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?

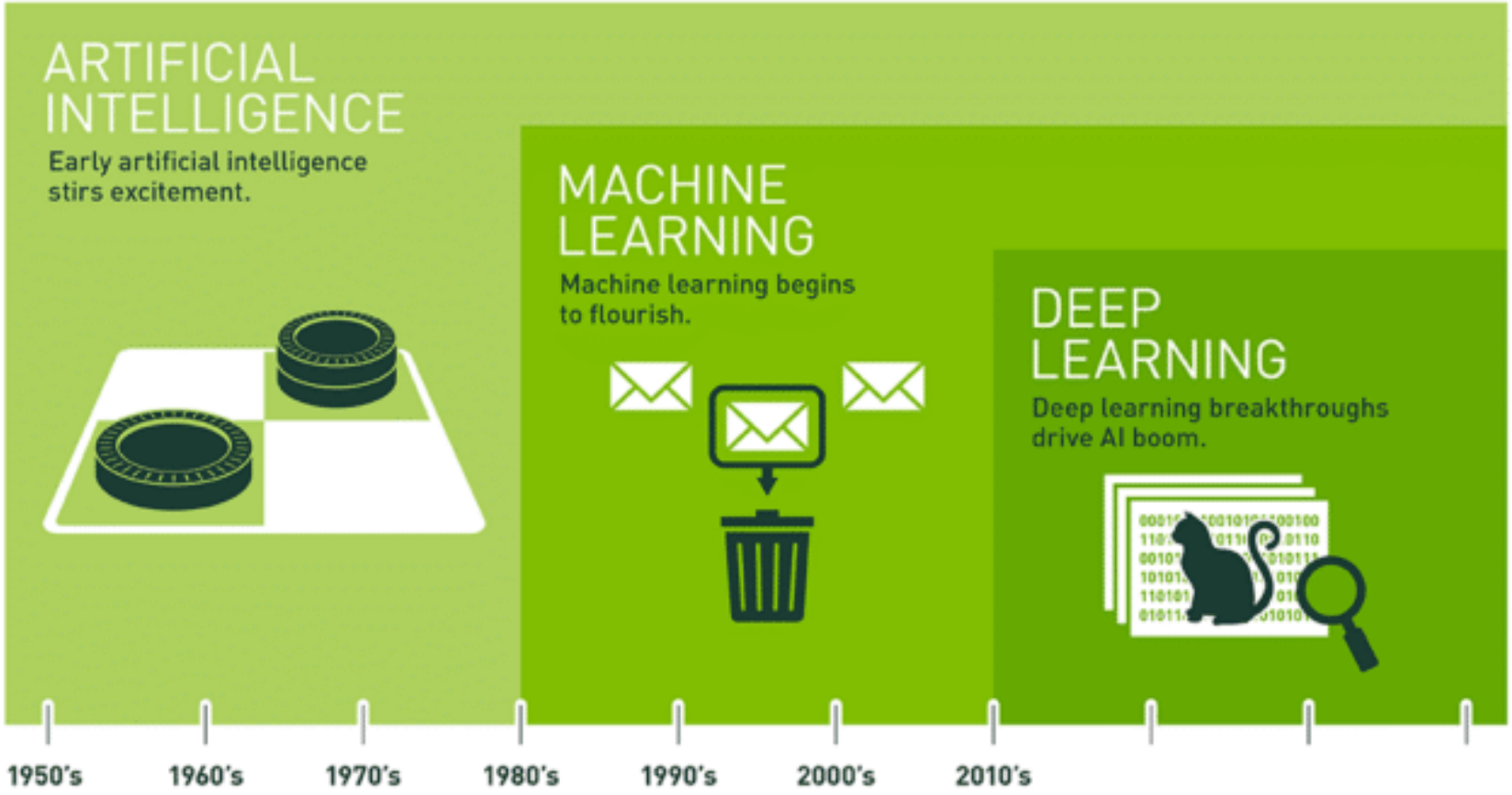
El Problema del Aprendizaje

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

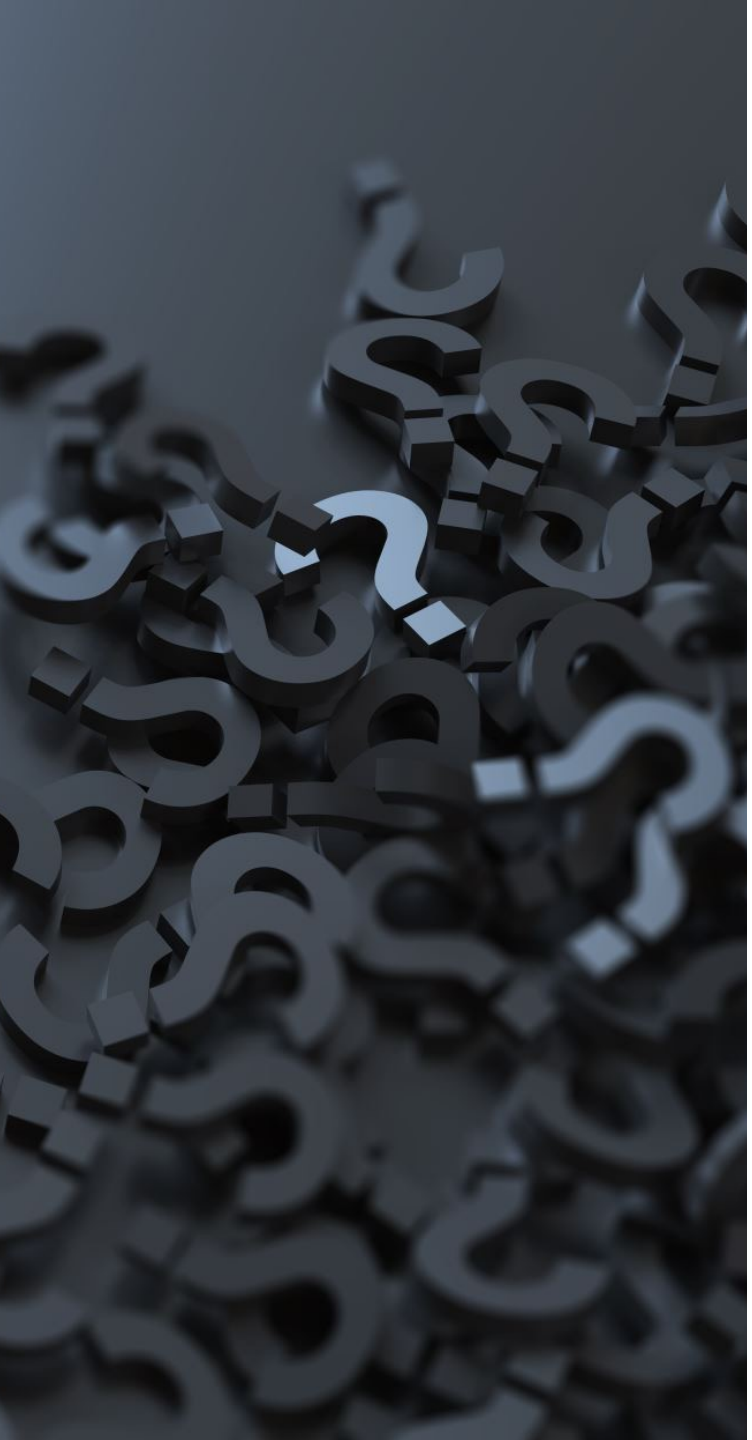
El Problema del Aprendizaje

- 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

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:

$$f: \chi \rightarrow \Upsilon$$

donde

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

Componentes del Aprendizaje

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

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

donde

$$y_n = f(x_n)$$

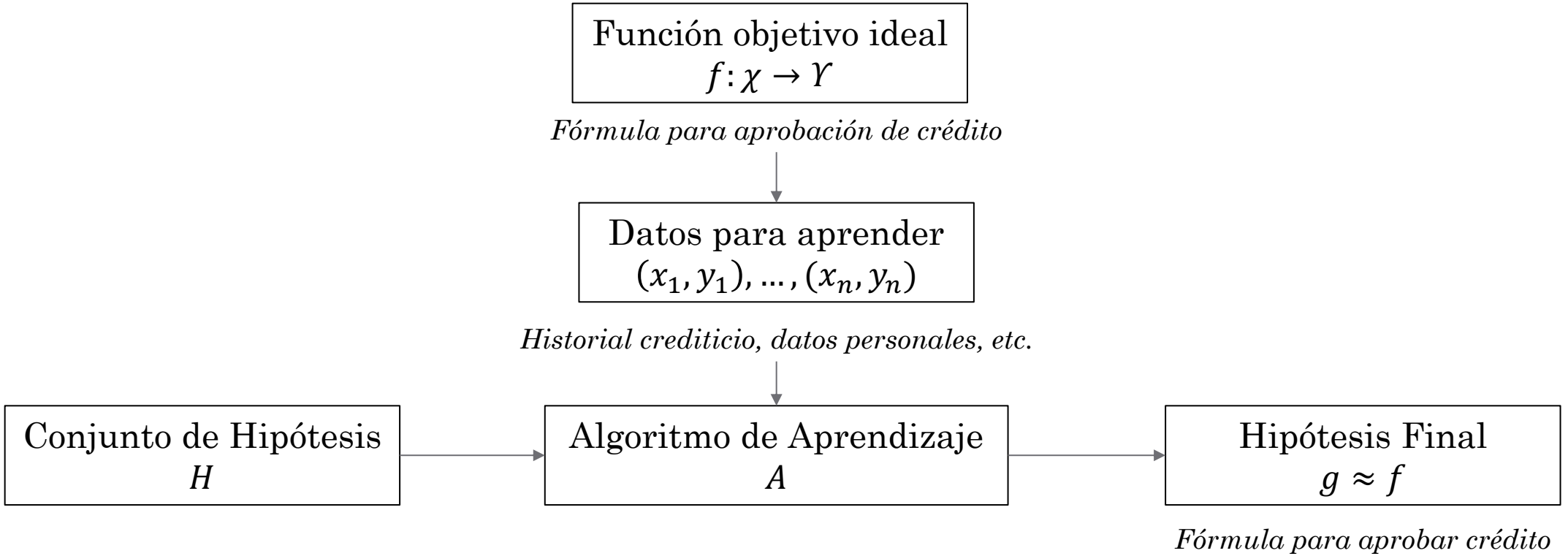
para $n = 1, \ldots, N$

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.

1ª Versión

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? (χ)
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.

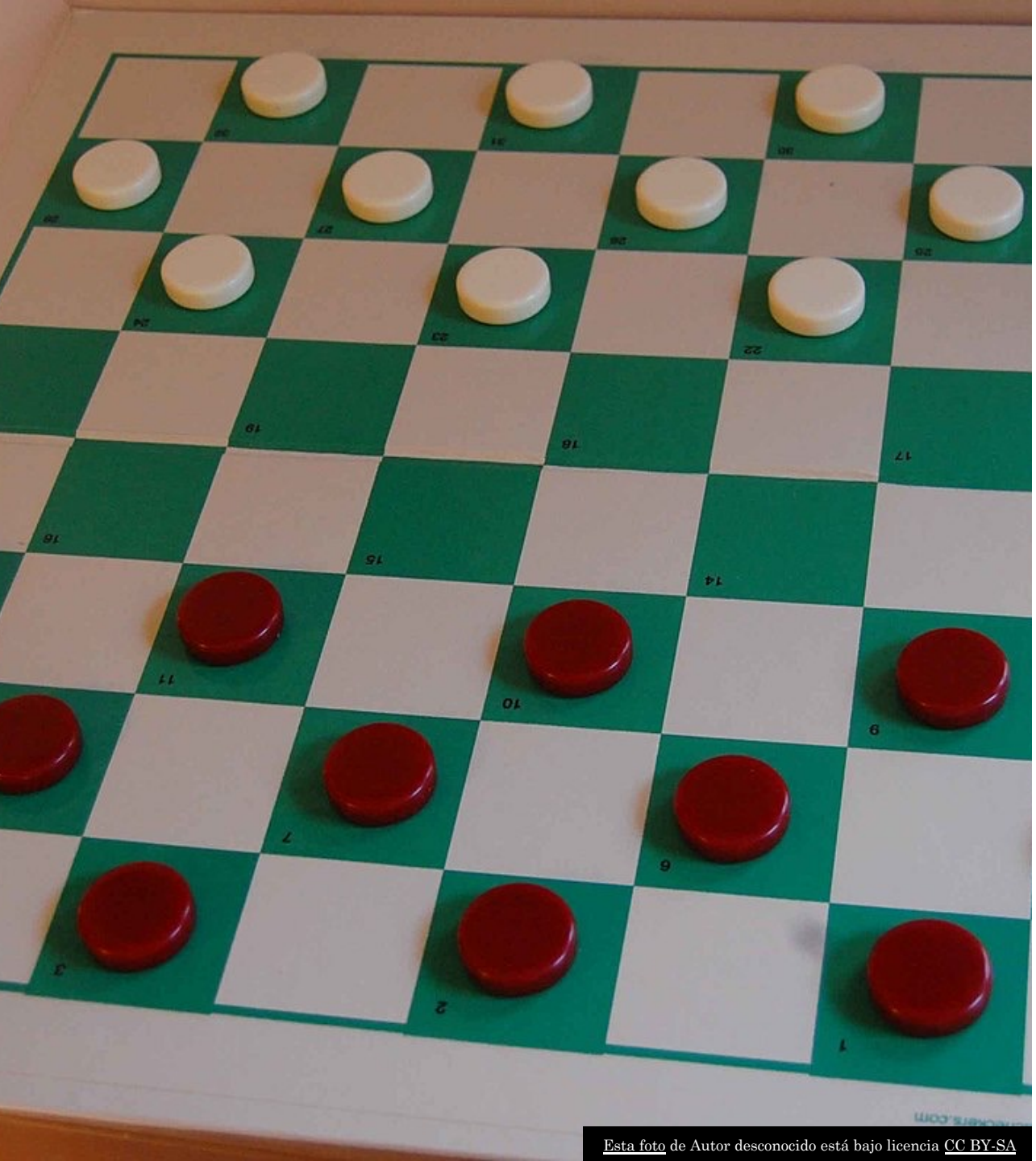
1. ¿Cuáles son los datos de entrada? (χ)
2. ¿Cuáles son las posibles salidas? (Y)
3. ¿Qué características debe tener el conjunto de datos?

