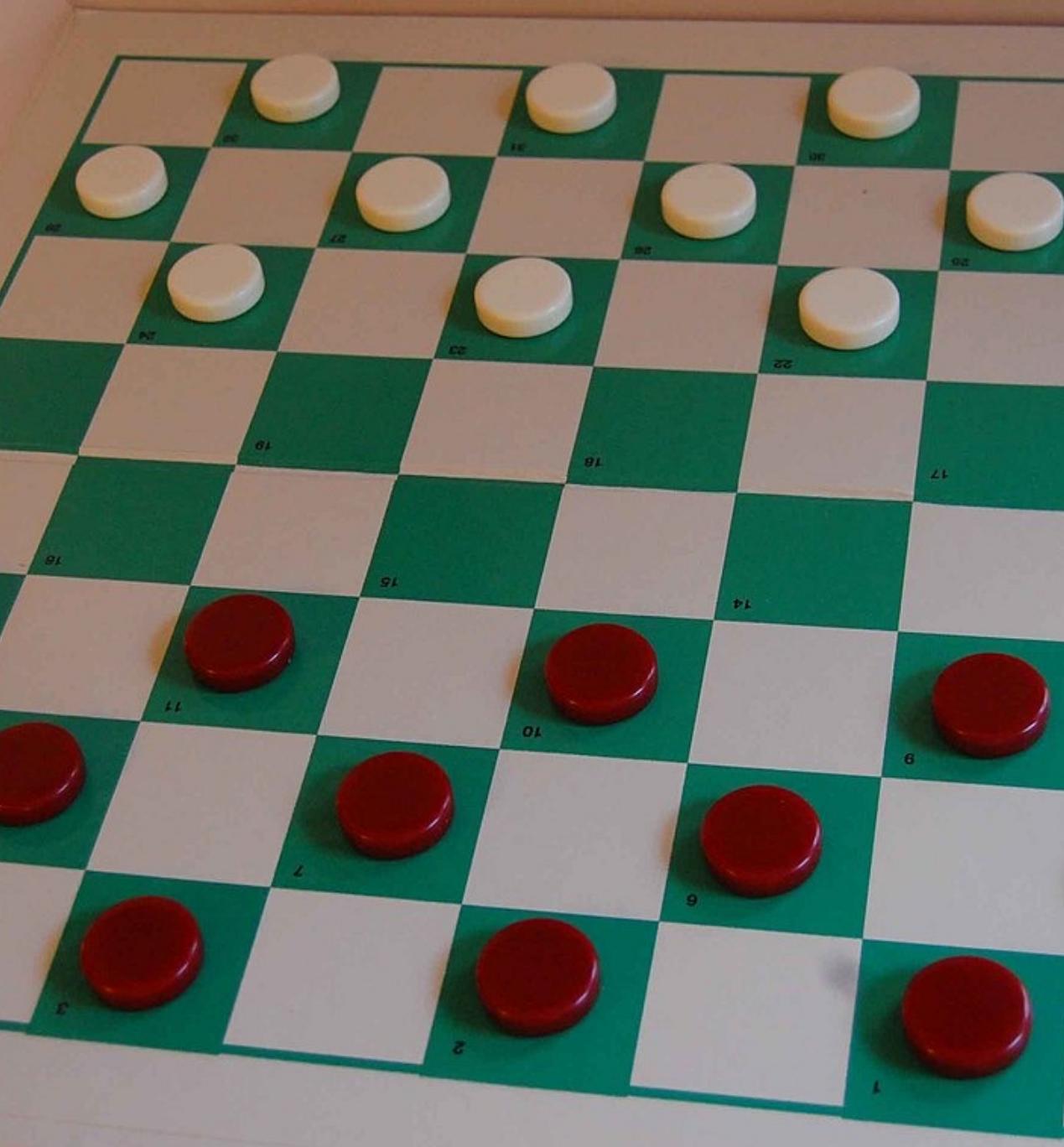
1. ¿Cuáles son los datos de entrada? (X)
2. ¿Cuáles son las posibles salidas? (Y)
3. ¿Qué características debe tener el conjunto de datos?



¿Qué es el Machine Learning? (informal)

Arthur Samuel (1959)

“Campo de estudio que permite que las computadoras sean capaces de aprender sin ser programadas explicitamente”



¿Qué es el Machine Learning? (informal)

- ¿Cómo mejoran su habilidad en un juego?
- Considerando la velocidad de aprendizaje, una computadora puede aprender más rápido que nosotros.
- Al final, puede resultar mejor que nosotros.



¿Qué es el Machine Learning? (formal)

Tom Mitchell (1998)

Un programa de computadora se dice que aprende de la experiencia E relacionada a una tarea T y una medida de rendimiento P , si su rendimiento en T , medida por P , mejora con la experiencia E .



¿Qué es Machine Learning?

Ejercicio #4:

Consideren el problema de jugar Checkers:

1. ¿Qué sería E ?
2. ¿Qué sería T ?
3. ¿Qué sería P ?



¿Qué es Machine Learning?

Ejercicio #5:

Consideren el problema de determinar si un correo es spam o no:

1. ¿Qué sería E ?
2. ¿Qué sería T ?
3. ¿Qué sería P ?

