Machine Learning en Profundidad

Delta Analytics construye capacidad técnica alrededor del mundo.

El contenido de este curso está siendo desarrollado activamente por Delta Analytics, una organización sin fines de lucro 501(c)3 del Área de la Bahía que apunta a capacitar a las comunidades para aprovechar sus datos.

Por favor comuníquese con cualquier pregunta o comentario a inquiry@deltanalytics.org.

Descubre más sobre nuestra misión aquí.

Module 2:
Bloques componentes de Machine Learning.

Course overview:

- ✓ Módulo 1: Introducción a Machine Learning
- ✓ Módulo 2: Machine Learning en Profundidad
- ❑ Módulo 3: Selección y Evaluación del Modelo
- ❑ Módulo 4: Regresión Lineal
- ❑ Módulo 5: Árboles de Decisión
- ❑ Módulo 6: Algoritmos de Conjunto
- ❑ Módulo 7: Algoritmos de Aprendizaje no Supervisados
- ❑ Módulo 8: Procesamiento del Lenguaje Natural Parte 1
- ❑ Módulo 9: Procesamiento del Lenguaje Natural Parte 2

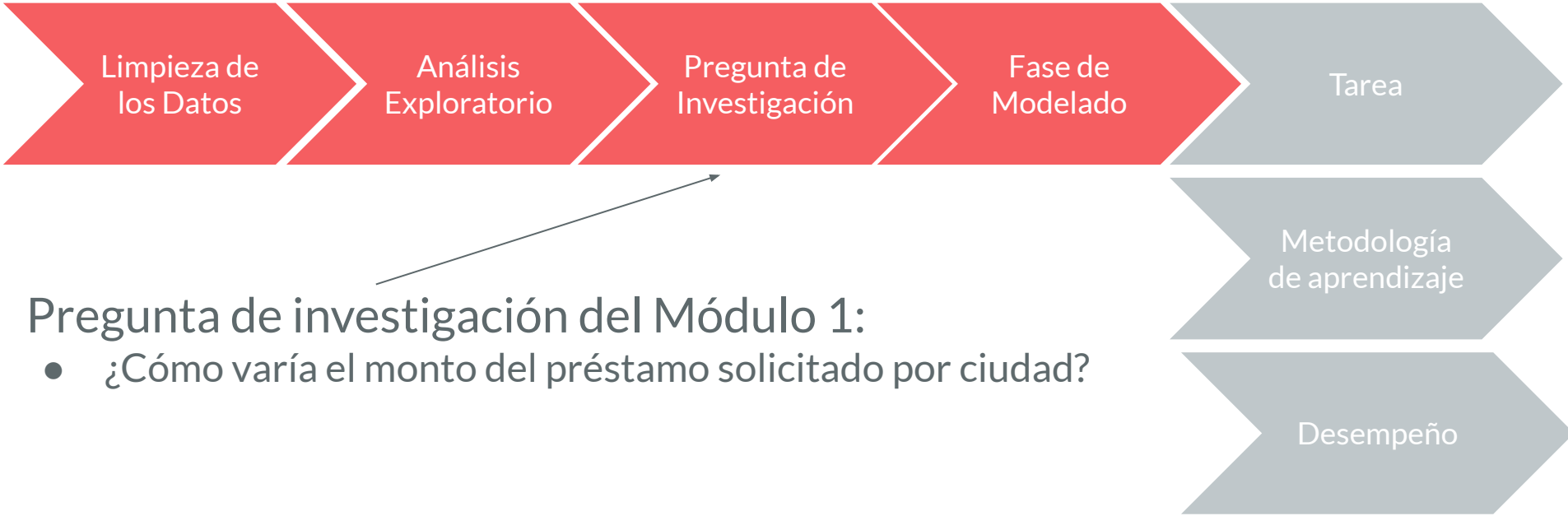
Ahora pasemos a los datos que usaremos ...

Module Checklist

- ❏ Desarrollo del Modelo
  - ❏ Definición de la tarea de machine learning
  - ❏ Medición del rendimiento del modelo
  - ❏ Métodos de aprendizaje supervisados y no supervisados
- ❏ Validación del Modelo

Fase de Modelado

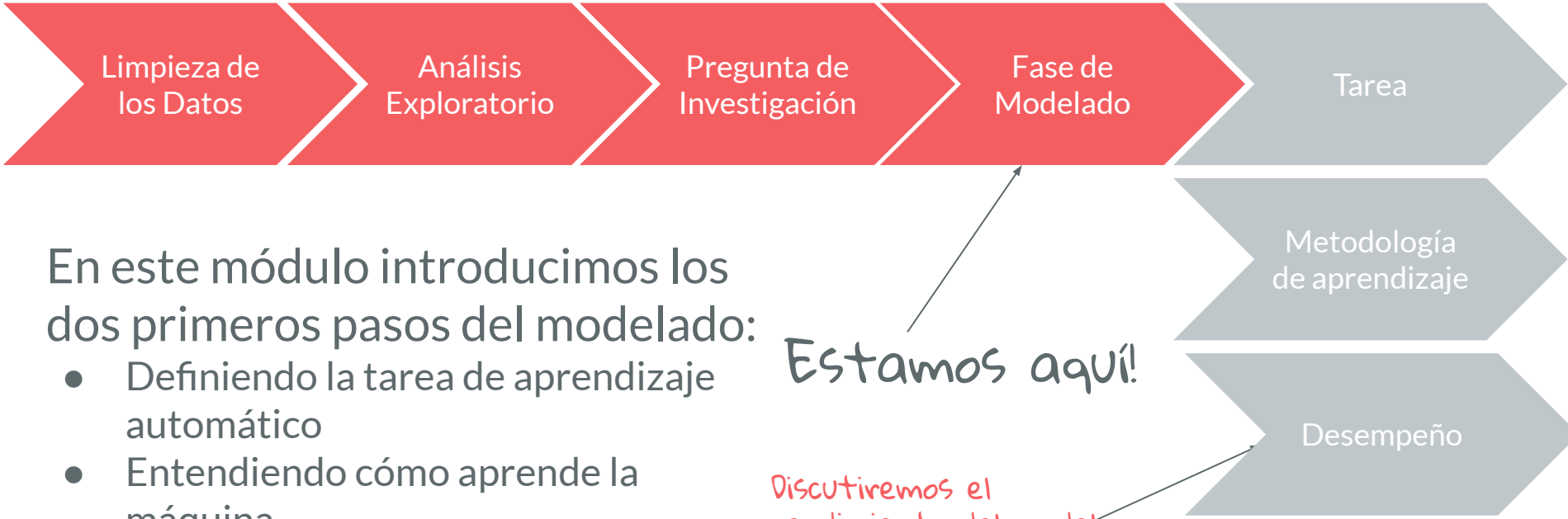
Ahora que tenemos nuestra pregunta de investigación, podemos comenzar a modelar



Pregunta de investigación del Módulo 1:

- ¿Cómo varía el monto del préstamo solicitado por ciudad?