Componentes del Aprendizaje

Ejercicio #3:

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

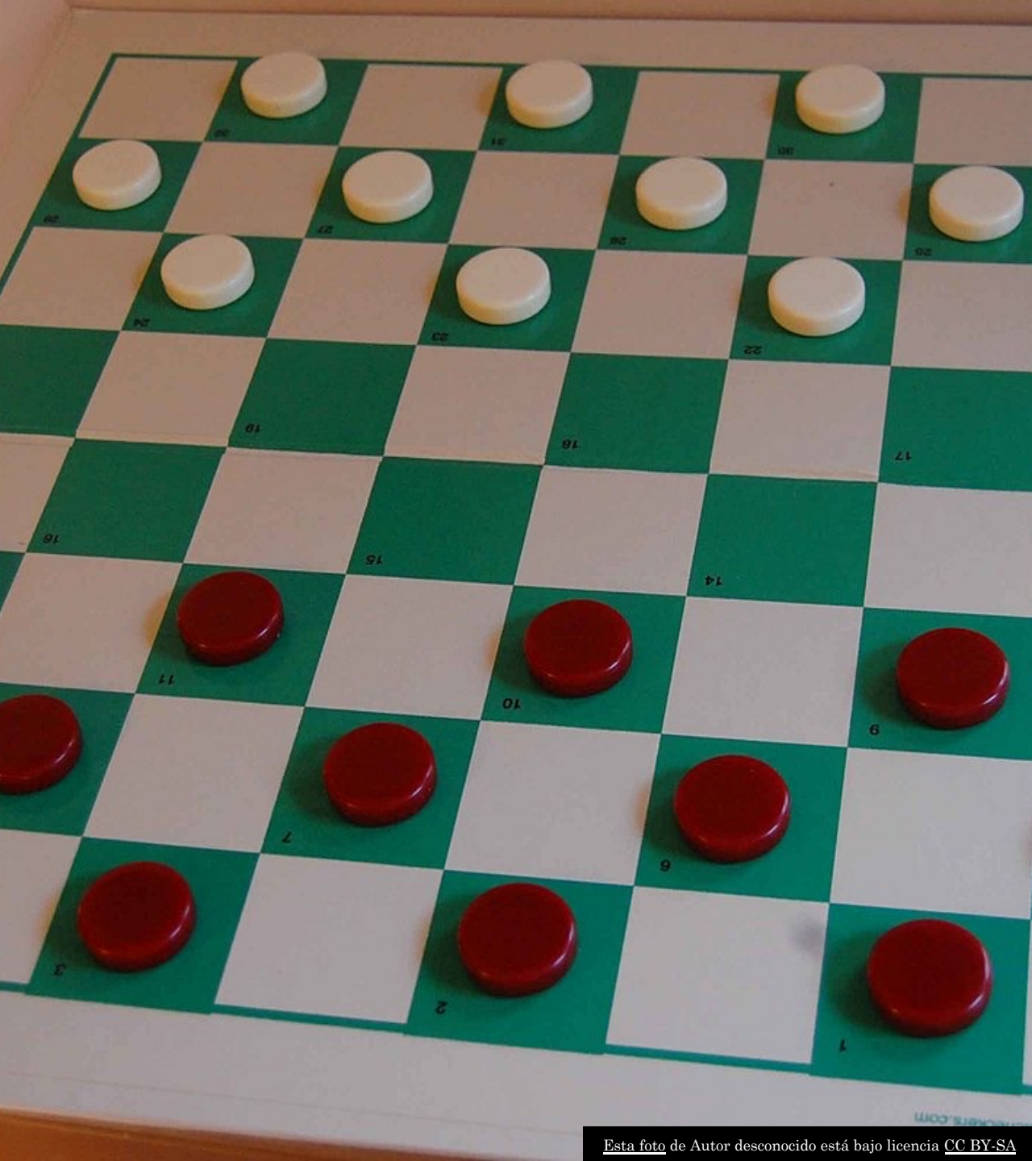
1. ¿Cuáles son los datos de entrada? (χ)
2. ¿Cuáles son las posibles salidas? (Y)
3. ¿Qué características debe tener el conjunto de datos?



¿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 experienca *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:

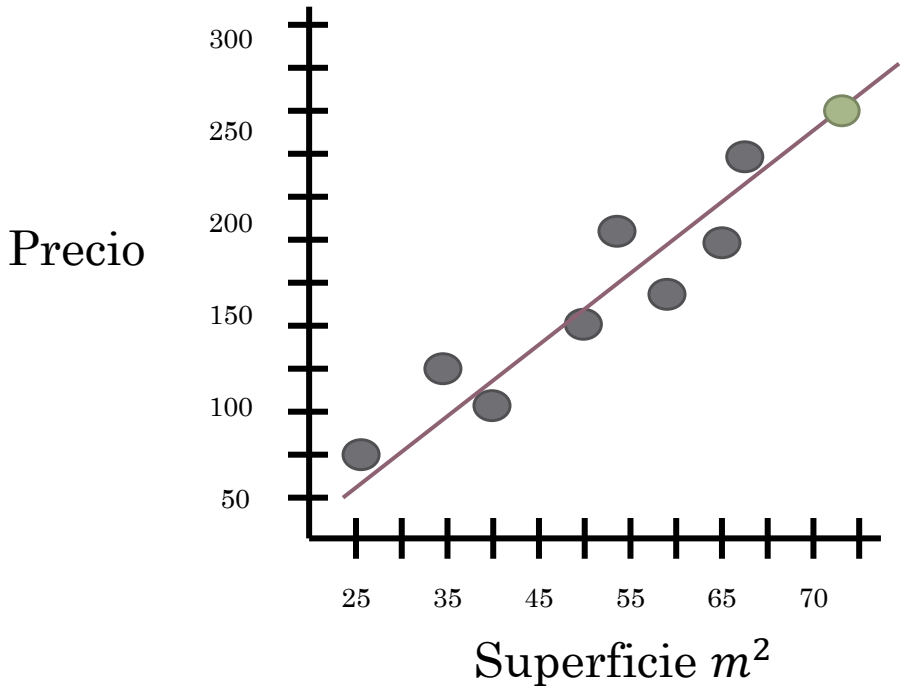
1. ¿Qué sería E?
2. ¿Qué sería T?
3. ¿Qué sería P?

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



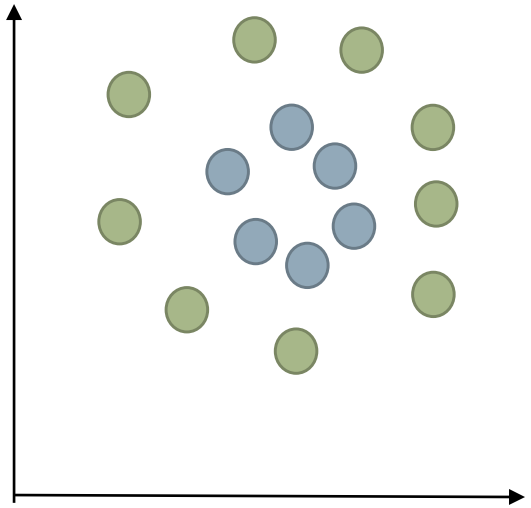
Supongamos que tenemos la siguiente información sobre el precio de la renta por metro cuadrado en una zona de la CDMX.

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

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

Este es un problema de **regresión.**

Aprendizaje Supervisado



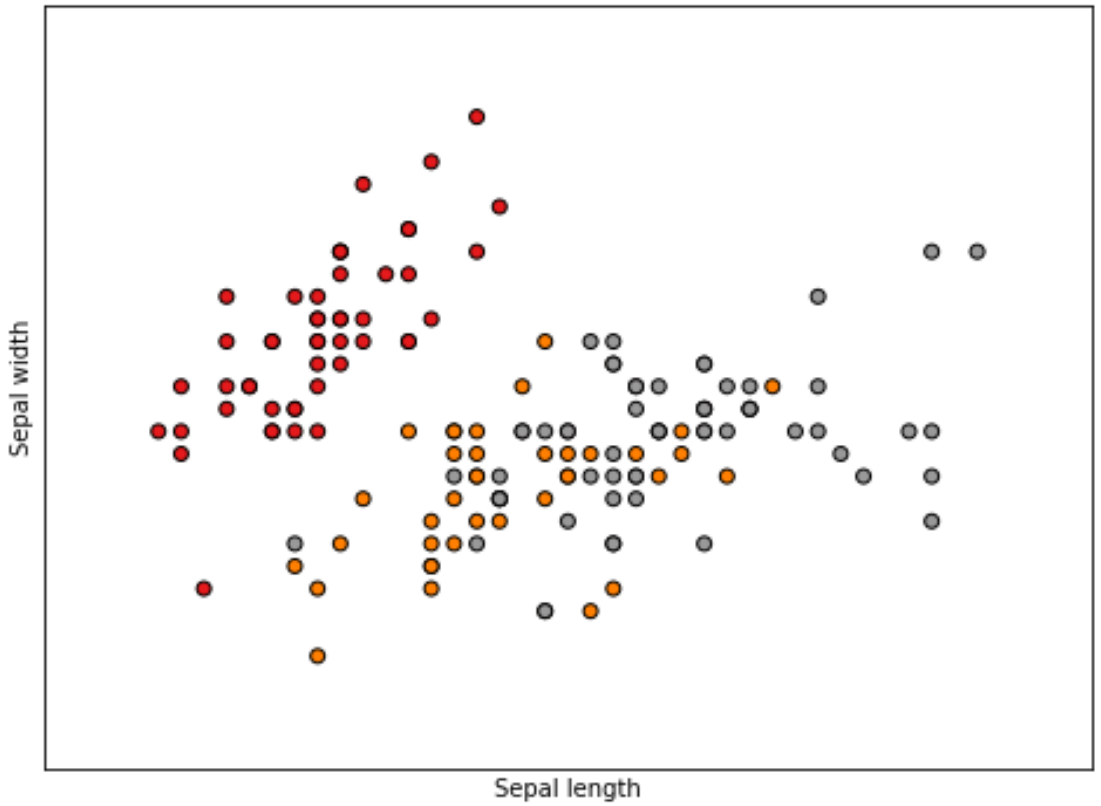
Consideremos ahora un problema de **clasificación**.

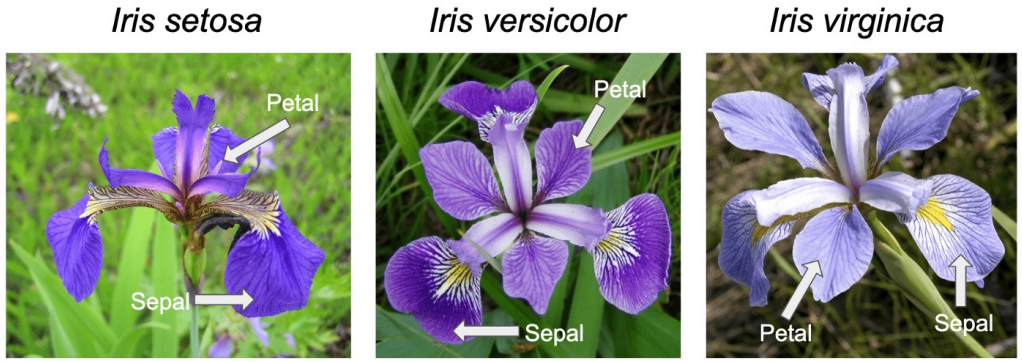
Se tienen dos (o más) clases de objetos a los cuales pertenecen cada elemento del conjunto de datos.