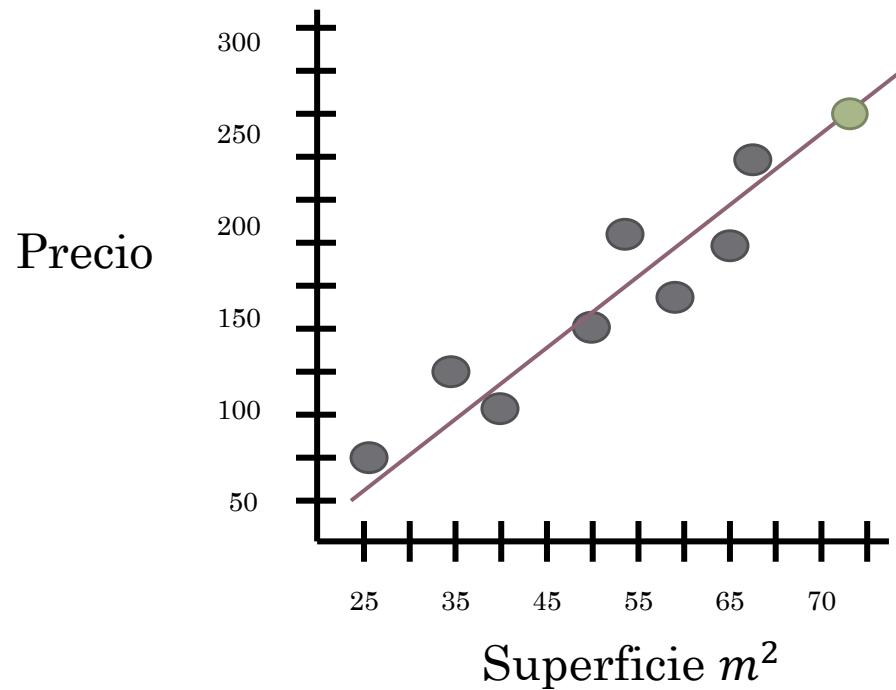
Tipos de Aprendizaje



Tipos de Aprendizaje

- La premisa de aprender de los datos es **utilizar observaciones para descubrir** qué es lo que sucede en un proceso.
- ¡Es muy amplio!

Aprendizaje Supervisado



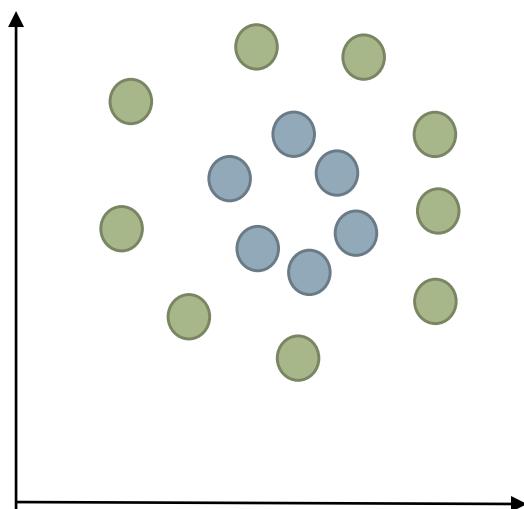
Este es un problema de **regresión**.

Supongamos que tenemos la siguiente información sobre el precio de la renta por metro cuadrado en una zona de la CDMX.

¿Cómo se podría determinar un nuevo valor considerando estos datos?

En este caso, estamos aprendiendo de los datos que tienen las «**respuestas correctas**».

Aprendizaje Supervisado



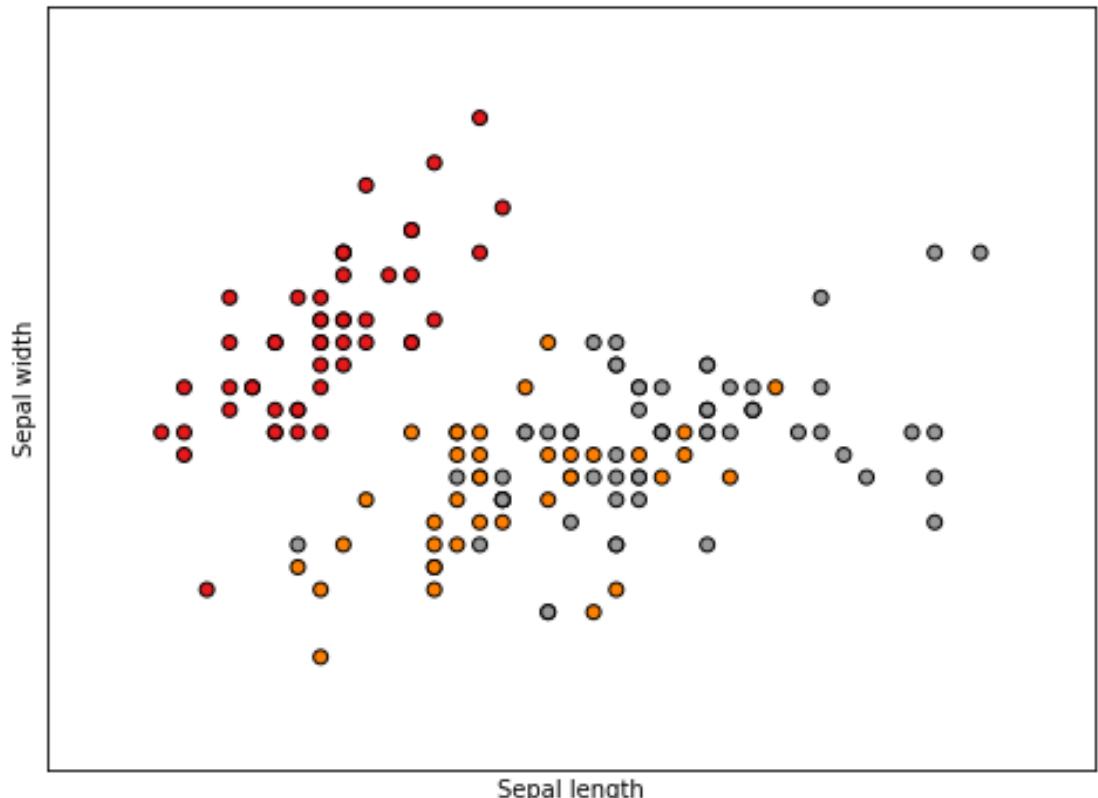
Consideremos ahora un problema de **clasificación**.