Ahora que tenemos nuestra pregunta de investigación, podemos comenzar a modelar

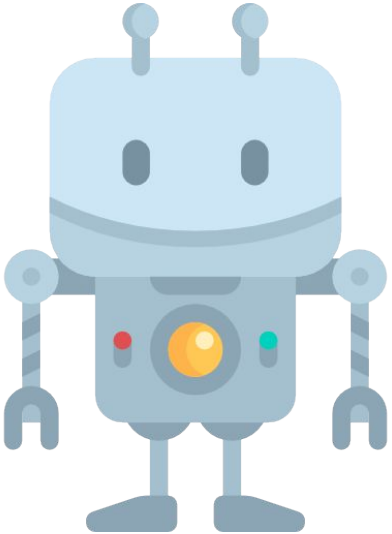


En este módulo introducimos los dos primeros pasos del modelado:

- Definiendo la tarea de aprendizaje automático
- Entendiendo cómo aprende la máquina.

Fase de Modelado

Empecemos por lo básico. ¿Por qué queremos construir un modelo?



Machine learning nos permite abordar tareas que son demasiado difíciles para codificar todos los enfoques posibles por nuestra cuenta.

Al permitir que las máquinas aprendan de la experiencia, evitamos la necesidad de que los humanos especifiquen todo el conocimiento que necesita una computadora.

Fuente: Deep Learning Book - Chapter 5: Introduction to Machine Learning

Intuición Humana

Basados en nuestra experiencia del mundo, entendemos las relaciones entre las características

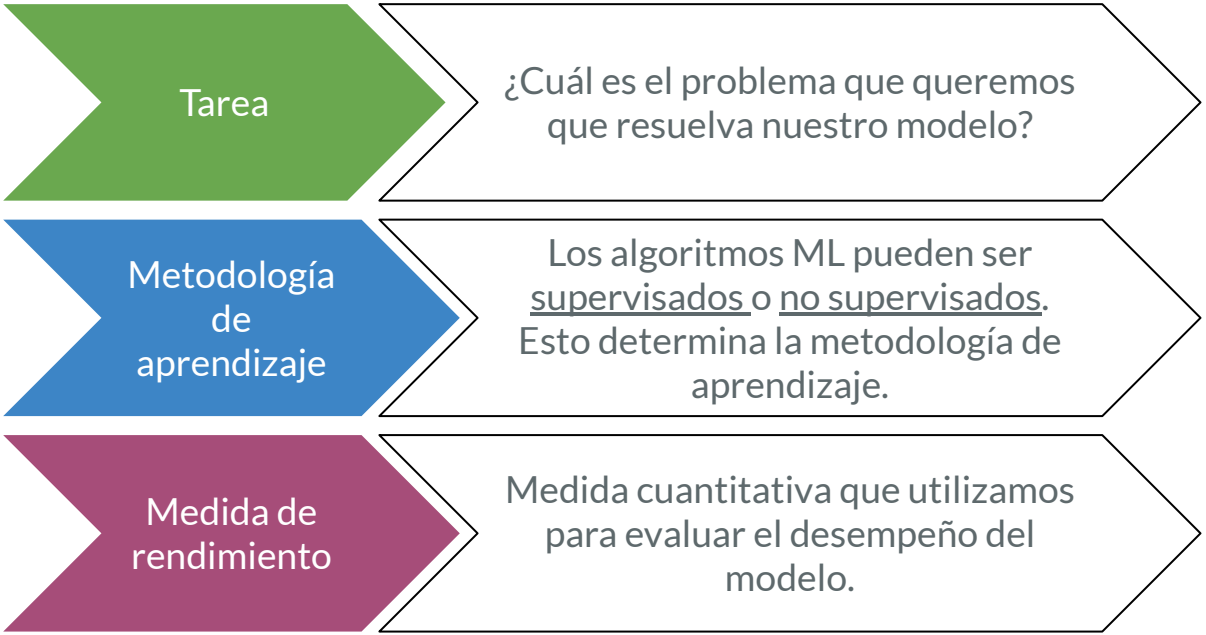
Modelo de Machine Learning

Las computadoras adquieren la intuición humana y la cuantifican, extrayendo patrones de datos en bruto

Los modelos de Machine learning cuantifican y aprenden los patrones que observamos en los datos..