Usualmente se etiquetan con valores numéricos:

- 1 y 0
- 1 y -1

Aprendizaje Supervisado





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?

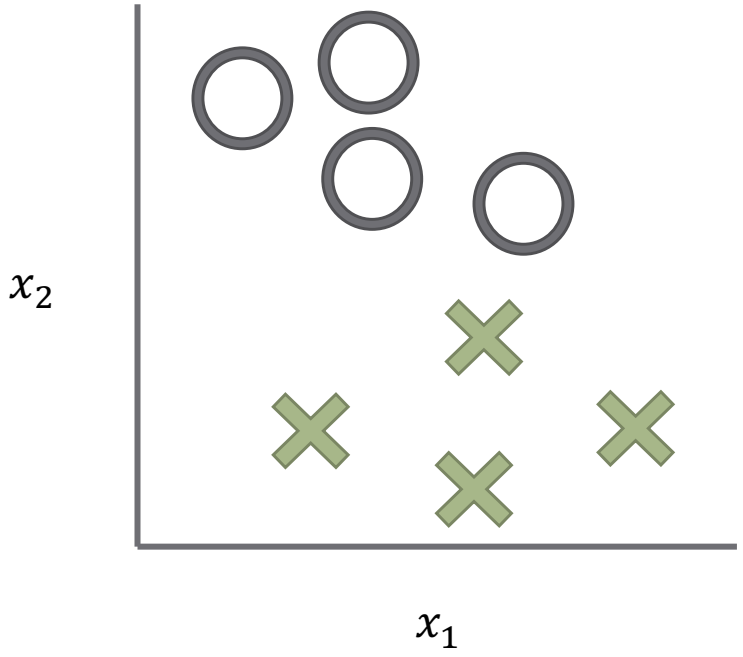
Ejercicio #7:

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

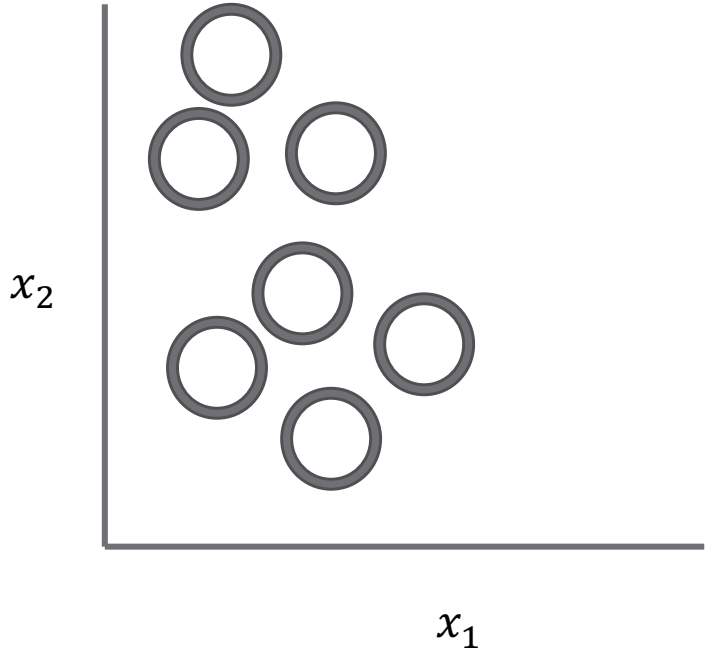
1. ¿Es un problema de regresión o clasificación?
2. ¿Cuáles serían los valores posibles para los precios?

Aprendizaje Supervisado

Aprendizaje No Supervisado



Aprendizaje Supervisado



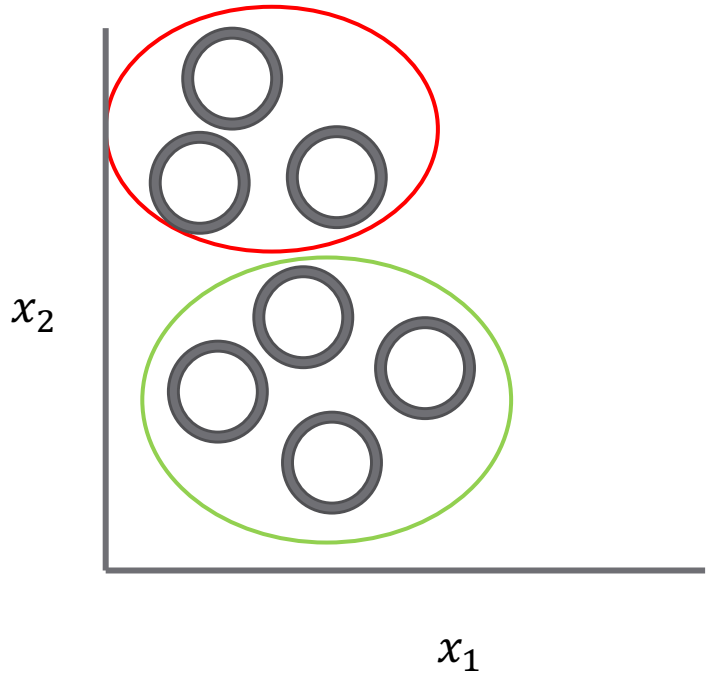
Aprendizaje No Supervisado

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

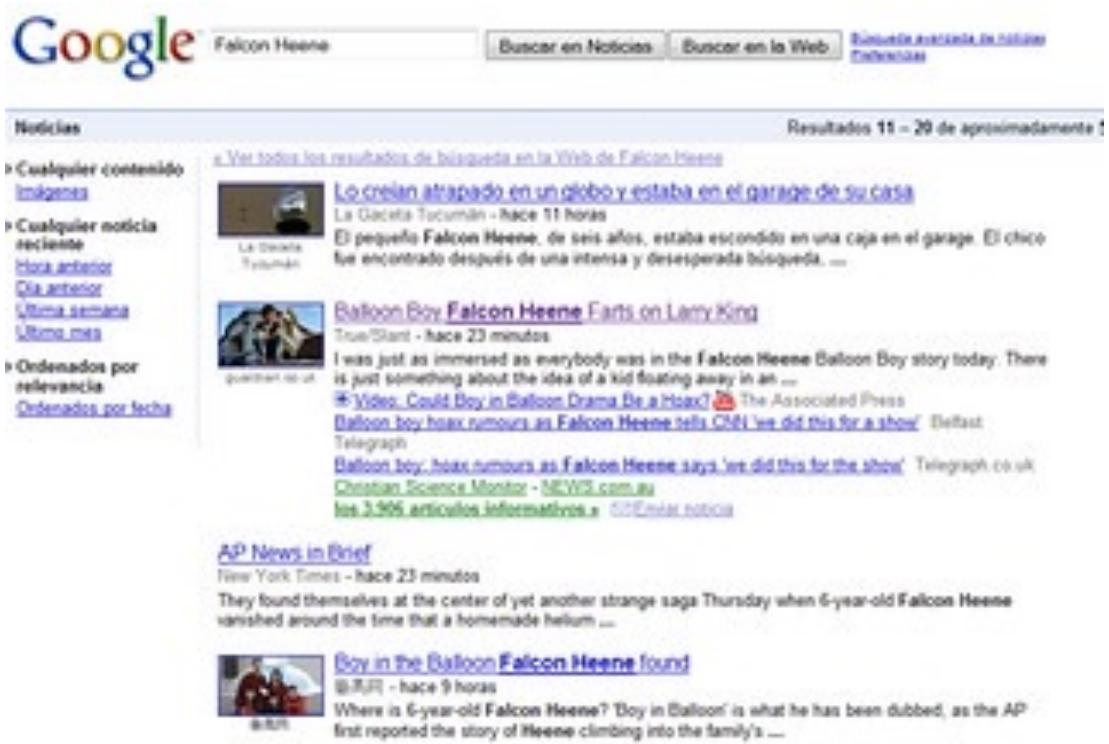
- En el aprendizaje no supervisado no se dan las clases o valores correctos de los datos.
- ¿Por qué? No siempre es posible determinar el número de clases de antemano, o es caro o difícil determinarlas.
- Aquí la tarea es encontrar estructuras o patrones en los datos.



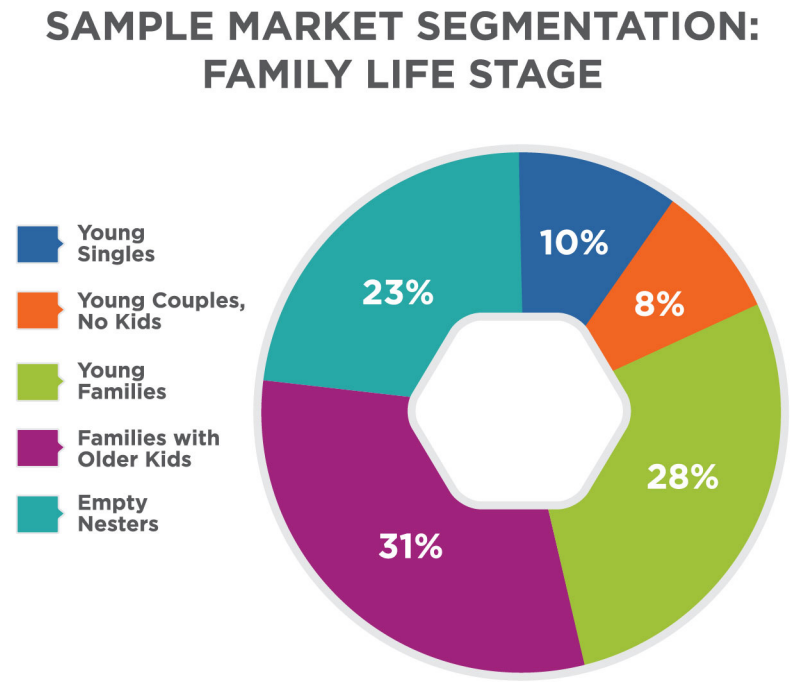
Aprendizaje No Supervisado

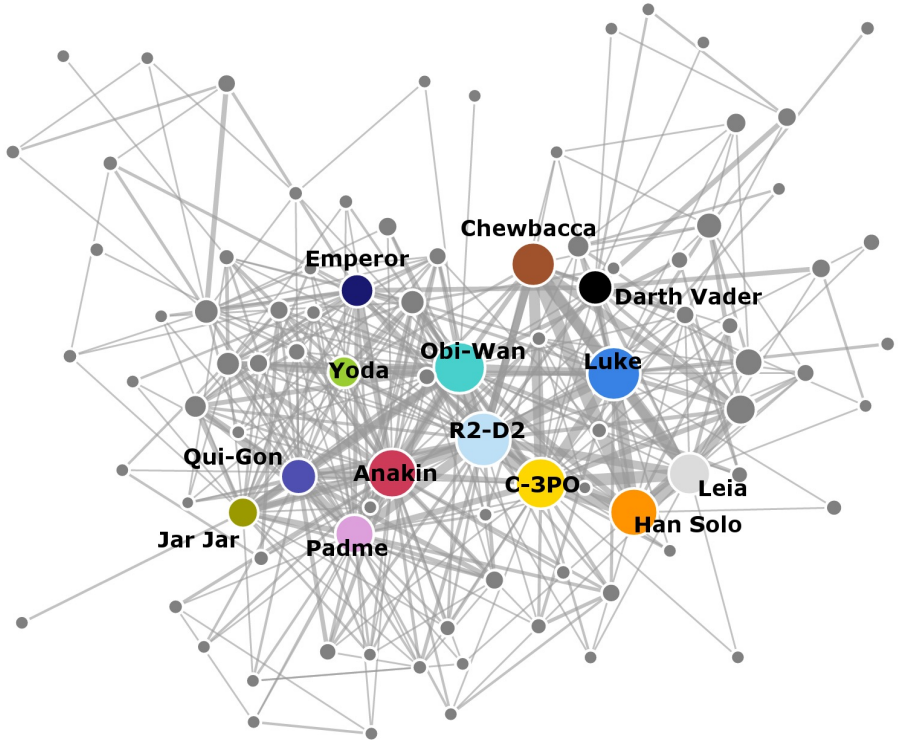
Una tarea común es en la clasificación de noticias:

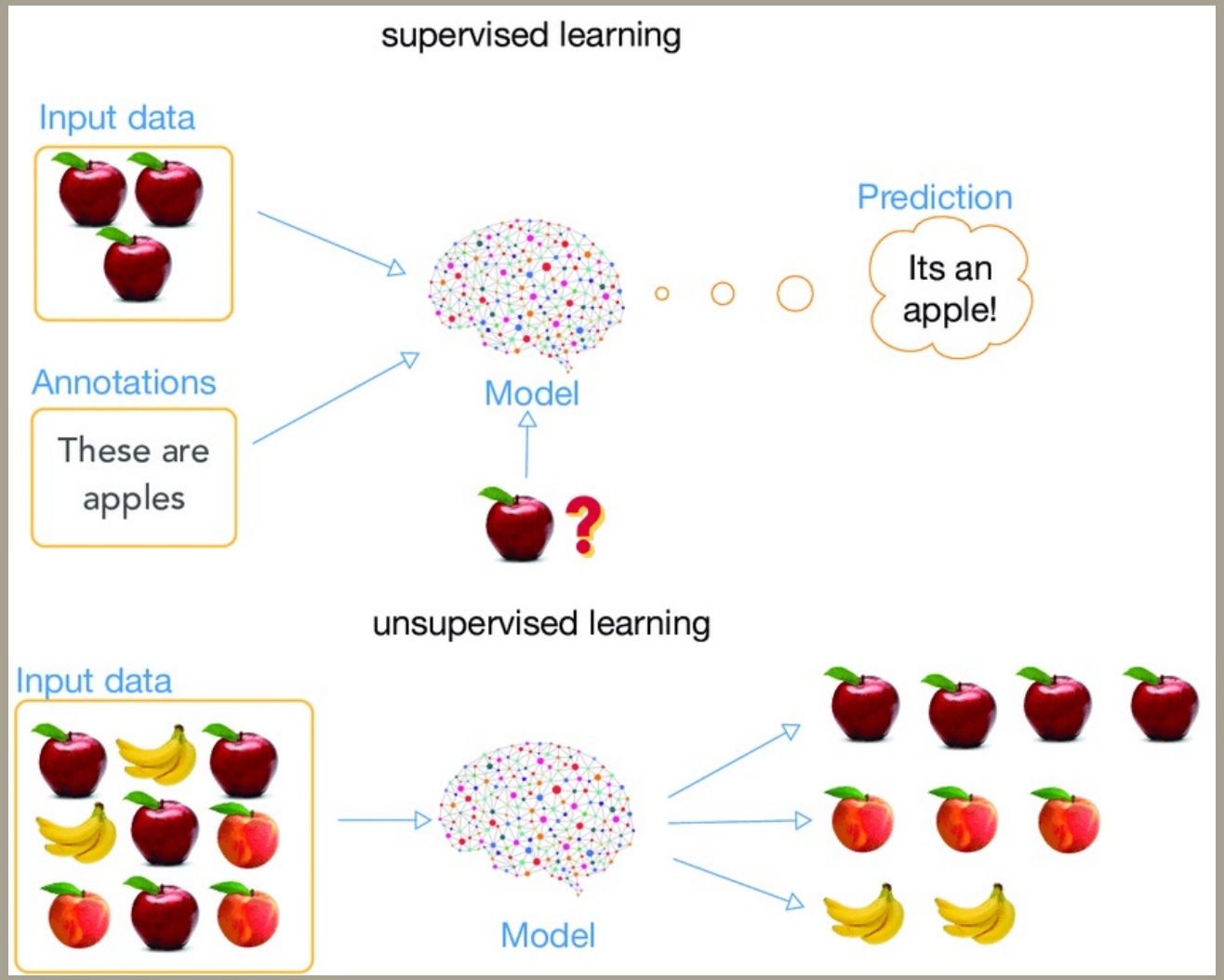
- No sabemos en cuantas clases separar cada noticias. E.g., deportes, sociales, nacional, internacional, etc.
- Una forma es determinar clústeres por medio de similitud en cuanto a temas o palabras.



Aprendizaje No Supervisado



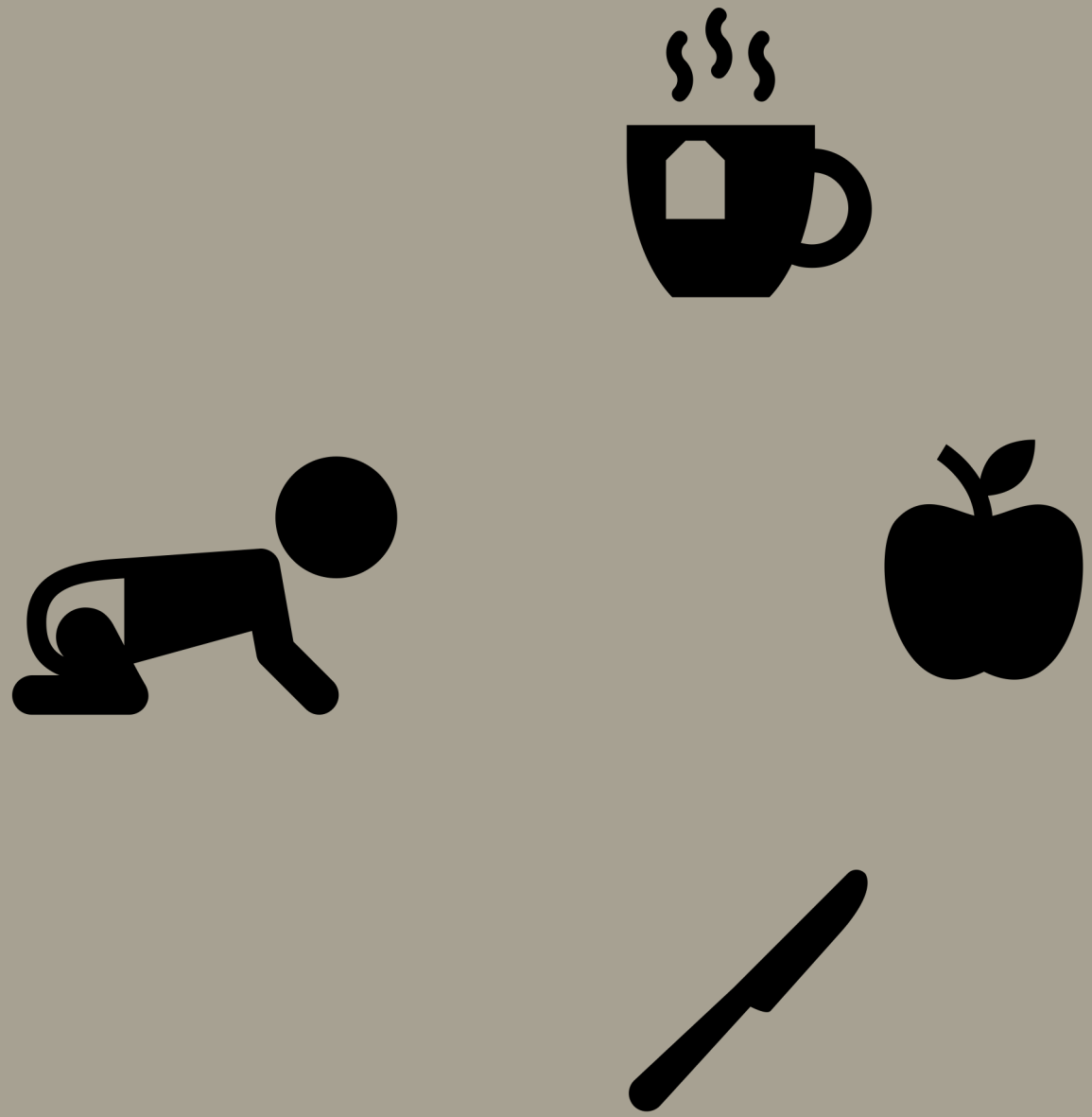




Aprendizaje Supervisado

vs

Aprendizaje No Supervisado



Aprendizaje por Refuerzo

Machine Learning

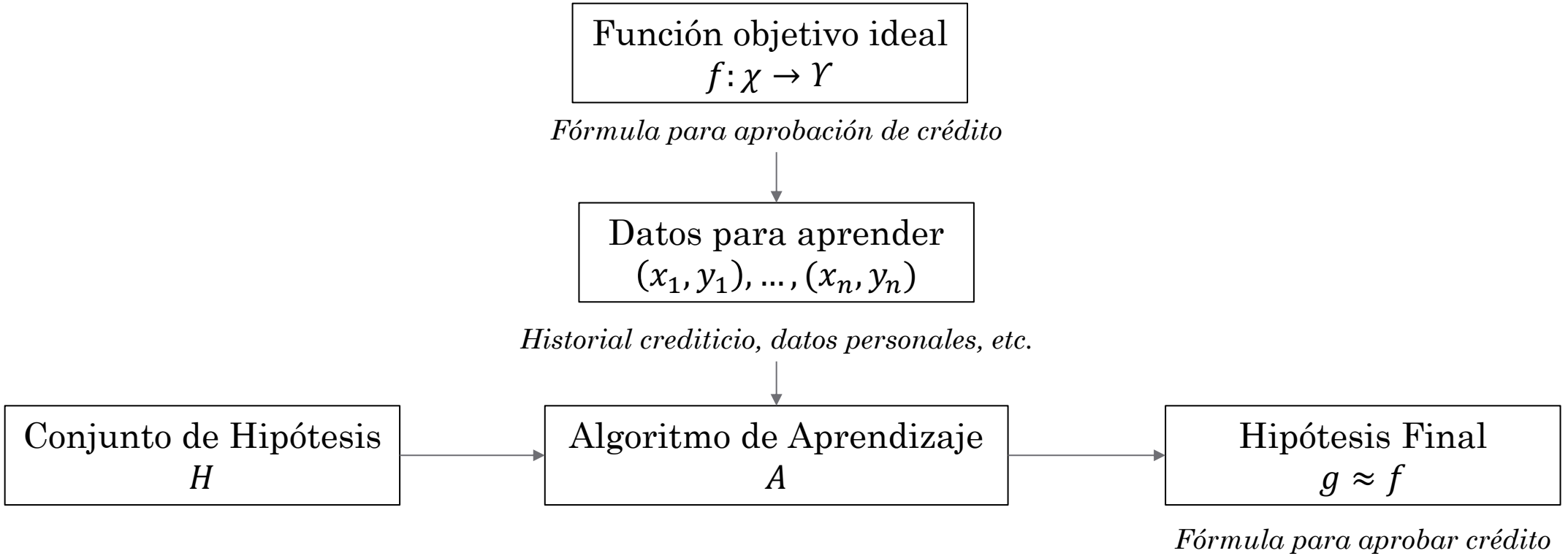
Machine Learning

- 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**.

Tarea

- Investigar sobre una aplicación de aprendizaje automático donde se utilice el enfoque de 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.

Componentes del Aprendizaje



Segunda Vuelta: Machine Learning

- Dado un problema específico, la función objetivo y los datos para entrenar los dicta el problema.
- Sin embargo, el algoritmo de aprendizaje y las hipótesis no.
- Lo anterior son **herramientas** que nosotros elegimos.
- Vamos a refrescar las cosas con un modelo muy sencillo.

Segunda Vuelta: Machine Learning

Sea $\mathrm{X} = \mathbb{R}^d$ el espacio de las entradas, y sea $y = \{-1, +1\}$ el espacio de las salidas, lo cual denota un espacio de decisión binario (sí/no, positivo/negativo, encendido/apagado, etc.).

Vamos a ponerlo en el contexto del problema para aceptar créditos bancarios:

- El vector $\boldsymbol{x} \in \mathbb{R}^d$ denota las características que se utilizan para modelar: salario, sexo, deudas actuales, estado en buró de crédito, etc.
- La variable de salida y correspondiente a cada vector $\boldsymbol{x}$ representa si se aceptó o se denegó el crédito.

Segunda Vuelta: Machine Learning

- Vamos a especificar el conjunto de hipótesis H mediane una forma funcional que todas las hipótesis $h \in H$ comparten, la cual asigna diferentes pesos a cada entrada de $\boldsymbol{x}$, lo cual refleja su importancia o contribución relativa para la decisión final.
- La ponderación se combina para determinar una calificación crediticia y se compara con un valor límite.
- Si pasa el límite, se acepta el crédito, si no, se niega:

$$\text{Se aprueba si} \qquad \sum_{i=1}^{d} w_i x_i > \text{límite}$$

$$\text{Se niega si} \qquad \sum_{i=1}^{d} w_i x_i < \text{límite}$$

Segunda Vuelta: Machine Learning

Ejercicio:

¿Cómo se puede escribir esto de forma compacta (i.e., en una sola expresión)?

Se aprueba si $\sum_{i=1}^{d} w_i x_i > -b$

Se niega si $\sum_{i=1}^{d} w_i x_i < -b$

Solución:

$$h(\boldsymbol{x}) = sign\left(\sum_{i=1}^{d} w_i x_i + b\right)$$

Segunda Vuelta: Machine Learning

$$h(\boldsymbol{x}) = sign\left(\sum_{i=1}^{d} w_i x_i + b\right)$$

- En este modelo, w_i son los pesos asociados a x_i.
- b es el límite que debe rebasar la ponderación, también llamado *bias*.
- $h(\boldsymbol{x}) = +1$ indica que el crédito es aprobado y $h(\boldsymbol{x}) = -1$ indica que es denegado.
- Este modelo de H se conoce como perceptrón.