Se tienen dos (o más) clases de objetos a los cuales pertenecen cada elemento del conjunto de datos.

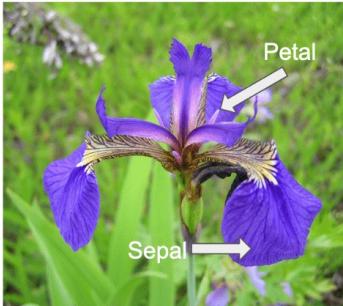
Usualmente se etiquetan con valores numéricos:

- 1 y 0
- 1 y -1

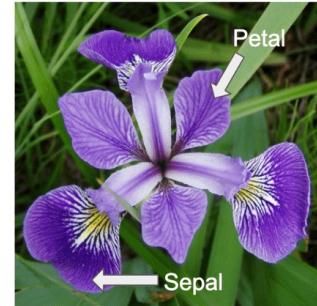
Aprendizaje Supervisado



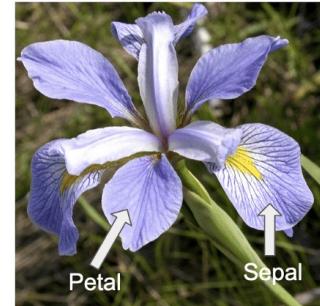
Iris setosa



Iris versicolor



Iris virginica



Cuatro características:

- Ancho y largo del pétalo
- Ancho y largo del sépalo



¿Qué es Machine Learning?

Ejercicio #6:

Consideren el problema de determinar si un correo es spam o no:

1. ¿Es un problema de regresión o clasificación?
2. ¿Cuáles serían las clases?



¿Qué es Machine Learning?

Ejercicio #7:

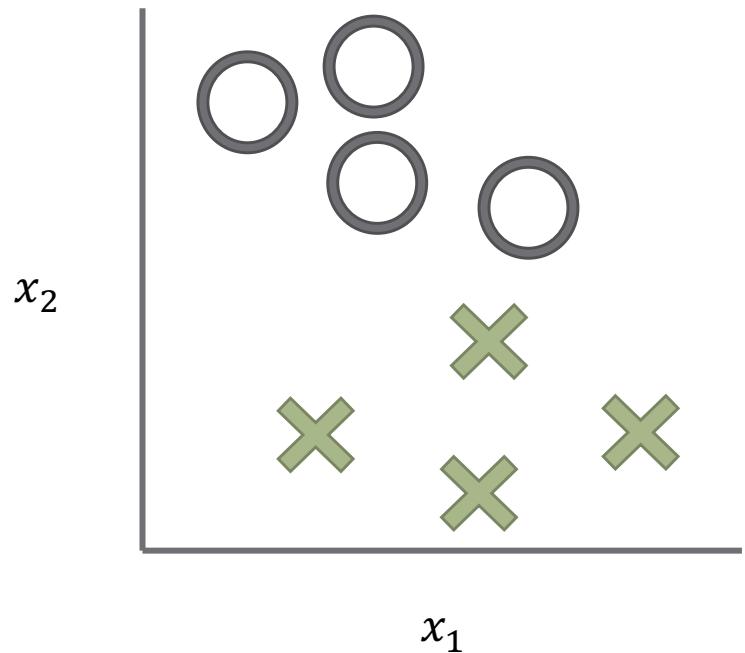
Consideren el problema de determinar el precio de un activo financiero:

1. ¿Es un problema de regresión o clasificación?
2. ¿Cuáles serían los valores posibles para los precios?