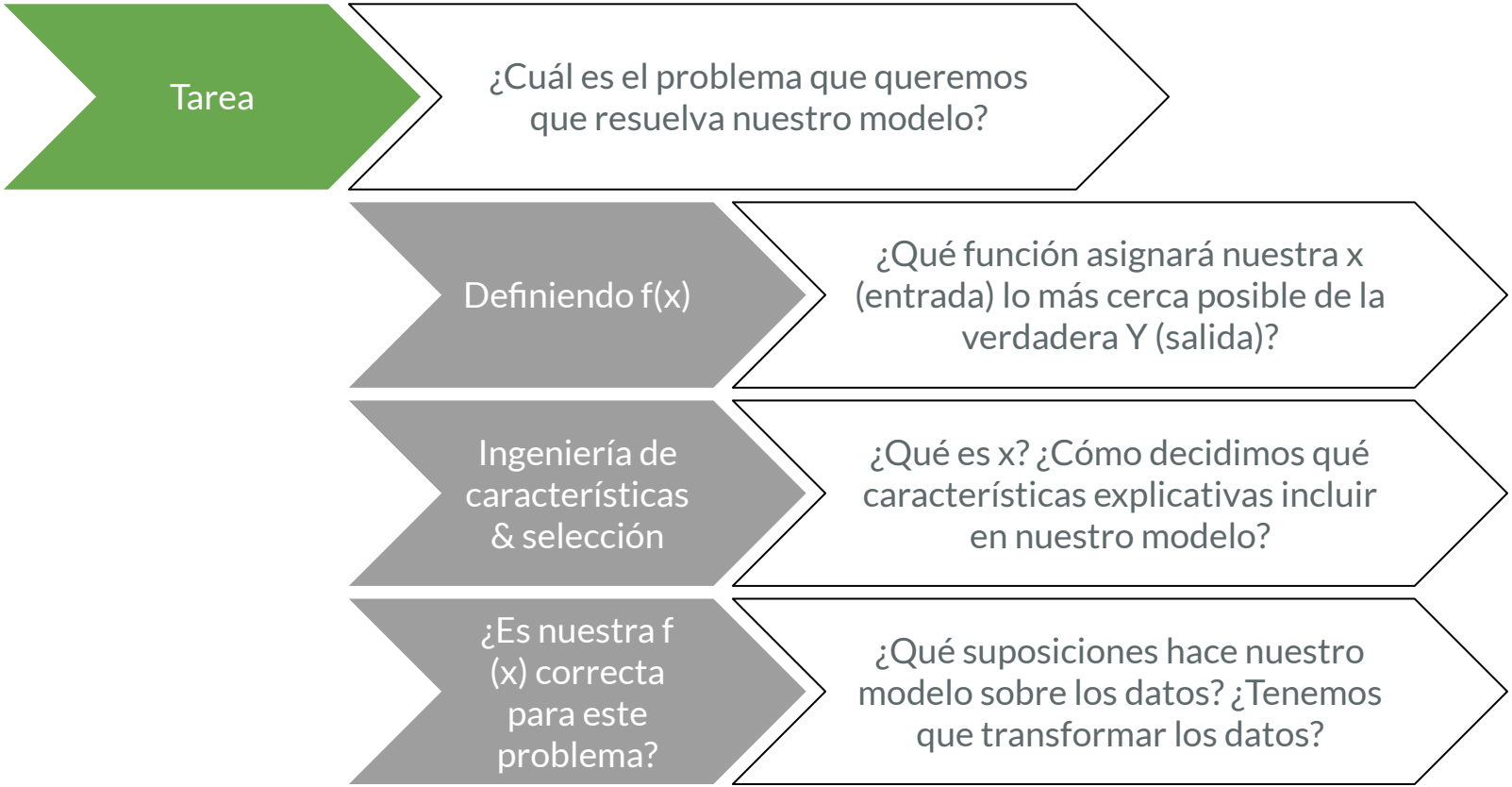
Fase de Modelado

Todos los modelos tienen 3 componentes clave: una tarea, una metodología de aprendizaje y una medida de rendimiento