Segunda Vuelta: Machine Learning

$$h(\boldsymbol{x}) = sign\left(\sum_{i=1}^{d} w_i x_i + b\right)$$

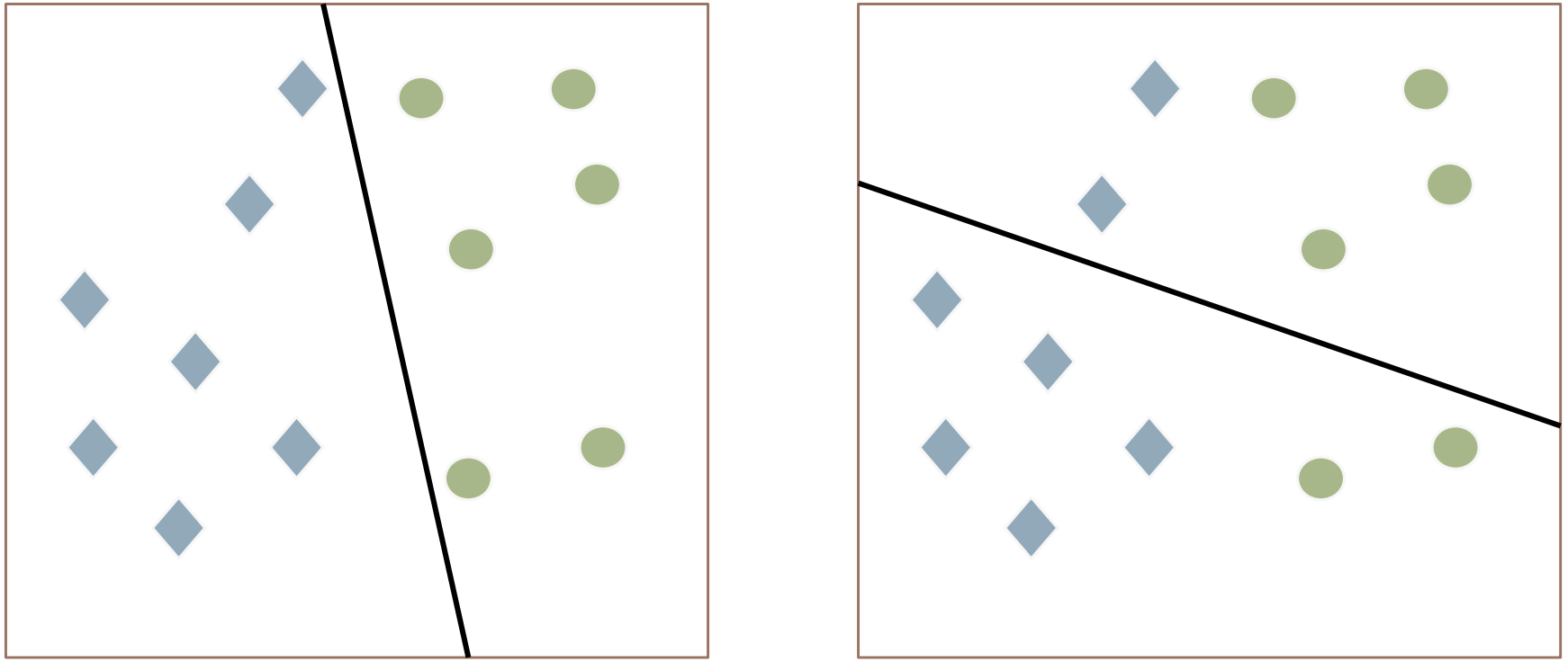
- El algoritmo de aprendizaje busca en H la mejor hipótesis mediante la comparación de pesos y bias que mejor desempeño tengan con el conjunto de datos dado.
- Los pesos w_i no deben ser necesariamente positivos. Si existe alguno negativo, indica que esa característica tiene un efecto adverso en el proceso de aprobación de crédito.
- El valor del bias puede ser grande o pequeño, depende de la actitud histórica del banco para la aprobación de créditos.

Segunda Vuelta: Machine Learning

$$h(\boldsymbol{x}) = sign\left(\sum_{i=1}^{d} w_i x_i + b\right)$$

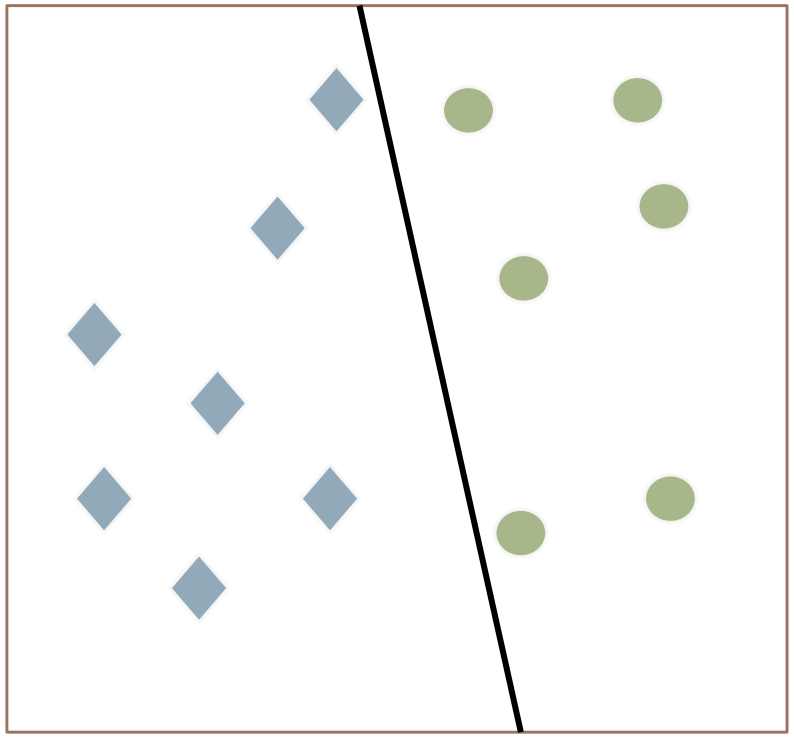
- La elección óptima de pesos y el sesgo definen la hipótesis final $g \in H$ que el algoritmo arroja como resultado.

Segunda Vuelta: Machine Learning



Clasificación realizada por el perceptrón para el caso datos linealmente separables en $\mathbb{R}^2$. Izquierda: separación perfecta. Derecha: Algunos puntos son clasificados mal.

Segunda Vuelta: Machine Learning



- El perceptrón parte el plano en dos regiones: la región +1 y la región -1.
- Valores distintos para w_1, w_2 y b dan como resultado distintas rectas $w_1x_1 + w_2x_2 + b = 0$.
- Si la información es linealmente separable, existen valores para los parámetros que separan perfectamente los puntos.

Segunda Vuelta: Machine Learning

Para simplificar la notación, vamos a tratar el bias b como un peso $w_0 = b$ y juntarlo con los otros pesos en un vector:

$$\boldsymbol{w} = (w_0, w_1, \dots, w_d) = (b, w_1, \dots, w_d)$$

y

$$\boldsymbol{x} = (x_0, x_1, \dots, x_d) = (1, x_1, \dots, x_d)$$

De esta manera, la fórmula se reduce a

$$h(x) = sign\left(\sum_{i=0}^{d} w_i x_i\right) = sign(\boldsymbol{w}^T \boldsymbol{x}).$$

Segunda Vuelta: Machine Learning

Con esto en mente, introducimos el **algoritmo de aprendizaje del perceptrón** (AAP), que determina el vector ***w*** basado en los datos disponibles.

Supuesto: Los datos o puntos de entrenamiento son linealmente separabales.

Segunda Vuelta: Machine Learning

El AAP busca determinar el vector $\boldsymbol{w}$ usando un método iterativo:

- En la t-ésima iteración, $t = 1,2, \ldots$, se tiene un valor de $\boldsymbol{w}$ en el paso t, llamémoslo $\boldsymbol{w}(t)$.
- El algoritmo elige un punto $\boldsymbol{x}$ mal clasificado, llamémoslo $(\boldsymbol{x}(t), y(t))$, para actualizar el peso $\boldsymbol{w}(t)$.
- Ya que el punto se encuentra mal clasificado,

$$y(t) \neq sign\left(\boldsymbol{w}^T(t)\boldsymbol{x}(t)\right)$$

- La regla de actualización es

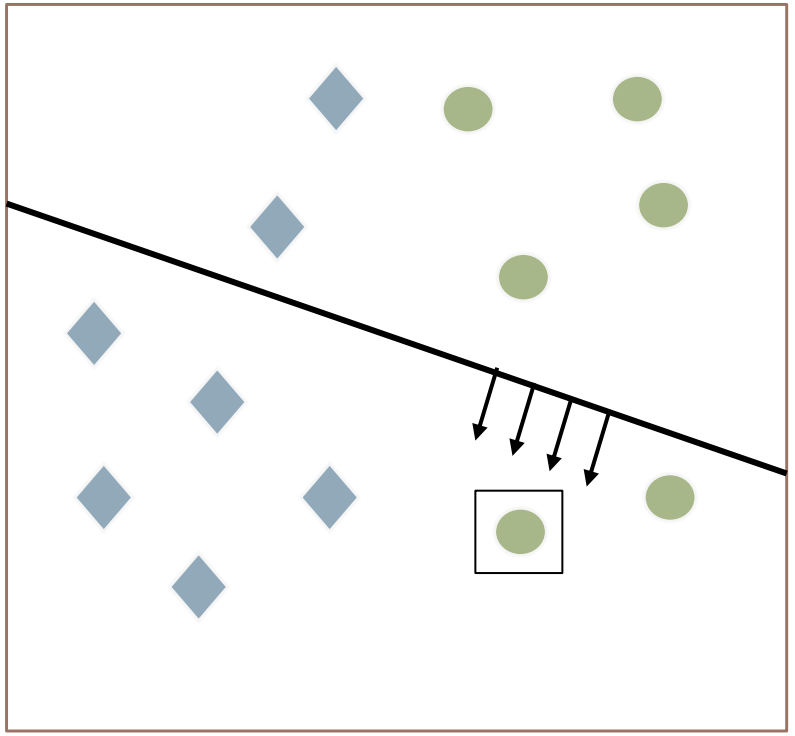
$$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + y(t)\boldsymbol{x}(t)$$

Segunda Vuelta: Machine Learning

- La regla de actualización
  $$\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + y(t)\boldsymbol{x}(t)$$

tiene el efecto de mover el borde o hiperplano de separación en la dirección de $\boldsymbol{x}(t)$ de tal manera que se clasifica correctamente.

- El algoritmo continúa con iteraciones subsecuentes hasta que no existen puntos mal clasificados.



Segunda Vuelta: Machine Learning

Tarea:

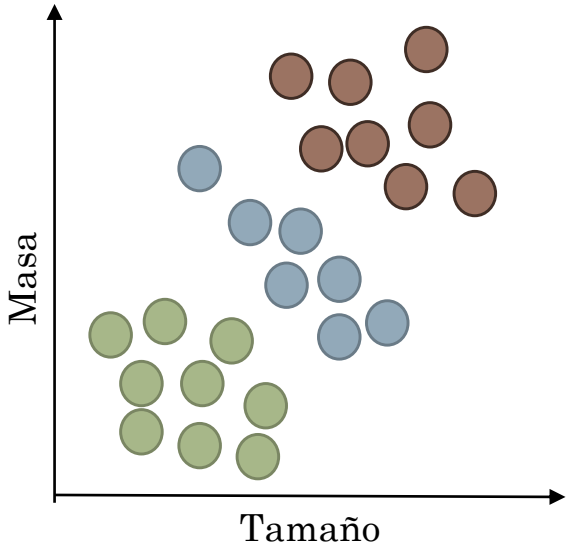
- Leer del capítulo 1 del libro *Learning From Data* de Mostafa et al., las secciones 1.1 y 1.2.
- Realizar el ejercicio 1.3, página 8 del libro *Learning From Data* de Mostafa et al.

Aprendizaje vs Diseño

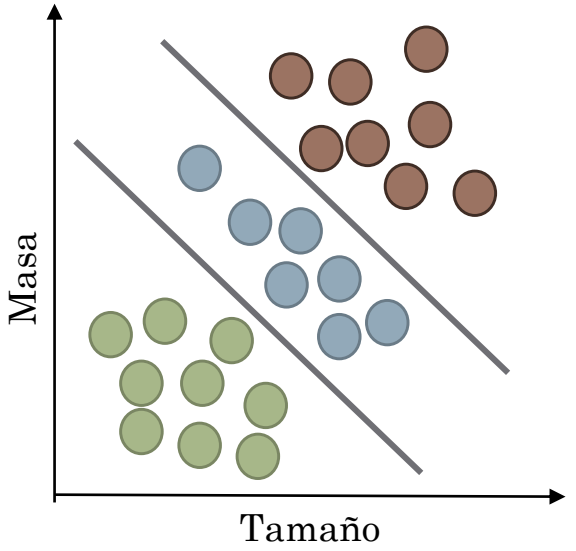
- A lo largo del curso aprendimos diversos modelos de aprendizaje supervisado y no supervisado.
- Noten que aprender comprende usar datos para entrenar los modelos.
- Otra forma para abordar problemas similares es mediante el **diseño** de soluciones basadas en especificaciones que se pueden observar en el problema.
- Ambas corrientes se suelen ver en cursos de reconocimiento de patrones.