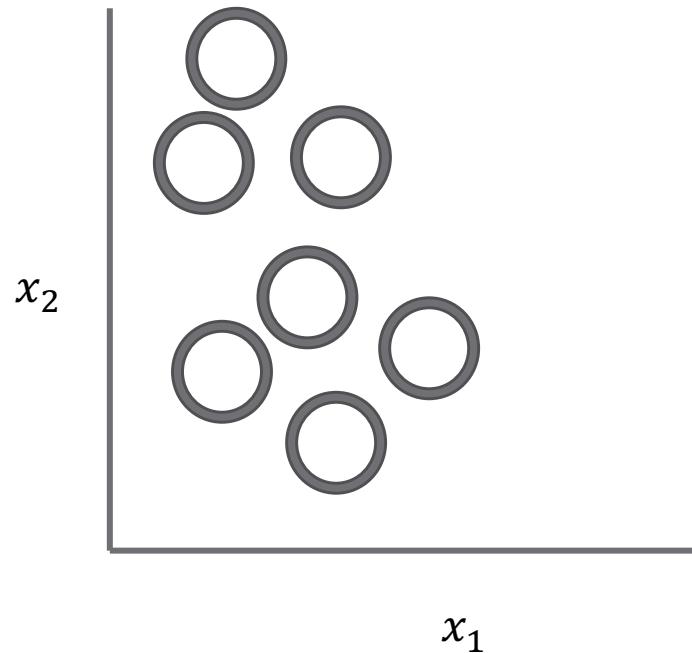


Aprendizaje
Supervisado

Aprendizaje No Supervisado



Aprendizaje Supervisado



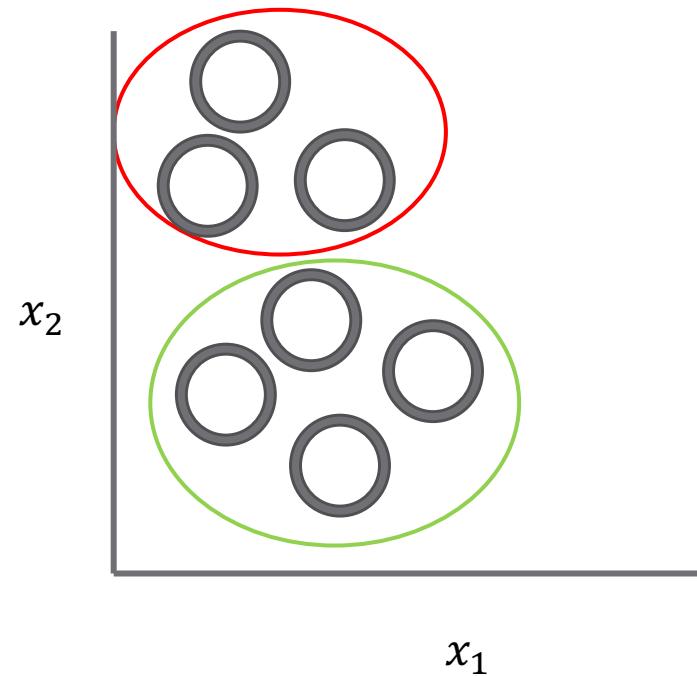
Aprendizaje No Supervisado

Aprendizaje No Supervisado

- En el aprendizaje no supervisado no se dan las clases o valores correctos de los datos.
- ¿Por qué? No siempre es posible determinar el número de clases de antemano, o es caro o difícil determinarlas.
- Aquí la tarea es encontrar estructuras o patrones en los datos.

Aprendizaje No Supervisado

- En el aprendizaje no supervisado no se dan las clases o valores correctos de los datos.
- ¿Por qué? No siempre es posible determinar el número de clases de antemano, o es caro o difícil determinarlas.
- Aquí la tarea es encontrar estructuras o patrones en los datos.



Aprendizaje No Supervisado

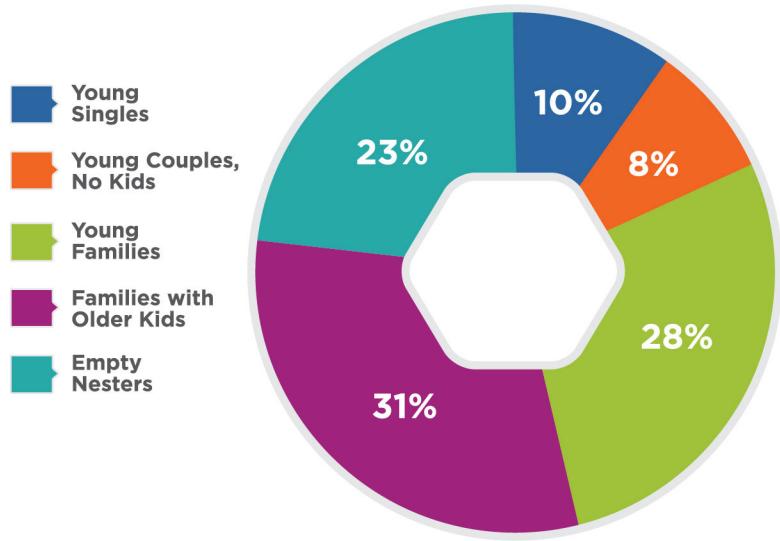
Una tarea común es en la clasificación de noticias:

- No sabemos en cuantas clases separar cada noticias. E.g., deportes, sociales, nacional, internacional, etc.
- Una forma es determinar clústeres por medio de similitud en cuanto a temas o palabras.

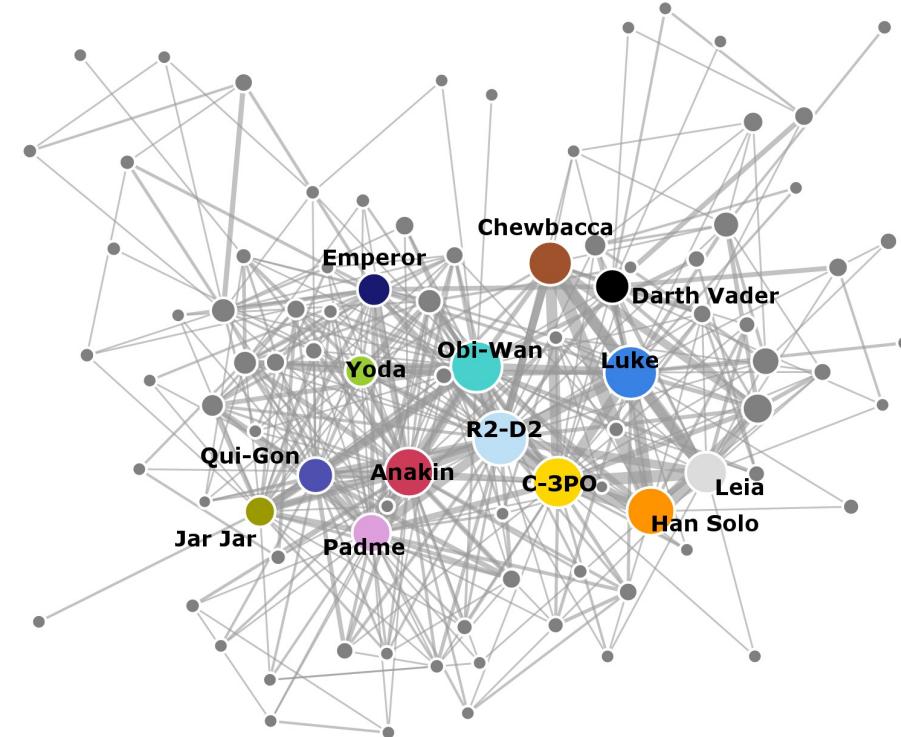
The screenshot shows a Google search results page for the query "Falcon Heene". The search bar at the top contains the text "Falcon Heene". Below the search bar are three buttons: "Buscar en Noticias" (Search News), "Buscar en la Web" (Search the Web), and "Búsqueda avanzada de noticias" (Advanced News Search). The main content area is titled "Noticias" (News) and displays 11 - 29 de aproximadamente 1 resultados (Results 11 - 29 of approximately 1). The first result is a news article from "La Gaceta Tucumán" with the headline "Lo creían atrapado en un globo y estaba en el garaje de su casa" (They believed he was trapped in a balloon and was in the garage of his house). The second result is a news article from "True/Slant" with the headline "Balloon Boy Falcon Heene Farts on Larry King" (Balloon boy Falcon Heene farts on Larry King). The third result is a news article from "Christian Science Monitor" with the headline "Balloon boy hoax rumours as Falcon Heene tells CNN 'we did this for a show'" (Balloon boy hoax rumours as Falcon Heene tells CNN 'we did this for a show'). There is also a section titled "AP News in Brief" which includes a headline from "New York Times" about a boy found in a balloon. The sidebar on the left provides navigation links for news categories like "Cualquier contenido" (Any content), "Cualquier noticia" (Any news), and "Ordenados por relevancia" (Sorted by relevance).