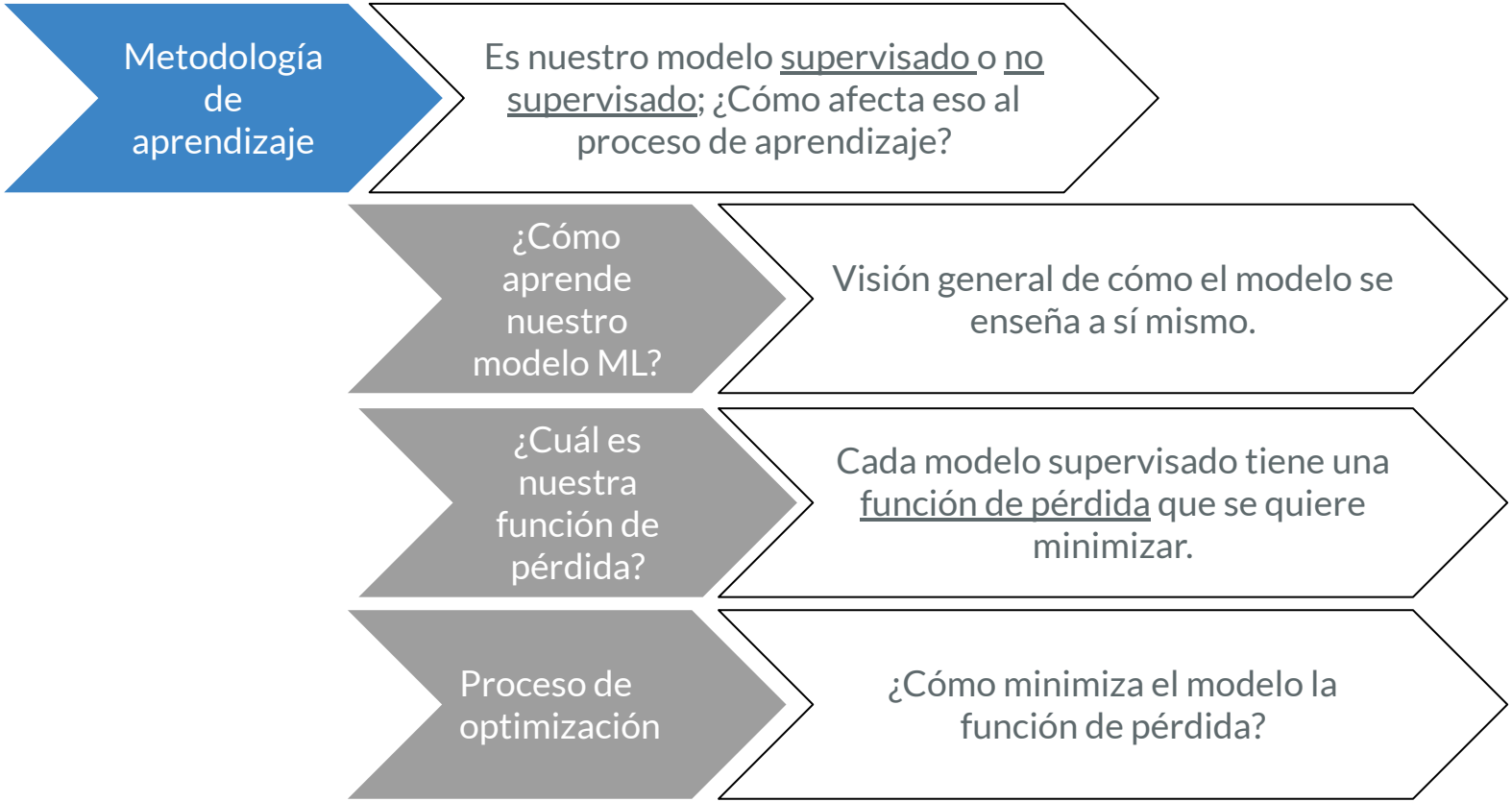


Fuente: Deep Learning Book - Chapter 5: Introduction to Machine Learning

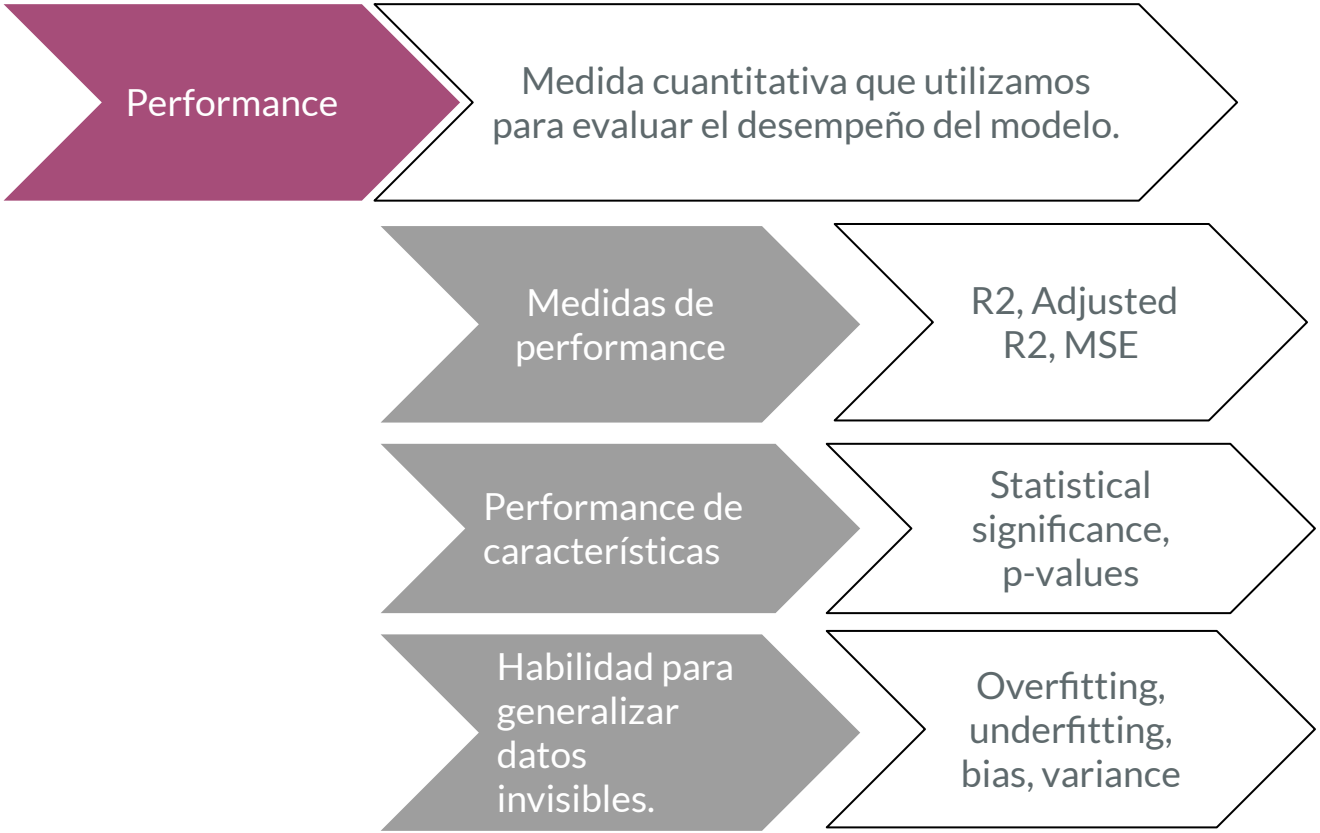
Aquí ya miramos más de cerca cada componente del dataframe:



Metodología de aprendizaje: ¿cómo aprende el modelo la función que mejor mapea x con la verdadera Y?

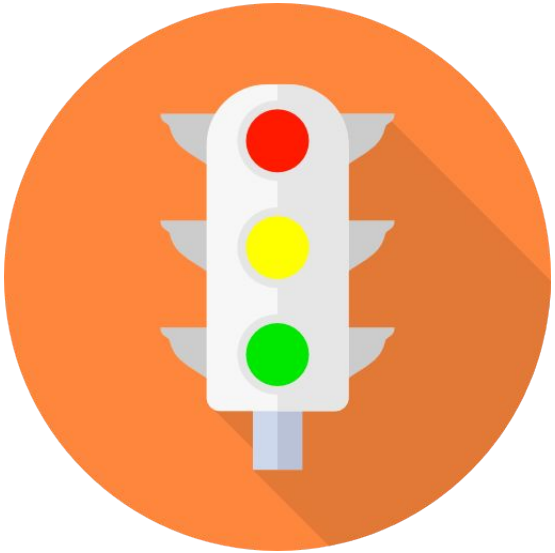


Desempeño: ¿Cómo evaluamos cuán útil es el modelo y cómo podemos mejorarlo?

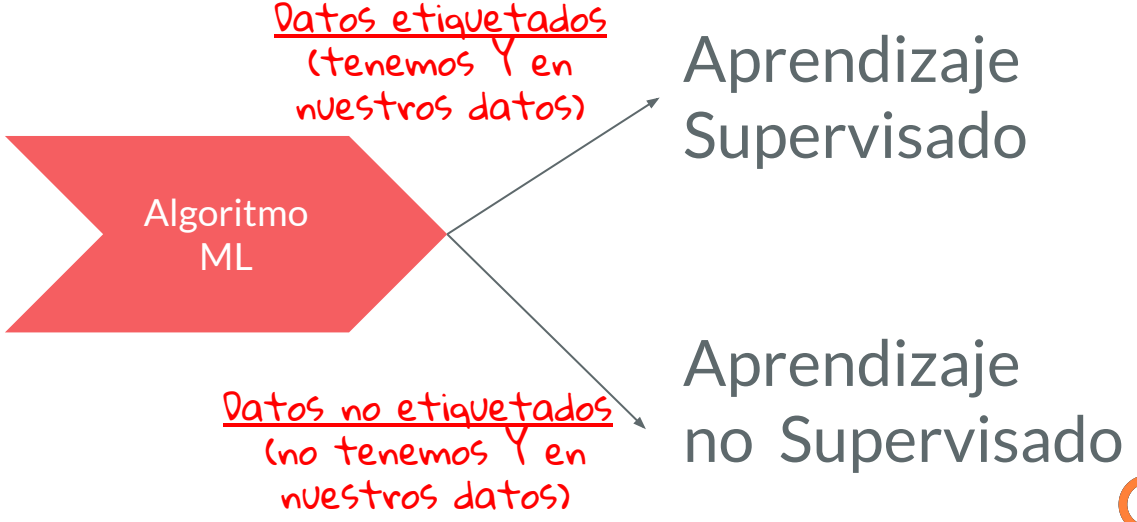


Tarea y metodología de aprendizaje.

¡Espera! Importante revelación:



Para las siguientes diapositivas, presentaremos la intuición detrás de los modelos de machine learning utilizando ejemplos de **aprendizaje supervisado**. Más adelante en este módulo, veremos cómo es diferente el aprendizaje no supervisado.