Aprendizaje vs Diseño

- Consideremos el problema de separar monedas según su denominación.
- Usaremos un modelo basado en el tamaño y la masa de la moneda.
- ¿Cómo se vería el problema desde la perspectiva de aprender de los datos?



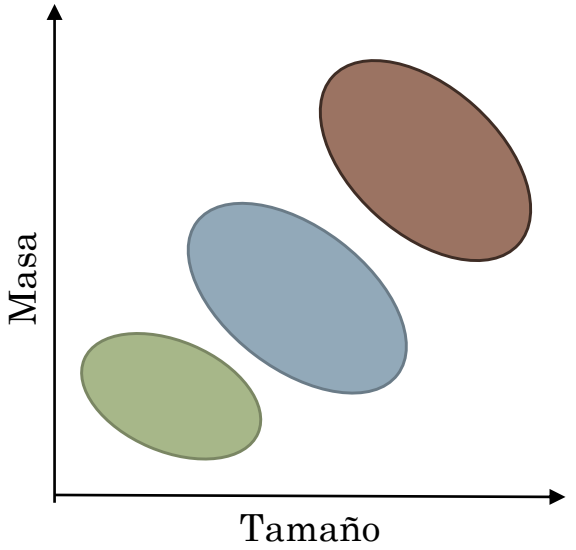
a) Info de las monedas.



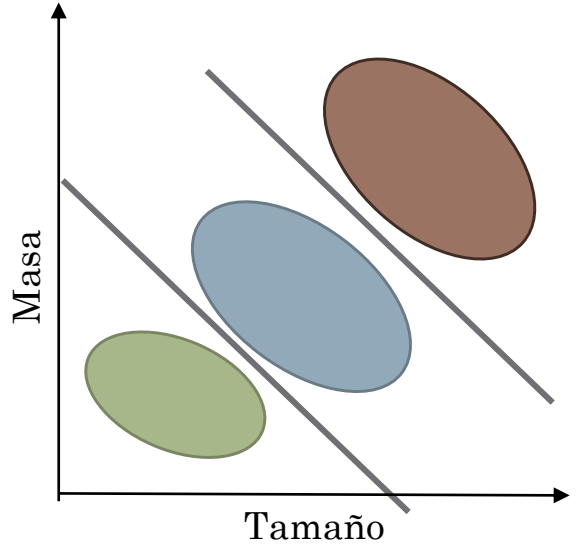
b) Modelo aprendido

Aprendizaje vs Diseño

- En el modelo con diseño buscamos información de expertos.
  - Especificaciones de cada moneda.
  - Número de monedas en circulación para cada denominación para crear un modelo frecuentista.
  - Modelo físico de variaciones en masa y tamaño debido a la erosión por los elementos y errores en fabricación.
- Con todo esto, se crea un modelo probabilístico para determinar la probabilidad conjunta dada la masa, el tamaño y la denominación de la moneda.
- Con este modelo, se crea la regla de clasificación que maximice la probabilidad dada la masa y el tamaño de la moneda.



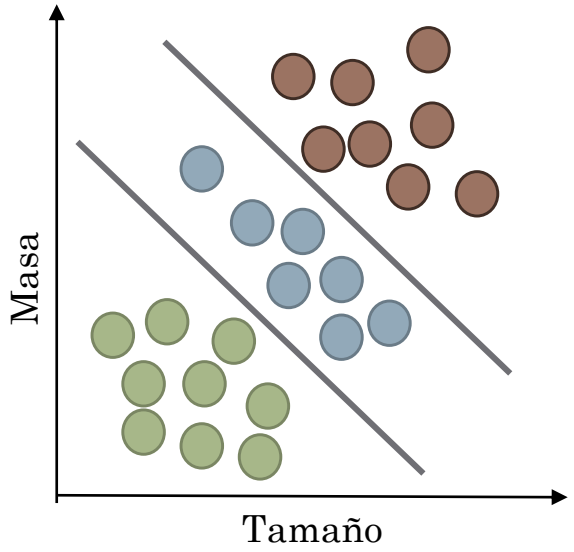
a) Modelo probabilístico



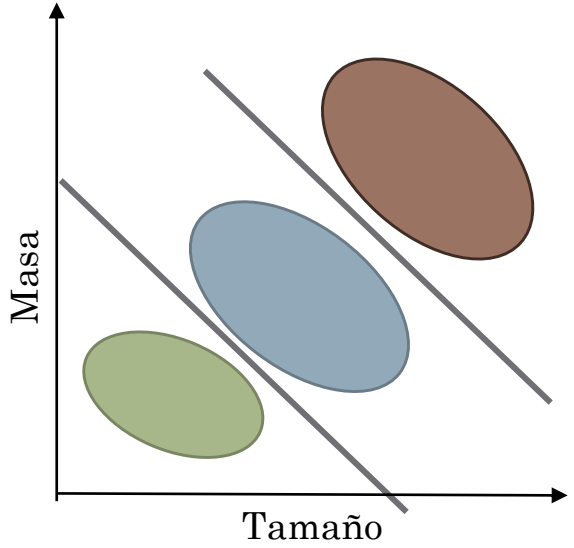
b) Clasificador inferido

Aprendizaje vs Diseño

- La diferencia entre ambas formas de resolver el problema es el rol de los datos.
- En el problema con diseño, el problema se encuentra bien definido por lo que se puede determinar f analíticamente sin usar los datos.
- La postura de aprender de los datos implica que el problema se encuentra menos especificado y necesita los datos para acercarse a f.
- Ambos son viables y tienen su mérito. Pero sólo aprender de los datos es viable cuando la función objetivo es desconocida.



a) Modelo aprendido



b) Clasificador inferido

¿Aprender es viable?

- El objetivo del Machine Learning como le hemos visto hasta el momento **es encontrar o aproximar** f.
- Además, f no se conoce. Es un total misterio.
- Esto da pie a formular la siguiente (crítica) pregunta:
  - ¿Cómo un conjunto de datos limitado contiene información suficiente para encontrar o aproximar f?

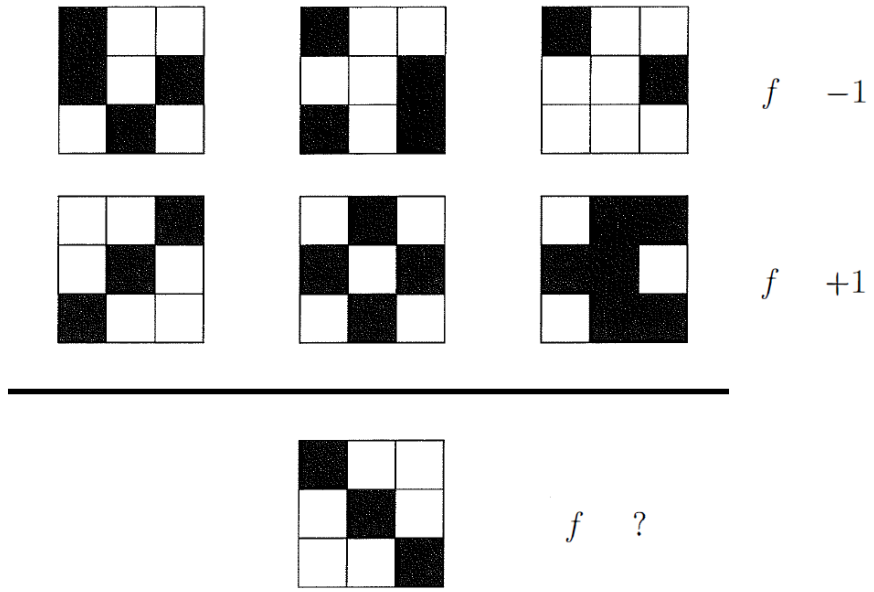


Figure 1.7: A visual learning problem. The first two rows show the training examples (each input $\mathbf{x}$ is a 9 bit vector represented visually as a 3×3 black and white array). The inputs in the first row have $f(\mathbf{x}) = -1$, and the inputs in the second row have $f(\mathbf{x}) = +1$. Your task is to learn from this data set what f is, then apply f to the test input at the bottom. Do you get -1 or $+1$?

¿Aprender es viable?

- Cuando tenemos el conjunto de datos D, conocemos los valores de f en esos puntos.
- Esto no quiere decir que ya aprendimos f, ya que no tenemos garantía que sabemos algo de f fuera de D.
- ¿El conjunto de datos D nos dice algo fuera de D?
  - Si sí, hemos logrado aprender.
  - Si no, entonces solo hemos memorizado e implica que el aprendizaje no es factible.

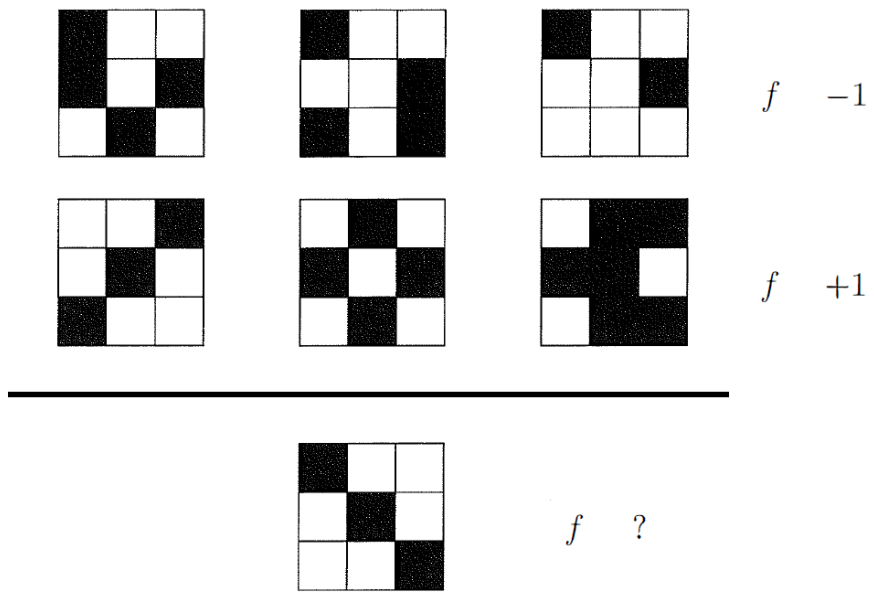


Figure 1.7: A visual learning problem. The first two rows show the training examples (each input $\mathbf{x}$ is a 9 bit vector represented visually as a 3×3 black and white array). The inputs in the first row have $f(\mathbf{x}) = -1$, and the inputs in the second row have $f(\mathbf{x}) = +1$. Your task is to learn from this data set what f is, then apply f to the test input at the bottom. Do you get -1 or $+1$?

¿Aprender es viable?

- Vamos a explorar la idea de que, ya que f no se conoce, f permanece desconocida fuera de D.
- Consideran la función Booleana de la derecha, donde $X \in \{0,1\}^3$.
- Además, nos dan un conjunto de datos con cinco elementos.
- La salida $y_n = f(x_n)$ es binaria, representada por ●/○, para $n = 1,2,3,4,5$.
- La ventaja es que solo existen $2^3 = 8$ vectores (datos) distintos.
- ¿Cuántas funciones f de tres entradas posibles existen?

$\mathbf{x}_n$	y_n
0 0 0	○
0 0 1	●
0 1 0	●
0 1 1	○
1 0 0	●

¿Aprender es viable?

- Vamos a enfocarnos en el problema de aprender f.
- Ya que f es desconocida salvo en D, cualquier función que coincida con D puede ser f.
- La imagen de la derecha muestra 8 candidatos a f y lo que producen en puntos fuera de D, i.e., la función g.
- La elección de g, la hipótesis final, depende de los cinco ejemplos de D.

x	y	g	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
0 0 0	○	○	○	○	○	○	○	○	○	○
0 0 1	●	●	●	●	●	●	●	●	●	●
0 1 0	●	●	●	●	●	●	●	●	●	●
0 1 1	○	○	○	○	○	○	○	○	○	○
1 0 0	●	●	●	●	●	●	●	●	●	●
1 0 1		?	○	○	○	○	●	●	●	●
1 1 0		?	○	○	●	●	○	○	●	●
1 1 1		?	○	●	○	●	○	●	○	●

¿Aprender es viable?