Aprendizaje No Supervisado

SAMPLE MARKET SEGMENTATION:
FAMILY LIFE STAGE



[Esta foto](#) de Autor desconocido está bajo licencia [CC BY](#)

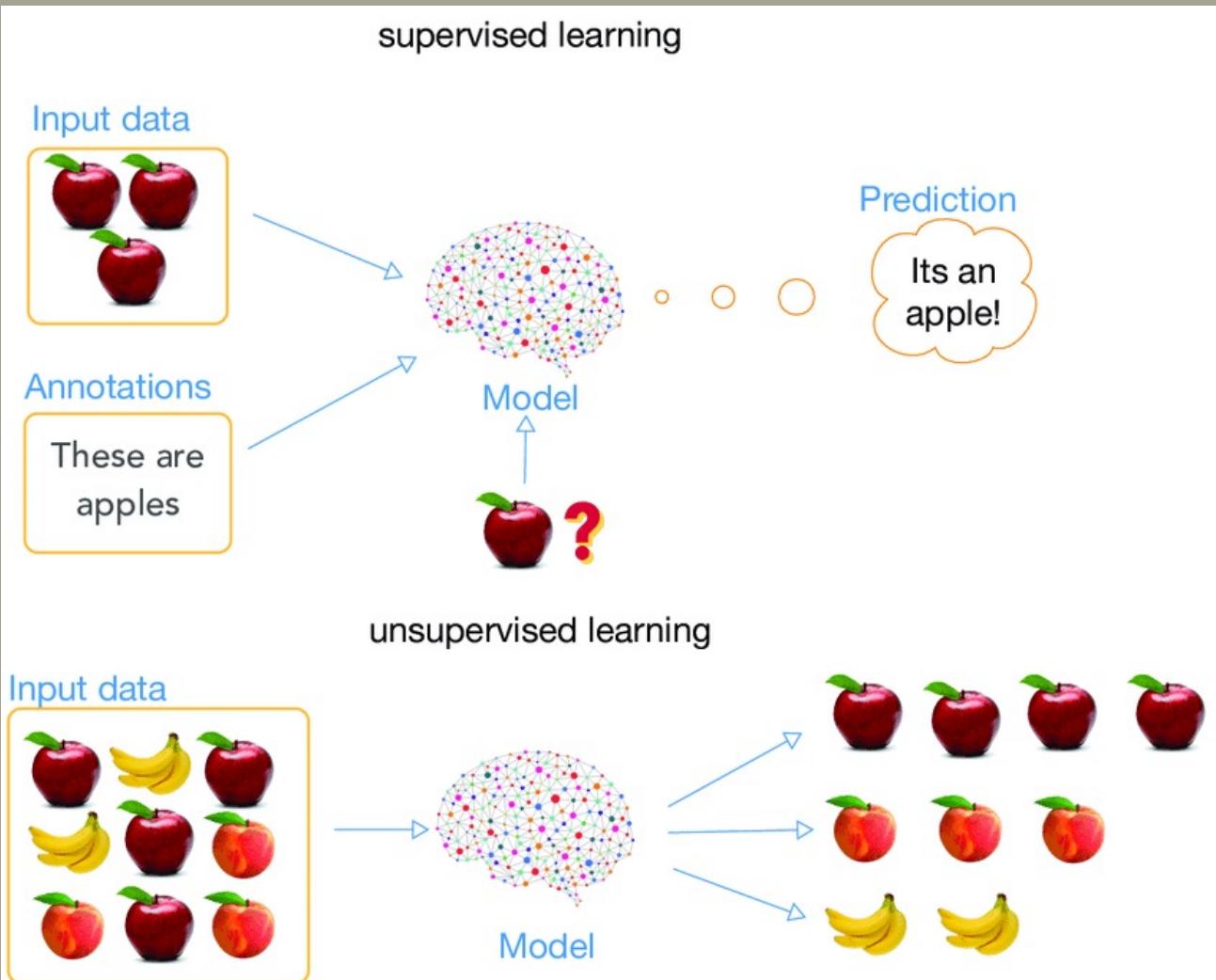


[Esta foto](#) de Autor desconocido está bajo licencia [CC BY-SA](#)