1. Tarea

Tarea

¿Cuál es el problema que queremos que resuelva nuestro modelo?

Pregunta de Investigación

Resumen: ¿Cómo varía el monto del préstamo solicitado en Kiva por ciudad en Kenia?

Tenemos datos de KIVA sobre el monto del préstamo solicitado por prestatarios en todo Kenia

Queremos saber cómo el monto del préstamo solicitado varía según la ciudad.

Tarea

Construir un modelo implica convertir tu pregunta de investigación en una pregunta de machine learning.

Pregunta de investigación → Tarea de Machine Learning

¿Cómo varía el monto del préstamo solicitado en Kiva por ciudad? → ??

Tarea

En primer lugar, establezcamos un vocabulario común para hablar sobre los datos.

Características

Observaciones

lender_count	loan_amount	location.country	location.country_code	location.geo.level	location.geo.pairs	location.geo.type	location.town	
7	225	Kenya	KE	town	-1.166667 36.833333	point	Kiambu	J
14	350	Kenya	KE	town	0.516667 35.283333	point	Eldoret	J
33	1075	Kenya	KE	town	1 38	point	Kakamega North	T C

Ubicación de la ciudad es un ejemplo de una característica. Cada columna en nuestro conjunto de datos es una característica

Cada fila de nuestro conjunto de datos es una observación.

Fuente:: Subset de Kenyan loans dataset.

Tarea

Una tarea de Machine Learning tiene características explicativas y una característica de salida o resultado.

Características explicativas

El prestatario del pueblo vive en

Característica de salida

Cantidad del préstamo solicitado

¿Cuáles serían las características de resultado y las características explicativas en las preguntas de investigación a continuación?

Intenta identificar algunas:

- ¿Cuál será el precio de una acción mañana?
- ¿Este paciente tiene malaria?
- ¿Esta persona compraría un carro?

La función de resultado podría ser una regresión (por ejemplo, $ 12) o una clasificación (por ejemplo, Sí o No). Hablaremos de esto más tarde

Soluciones:

Características explicatorias	Características de salida
Precio de un índice del mercado de valores hoy	El precio de las acciones de la compañía X mañana
Edad, síntomas, historia de viaje	Si un paciente tiene malaria o no
Ingresos, ubicación	Si una persona compraría o no un coche

Tarea

Definamos nuestras características explicativas y de resultados para esta tarea

***Problema*:**
Soy el alcalde de una ciudad de 30,000 personas y necesito justificar el presupuesto de gasto en mosquiteras.

Quiero pruebas de cómo el número de mosquiteros afecta al número de casos de malaria. ¿Puedes ayudar?

1.Pregunta de investigación

2.Tarea

Comencemos por identificar la pregunta de investigación

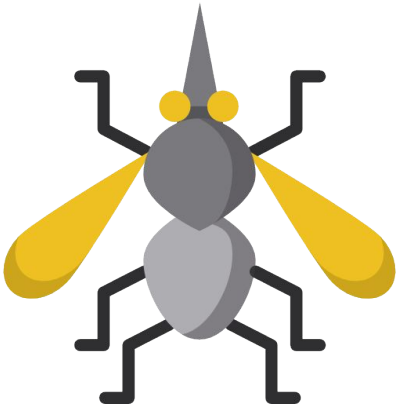
La pregunta de investigación es qué queremos averiguar a partir de los datos, formalmente establecidos.

¿Cómo cambia el número de casos de malaria cuando cambia el número de mosquiteros?

Tarea

A continuación definamos nuestra tarea

1. Definir característica explicativa y de salida.
2. Definir f(x)
3. Juntar todo



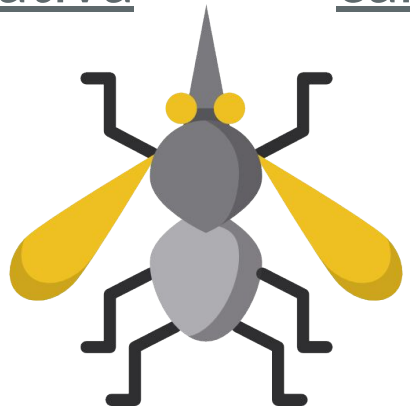
Definir característica explicativa y de resultado.

¿Cómo cambia el número de casos de malaria cuando cambia el número de mosquiteros?

característica(s) explicativa

x Número de mosquiteras

2007: 1000
2008: 2200
2009: 6600
2010: 12600



característica de resultado

Y Número de personas con malaria

2007: 80
2008: 40
2009: 42
2010: 35

También llamamos a nuestras características explicativas x, y nuestra característica de resultado Y.
Parece que a medida que las mosquiteras aumentan, el número de casos de malaria disminuye.

Tarea

¿Qué concluirías al mirar estos datos? ¿Cuántas redes recomendarías?

x	Número de mosquiteras
2007: 1000	
2008: 2200	
2009: 6600	
2010: 12600	

Y	Número de personas con malaria
2007: 80	
2008: 40	
2009: 42	
2010: 35	

Llegaste a una conclusión al **reconocer un patrón en los datos**. Esto es similar a cómo un algoritmo de aprendizaje automático abordaría el mismo problema.

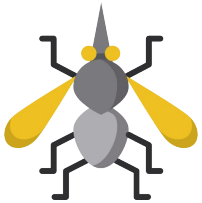
Tarea

El aprendizaje automático nos permite aprender de patrones históricos.

Si el Sr. Alcalde no tuviera métodos de aprendizaje automático, podría encontrar una respuesta probando un número diferente de redes año tras año.

Pero esto tiene un **costo humano** obvio, y sería muy difícil actualizar el modelo para dar cuenta, por ejemplo, de los nuevos residentes de su ciudad.

Los algoritmos de Machine learning ayudan a responder preguntas sin este costo humano - estamos **aprendiendo desde los datos**, o en otras palabras, **aprendiendo de la historia**!



Intuición humana

"Durante cuatro años, un número creciente de mosquiteros disminuye el número de casos de malaria".

Modelo de ML

Un aumento en x (mosquiteros) causa una disminución en Y (casos de malaria).

- Los seres humanos forman reglas basadas en la observación y el reconocimiento de patrones..
- El modelo ML toma la entrada x y la asigna a la salida Y.

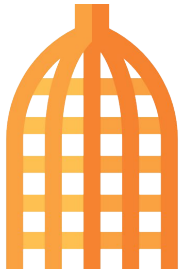
Definir f(x)

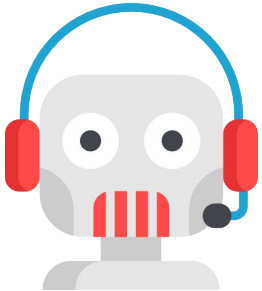
Nuestro modelo f (x) es una función que asigna nuestra entrada x a una Y * predicha.