- Ya que no conocemos f, no podemos excluir ninguna función de $f_1, \ldots, f_8$.
- ¡Tenemos un dilema!
- El objetivo de encontrar o aproximar f es inferir o predecir el valor de la función en puntos fuera de D.
- La calidad del aprendizaje se determina por qué tan buenas sean las predicciones comparadas con valores reales.

x	y	g	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
0 0 0	○	○	○	○	○	○	○	○	○	○
0 0 1	●	●	●	●	●	●	●	●	●	●
0 1 0	●	●	●	●	●	●	●	●	●	●
0 1 1	○	○	○	○	○	○	○	○	○	○
1 0 0	●	●	●	●	●	●	●	●	●	●
1 0 1		?	○	○	○	○	●	●	●	●
1 1 0		?	○	○	●	●	○	○	●	●
1 1 1		?	○	●	○	●	○	●	○	●

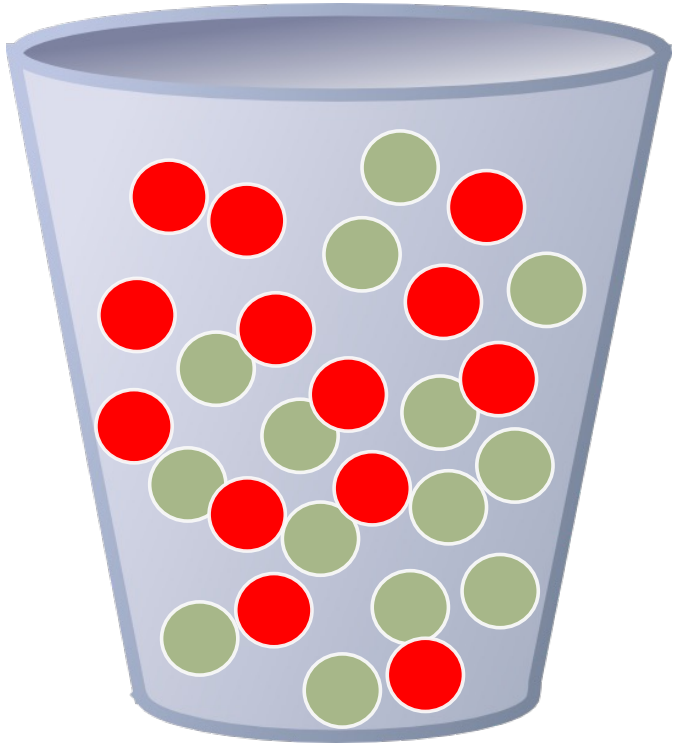
¿Aprender es viable?

- No importa que algoritmo A se use o qué conjunto de hipótesis H se considere.
- Ya sea que H contenga una hipótesis que concuerde con D o no, o el algoritmo A elija dicha hipótesis o no, no hace diferencia alguna en lo que respecta al rendimiento fuera de D.
- ¡Pero el rendimiento fuera de D es lo único que importa en el aprendizaje!
- Mientras f sea una función desconocida, al conocer D no se puede excluir ningún patrón de valores para f fuera de D.
- Por lo tanto, las predicciones de g fuera de D no tienen sentido.

x	y	g	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
0 0 0	○	○	○	○	○	○	○	○	○	○
0 0 1	●	●	●	●	●	●	●	●	●	●
0 1 0	●	●	●	●	●	●	●	●	●	●
0 1 1	○	○	○	○	○	○	○	○	○	○
1 0 0	●	●	●	●	●	●	●	●	●	●
1 0 1		?	○	○	○	○	●	●	●	●
1 1 0		?	○	○	●	●	○	○	●	●
1 1 1		?	○	●	○	●	○	●	○	●

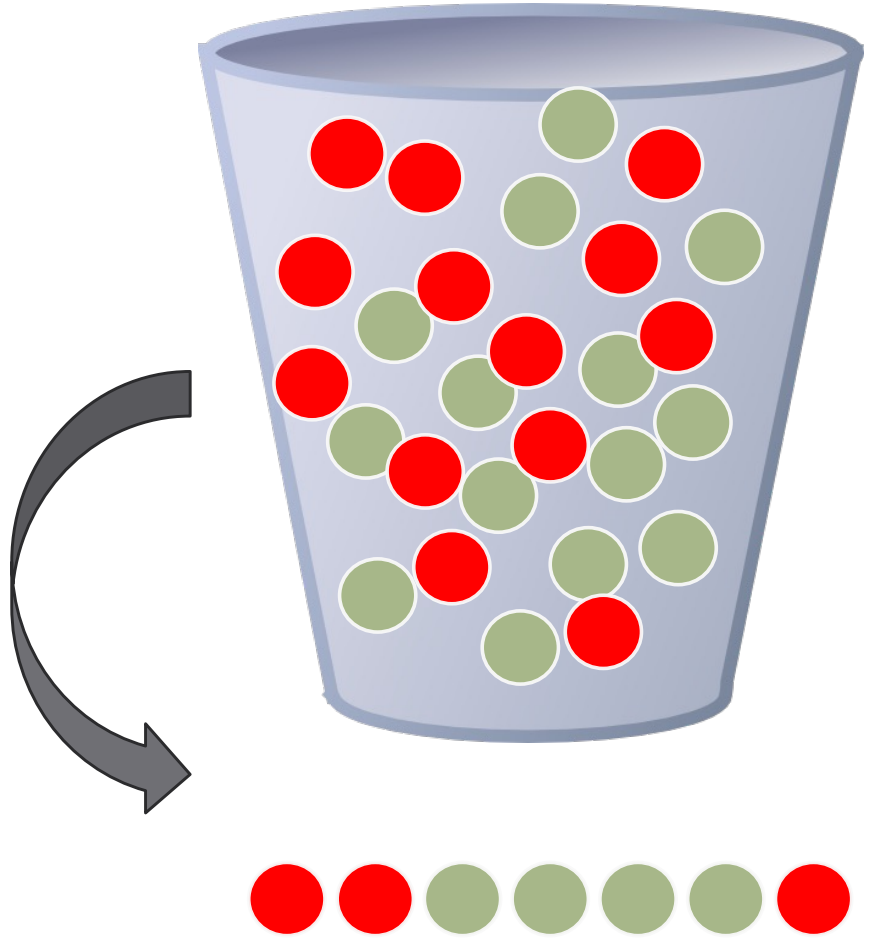
¿Aprender es viable?

- Vamos a marcar el camino para mostrar que se puede inferir algo afuera de D usando únicamente D.
- Consideremos el caso de crear una muestra y qué se puede decir de los objetos fuera de esa muestra.
- Consideremos un jarrón con bolitas verdes y rojas, posiblemente una infinidad de ellas.
- La probabilidad de agarrar una roja es μ y la de agarrar una verde es $1 - \mu$.
- El valor de μ es desconocido para nosotros.



¿Aprender es viable?

- Se elige una muestra de N bolas independientes con reemplazo y observamos la fracción v de bolas rojas de la muestra.
- ¿Qué nos dice v de μ?
- Una respuesta es que, sin importar los colores de las bolas que elegimos en N, no sabemos el color de una bola que no elegimos.
- Podemos obtener una mayoría de bolas verdes cuando la gran mayoría de bolas son rojas (en la cesta). Es posible, pero no es probable.

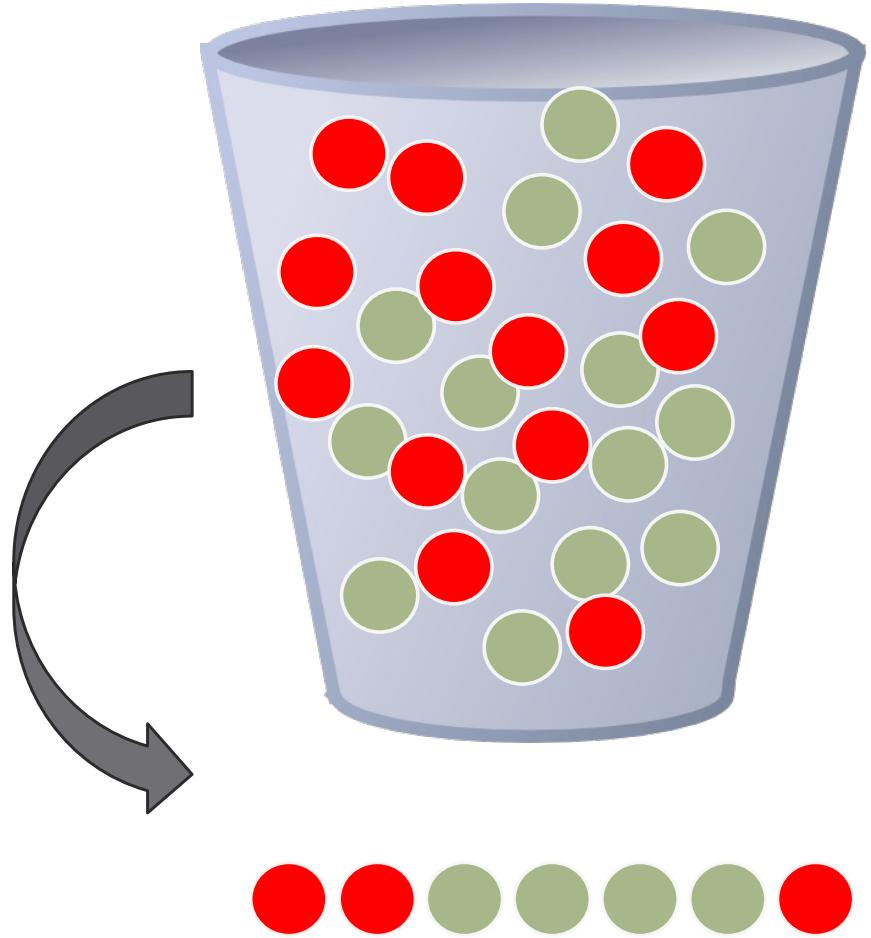


¿Aprender es viable?

Ejercicio

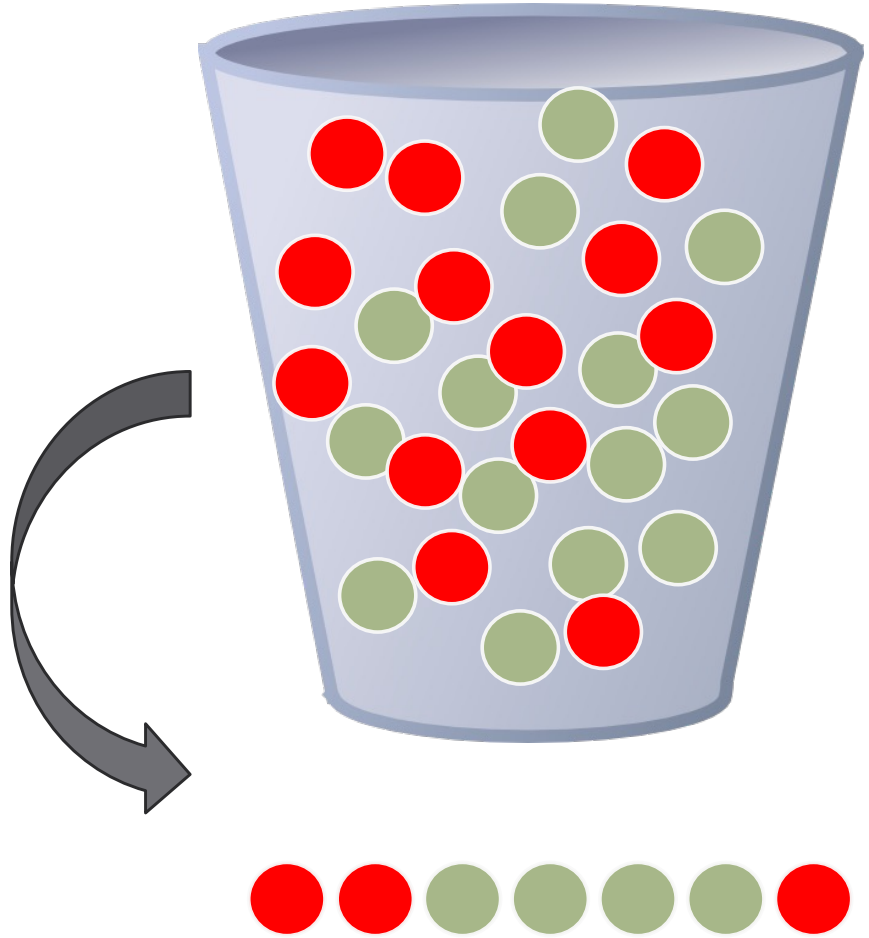
Si $\mu = 0.9$, ¿cuál es la probabilidad de que una muestra de 10 bolas tenga un valor de $v \leq 0.1$?

Pista: consideren una distribución binomial.



¿Aprender es viable?

- Esta situación **es similar** a realizar una **encuesta**.
- Una muestra aleatoria de una población tiende a convenir con las opiniones de la población.
- Cuando la población es grande, v tiende a μ.



¿Aprender es viable?

Para cuantificar la relación entre v y μ, podemos usar una cota simple llamada *Desigualdad de Hoeffding*. Esta nos dice que para cualquier muestra de tamaño *N:*

$$P[|v - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

para cualquier valor de ϵ.