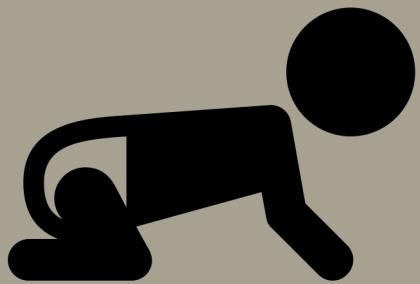
Aprendizaje Supervisado

VS

Aprendizaje No Supervisado



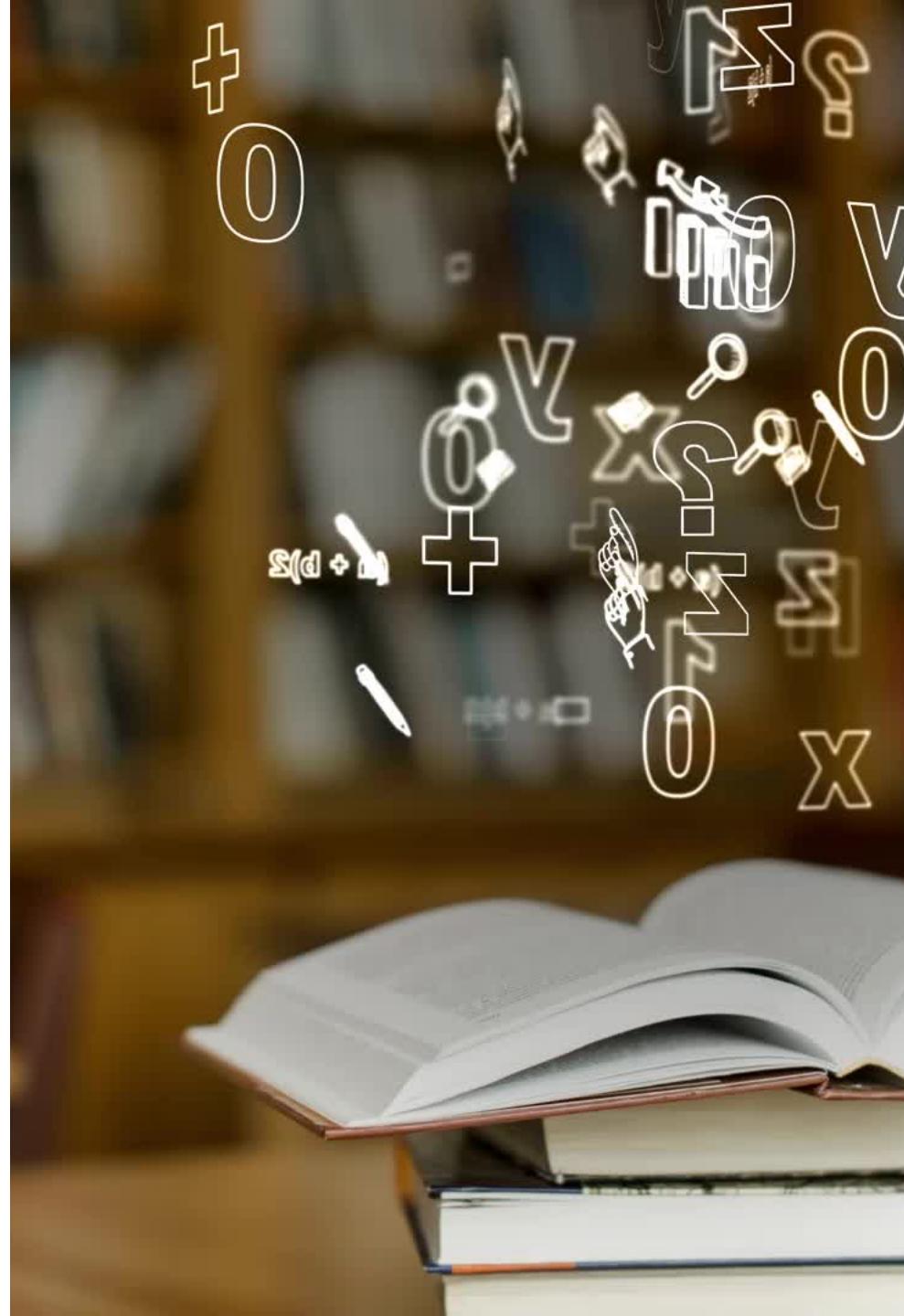
Aprendizaje
por Refuerzo



Machine Learning

Machine Learning

- Es mejor el **aprendizaje supervisado** (dar ejemplos de datos y su valor).
- Esto requiere que **se armen conjuntos de datos “grandes” y su anotación**, usualmente de forma manual.
- El científico de datos **debe proponer las características** para ayudar a dar forma a las reglas que permiten descifrar la estructura de cada dato.
 - Debe ser experto en el tema o área de aplicación.
 - Trabajar en conjunto con un experto.
- El Machine Learning **consta de métodos de aprendizaje** no tan complicados como el Deep Learning.
- Aprender es **optimizar**.



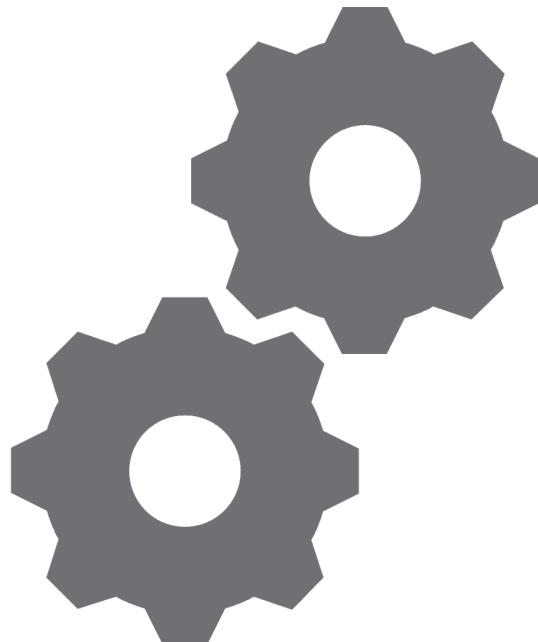
Algoritmos de Aprendizaje Supervisado

- Regresión Lineal,
Polinomial y Logística
- Árboles de Decisión
- Redes Neuronales
(Perceptrón Multicapa)