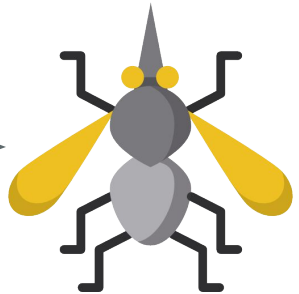
característica(s) explicativa(s) — modelo — resultado predicho

x Número de mosquiteras → f(x) + e → Y* Número previsto de personas con malaria

error irreducible







Definir f(x)

El objetivo de f (x) es predecir una Y * tan cerca de la verdadera Y como sea posible.

Nuestra función f(x) asigna una entrada x a una Y predicha, a la que nos referimos como Y*. Queremos elegir una f (x) que asigne x lo más cerca posible de la verdadera Y.

$$f(x) + e = Y^*$$

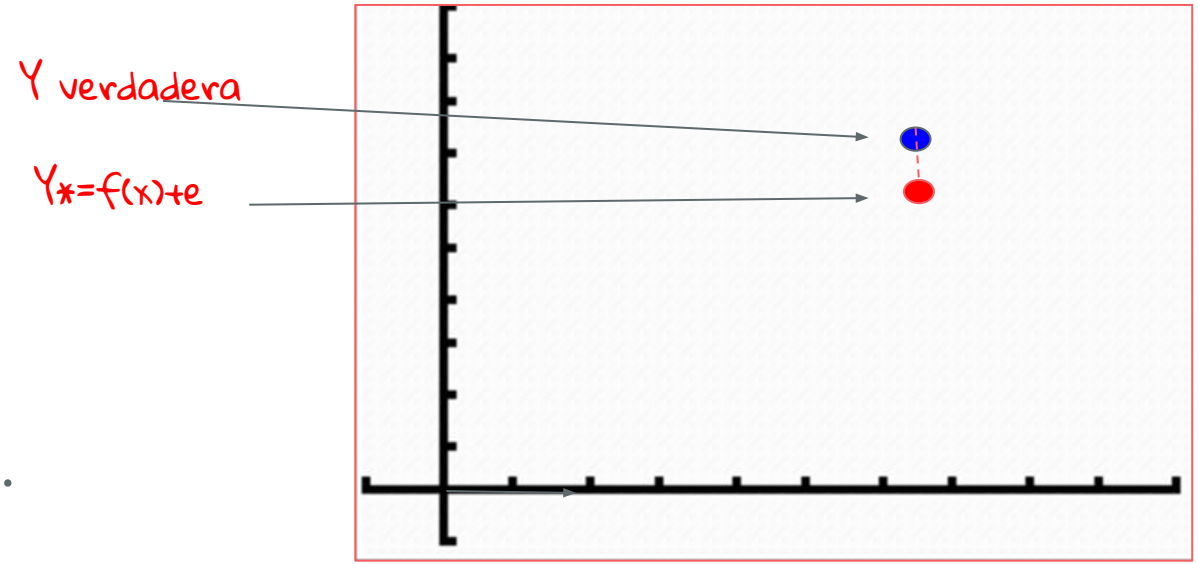
Número previsto de personas con malaria

e es un error irreducible. Esto captura el error causado por factores como el error de medición, la aleatoriedad en los datos y la elección inadecuada del modelo. No importa qué tan bien optimice su modelo, esto nunca se reducirá a 0.

Definir f(x)

Queremos que Y * esté cerca de la Y verdadera porque queremos que la función genere predicciones útiles.

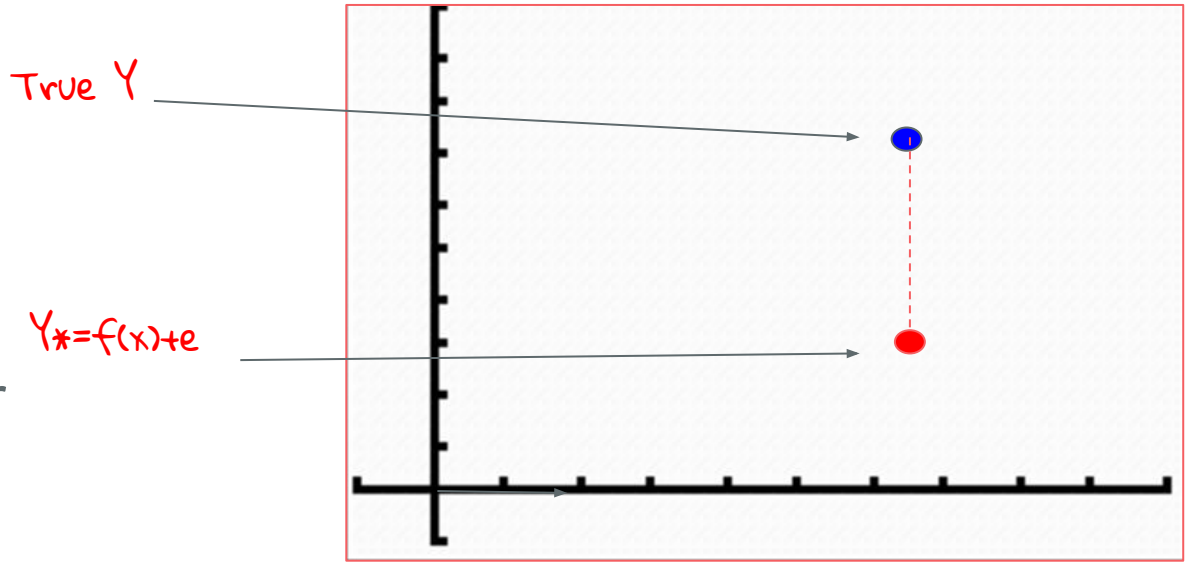
En este ejemplo, la Y predicha aparece cerca de la Y verdadera. Hablaremos sobre cómo cuantificar esto en la siguiente sección.



Definir f(x)

Queremos que Y * esté cerca de la Y verdadera porque queremos que la función genere predicciones útiles.

En este ejemplo, la Y predicha aparece lejos de la Y verdadera. ***Esto probablemente no sea muy útil.*** Hablaremos sobre cómo cuantificar esto en la siguiente sección.