¿Qué quiere decir lo anterior?

¿Aprender es viable?

Para cuantificar la relación entre v y μ, podemos usar una cota simple llamada *Desigualdad de Hoeffding*. Esta nos dice que para cualquier muestra de tamaño N:

$$P[|v - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

para cualquier valor de ϵ.

La desigualdad anterior nos dice que, conforme N aumenta, se vuelve exponencialmente poco probable que la diferencia entre v y μ sea mayor que nuestra tolerancia ϵ.

¿Aprender es viable?

$$P[|v - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

- La única cantidad aleatoria en esta desigualdad es el valor de v, ya que depende de la muestra obtenida.
- Por otro lado, μ no es aleatoria. Es una constante, aunque desconocida.
- La utilidad de la desigualdad es inferir el valor de μ dado v, aunque v sea afectado por μ.
- Esta afectación es que v tiende a estar cerca de μ, podemos inferir que μ tiende a v.

¿Aprender es viable?

$$P[|v - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

- A pesar de que $P[|v - \mu| > \epsilon]$ depende de μ, podemos acotar la probabilidad con $2e^{-2\epsilon^2 N}$, que no depende de μ.
- Solo N afecta este valor de la cota, ni siquiera el valor del tamaño de la cesta (población).

¿Aprender es viable?

$$P[|v - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

Ejercicio

Si $\mu = 0.9$, usen la desigualdad de Hoeffding para acotar la probabilidad de que una muestra de 10 bolas tenga $v \leq 0.1$ y comparen con el ejercicio anterior.

¿Aprender es viable?

$$P[|v - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

Respuesta

Dado $\mu = 0.9, N = 10$, queremos $v \leq 0.1$. Es decir,

$$|\mu - v| = \mu - v \geq 0.9 - 0.1 = 0.8$$

En total se tiene que

$$P[v \leq 0.1] = P[|\mu - v| \geq 0.8] \leq P[|\mu - v| > 0.7] \leq 2e^{-2\epsilon^2 N}$$
$$\approx 0.0001109032$$

¿Aprender es viable?

$$P[|v - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

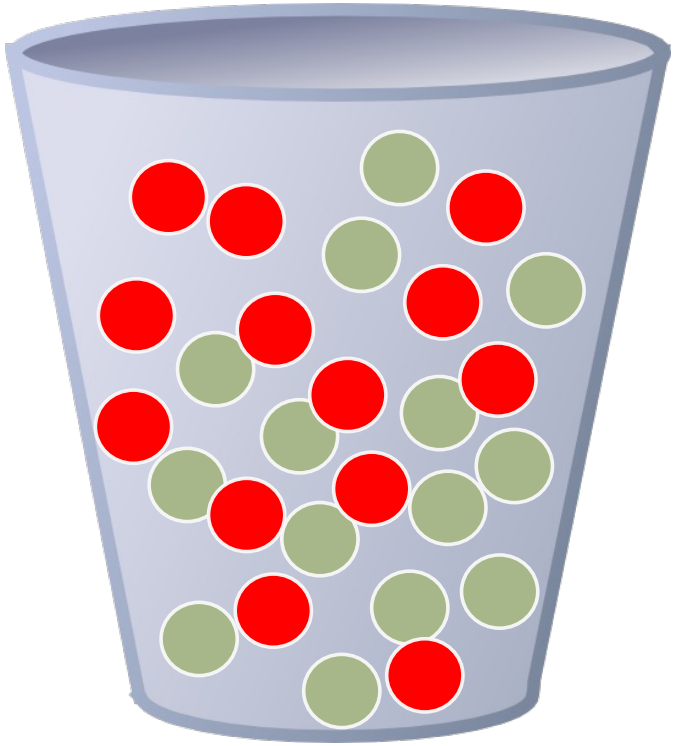
- Si queremos que esta diferencia entre v y μ sea pequeña, i.e., un valor de ϵ pequeño, necesitamos que N sea grande.
- Aunque no sepamos el valor exacto de μ, sabemos que estamos a $\pm\epsilon$ de μ la mayor parte del tiempo.
- Noten que esto funciona al realizar un muestreo aleatorio. Si la muestra no se crea de esta manera, no se puede aplicar este análisis probabilístico.

¿Aprender es viable?

¿Cómo se relaciona esto con el problema del aprendizaje?

- En el problema de las bolas, lo que no conocemos es el valor de μ, mientras que en el problema de aprendizaje es encontrar $f: X \to Y$.
- ¿Cómo las conectamos?

¿Aprender es viable?



Conjunto de Hipótesis
H

h

Comparamos h con f en cada punto $\boldsymbol{x} \in X$.

Si $h(x) = f(x)$, coloreamos $\boldsymbol{x}$ de verde.
Si $h(x) \neq f(x)$, coloreamos $\boldsymbol{x}$ de rojo.

No sabemos el color de cada punto dado que no conocemos f.

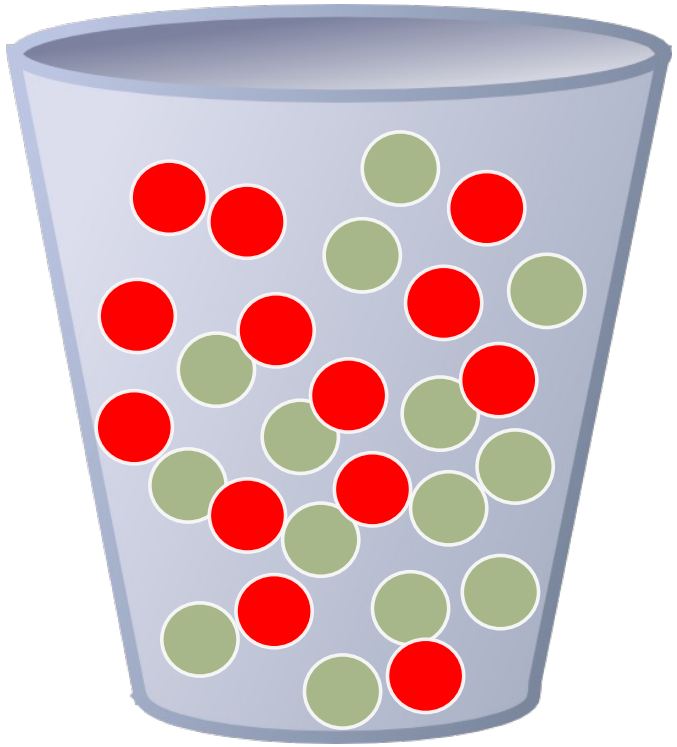
$\longrightarrow$

x es rojo con probabilidad μ

x es verde con probabilidad 1- μ

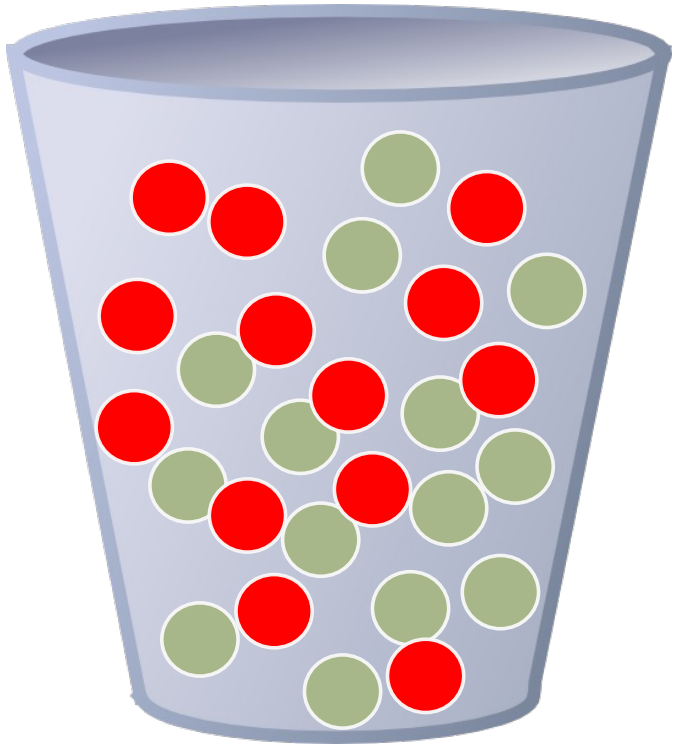
Sin embargo, si elegimos x al azar según una distribución P sobre X…

¿Aprender es viable?



- Los datos de entrenamiento funcionan como la muestra de la cesta.
- Si los datos $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ en D se eligen según P, se tiene una muestra aleatoria de puntos rojos con probabilidad μ, y verdes con probabilidad $1 - \mu$.
- Ahora, tanto $f(\boldsymbol{x}_n)$ como $\mathrm{h}(\boldsymbol{x}_n)$ son conocidos para $n = 1, \ldots, N$.
- Ya se redujo el problema del aprendizaje al de la cesta.

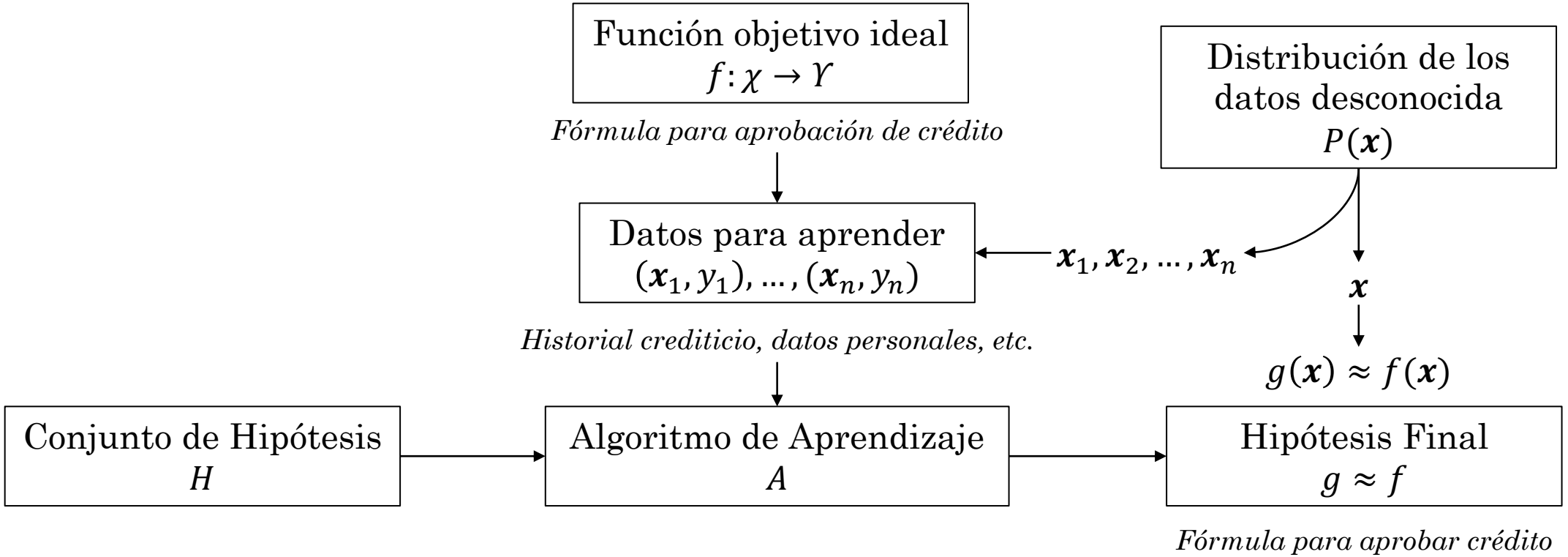
¿Aprender es viable?



- La desigualdad de Hoeffding ahora nos permite hacer una predicción fuera de D.
- Usar v para predecir μ nos dice algo acerca de f, pero no qué es f.
- Lo que nos dice μ es la razón de error que h implica al aproximar f.
- Si v se acerca a cero, podemos decir que h aproxima f bien sobre todo X.
- Si no, pues, mala suerte.

2ª Versión

Componentes del Aprendizaje



¿Aprender es viable?

- Desfortunadamente, no tenemos control sobre v ya que se basa en una hipótesis h en particular.
- Debemos buscar y probar varias hipótesis de H y elegir aquella con el menor error.
- Al usar sólo una, sólo estamos verificando si esa hipótesis es buena o mala.
- ¿Cómo se extiende la equivalencia de la cesta con varias hipótesis?*

¿Aprender es viable?

Lo anterior nos lleva a las siguientes conclusiones:

1. No se puede aprender nada fuera de D → punto de vista determinista.
2. Se puede aprender algo fuera de D → punto de vista probabilista.

- Al considerar el punto de vista probabilista, debemos tener cuidado ya que D se genera según una distribución de probabilidad.
- Esa misma distribución se debe usar para evaluar que tan bien g aproxima f.

Fin de la presentación

¡Gracias por su tiempo!