Algoritmos de Aprendizaje No Supervisado

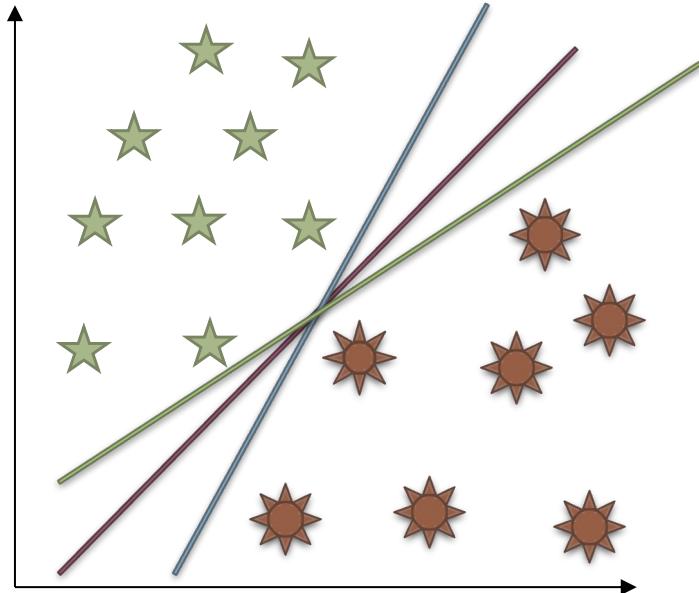
- Métodos de clústering
 - K-Nearest Neighbourghs
 - K-Means
 - Gaussian Mixtures



Tarea

- Investigar sobre un problema de aprendizaje automático donde se aplique aprendizaje no supervisado como parte de su solución.
- Investigar sobre un problema de aprendizaje automático donde se aplique aprendizaje supervisado como parte de su solución.
- Investigar sobre qué trata el algoritmo Cocktail Party.

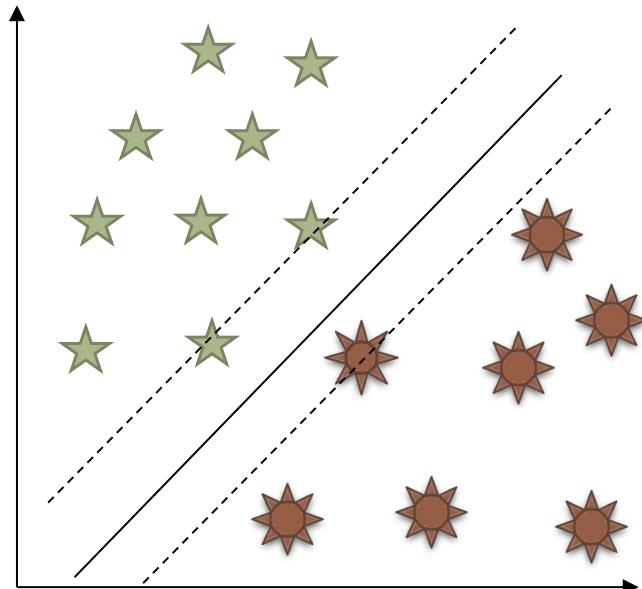
Máquina de Vectores de Soporte



¿De cuántas formas se puede separar los objetos de la figura?

¿Cuál es la correcta?

Máquina de Vectores de Soporte

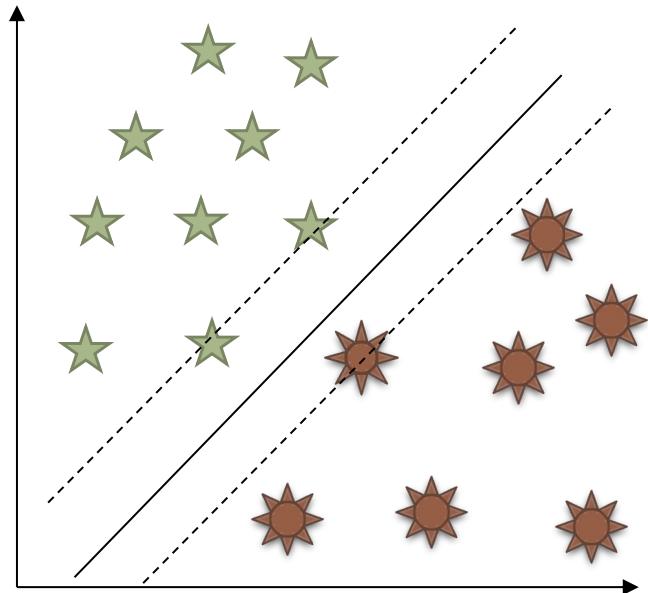


Idea

- Utilizar algunos puntos para generar márgenes de apoyo.
- Hacer ese margen lo más amplio posible.
- Dibujar la recta de separación en medio.