Definir f(x)

¿Qué es f (x)? Depende del algoritmo de aprendizaje automático que elijamos.

x → **f(x)** → **Y***

característica(s) explicativa(s), como el número de mosquiteras

resultado predicho, por ej. Número de personas con malaria

Ejemplo de f(x):

Algoritmos de aprendizaje supervisado:

- Regresión lineal
- Árbol de decisiones
- Random forest
- ...

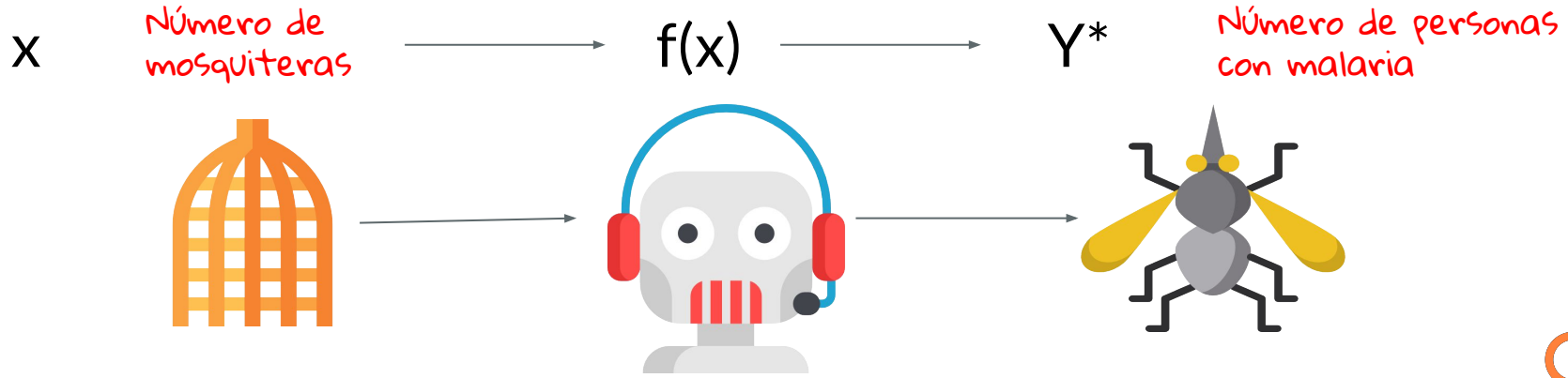
Juntar todo

Vamos a juntar todo.

Pregunta de investigación

¿Cómo cambia el número de casos de malaria cuando cambia el número de mosquiteros?

Tarea de Machine Learning



Tarea de aprendizaje supervisado

La función de tarea depende del tipo de datos que desea predecir. Los problemas de aprendizaje supervisados se dividen en dos categorías principales: regresión y clasificación.

La Tarea

Regresión

Variable continua

Un problema de regresión es cuando estamos tratando de predecir un **valor numérico**, como "costo" o "peso".

Clasificación

Variable categórica

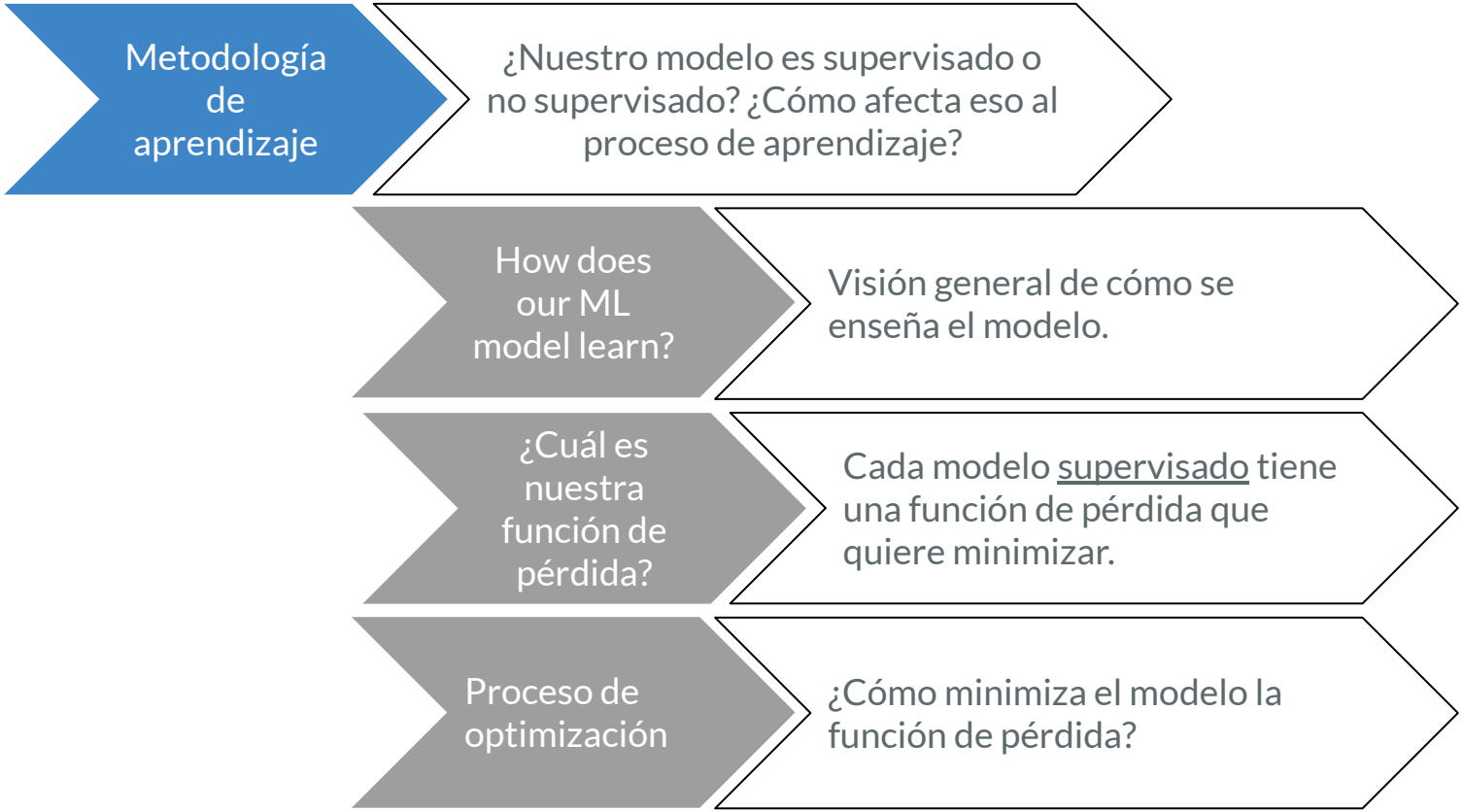
Un problema de clasificación es cuando estamos tratando de predecir si algo **pertenece a una categoría**, como "rojo" o "azul" o "enfermedad" y "no enfermedad".