Máquina de Vectores de Soporte

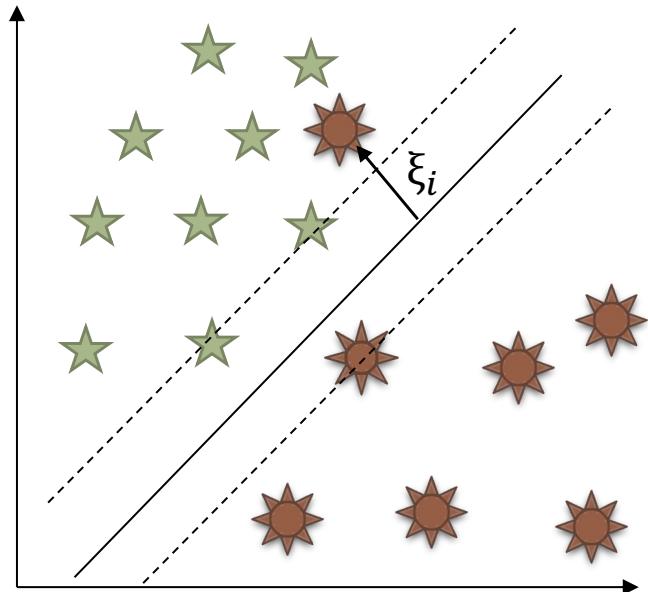
Formulación



$$\min \frac{1}{2} \|w\|^2$$

$$\text{s. a. } y_i(w^T \cdot x_i + b) \geq 1$$

Máquina de Vectores de Soporte

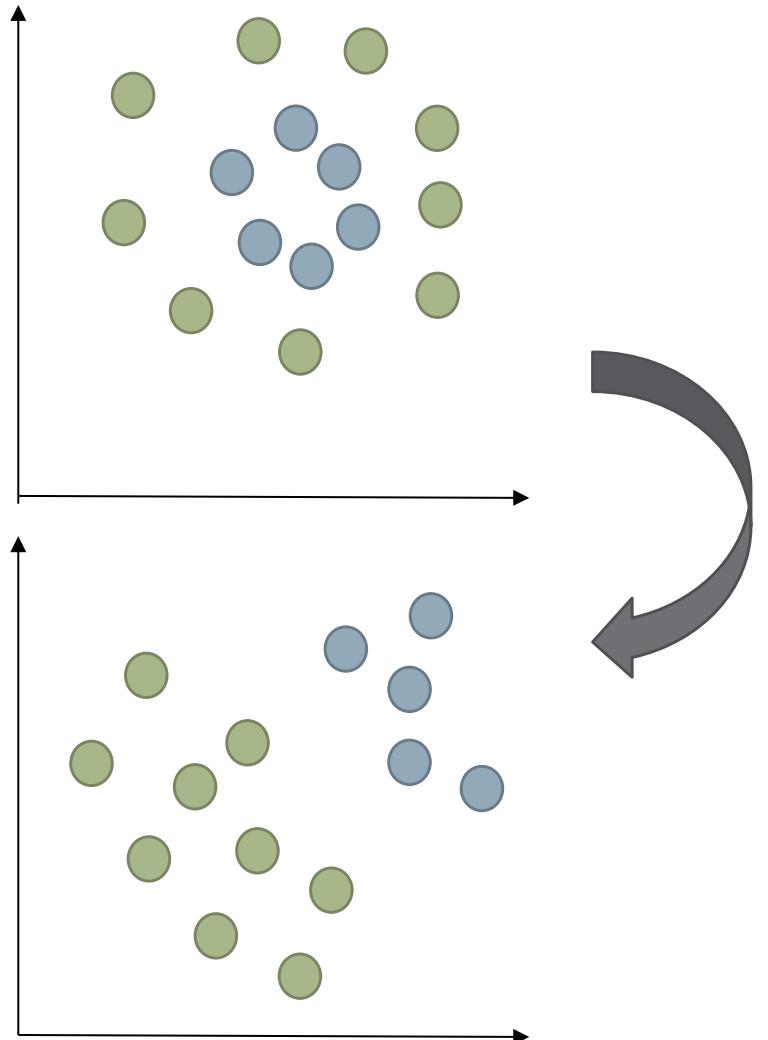


Margen Suave

$$\min \frac{1}{2} \|w\|^2 + c \sum \xi_i$$

s. a. $y_i(w^T \cdot x_i + b) \geq 1 - \xi_i$

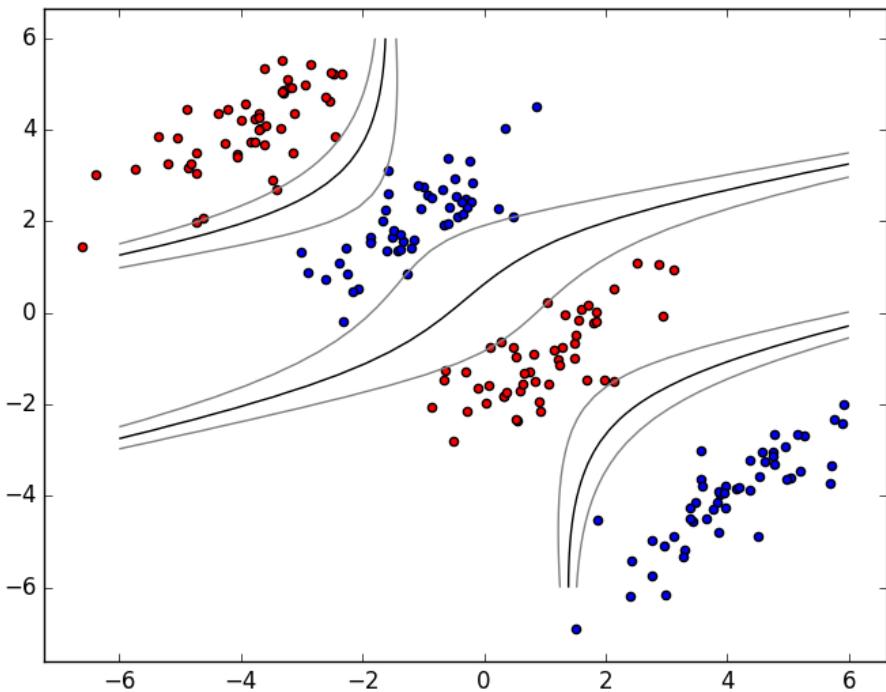
Máquina de Vectores de Soporte



Truco del Kernel

¿Qué pasa cuando los datos tiene formas más complejas? Una recta o hiperplano no las puede separar perfectamente.

Máquina de Vectores de Soporte



Truco del Kernel

Permite dibujar funciones de separación con formas más complejas.

Naïve Bayes

- Es una técnica de clasificación que se basa en el teorema de Bayes con el supuesto de que todas las características que predicen el valor objetivo son independientes entre sí.
- Calcula la probabilidad de cada clase del conjunto Y y luego elige la que tiene mayor probabilidad para cada dato x_i .
- Funciona particularmente bien con problemas de procesamiento del lenguaje natural.

$$y = \operatorname{argmax}_y P(y) \prod_{i=1}^n P(x_i|y_i)$$