Fuente: Andrew Ng, Stanford CS229 Machine Learning Course

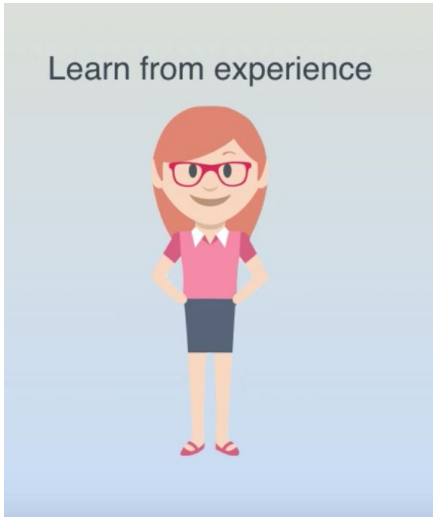
Metodología de aprendizaje

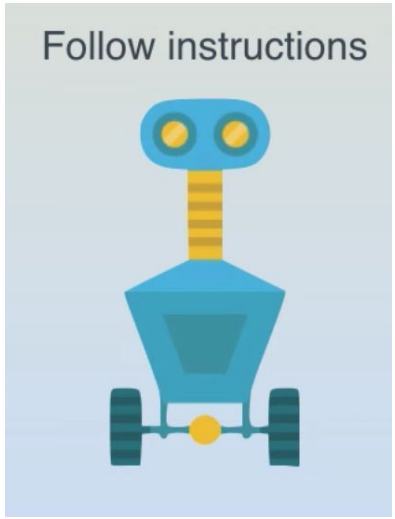
Metodología de aprendizaje	Los algoritmos ML pueden ser supervisados o no supervisados. Esto determina la metodología de aprendizaje.

Metodología de aprendizaje: ¿cómo aprende el modelo la función que mejor mapea x con la verdadera Y?

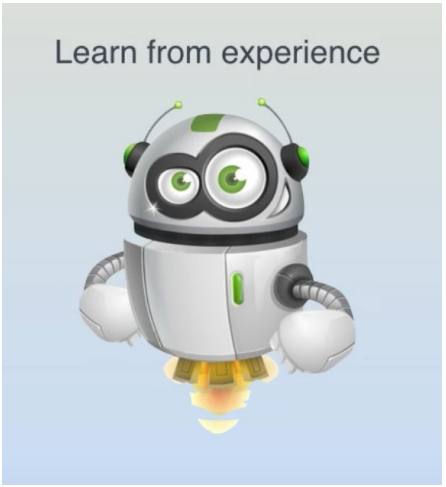


Recuerda que el Machine Learning es un subconjunto de la inteligencia artificial que permite a las máquinas aprender desde los datos.





Machine learning



La programación de software tradicional implica dar a las máquinas las instrucciones para las acciones que realizan. **Machine Learning implica permitir que las máquinas aprendan de datos en bruto para que el programa computacional pueda cambiar cuando se expone a nuevos datos (aprendiendo de la experiencia).**

Source: https://www.youtube.com/watch?v=IpGxLWOIZy4

Metodología de aprendizaje

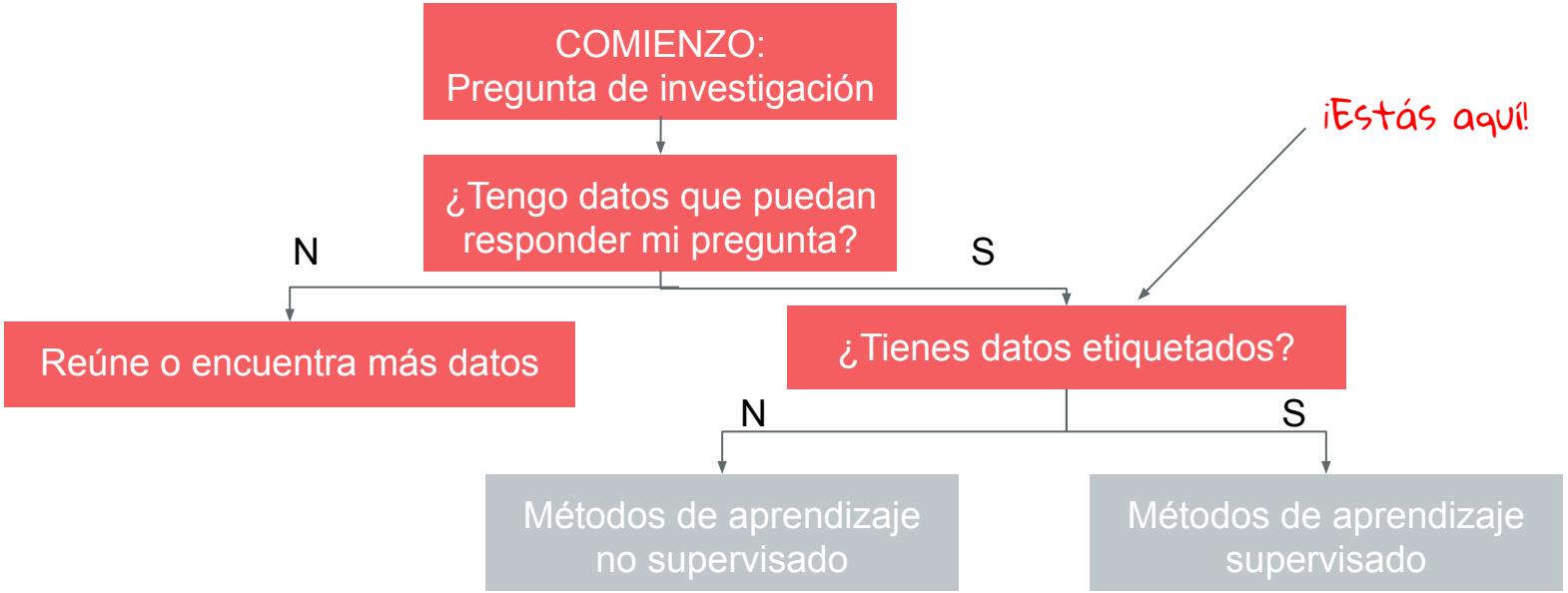
¿Qué queremos decir cuando decimos que una máquina "aprende de la experiencia"?

Machine learning es un subconjunto de la IA que permite a las máquinas aprender de datos sin procesar.

¿Cómo aprende el modelo de los datos en bruto?

Metodología de aprendizaje

La forma en que el algoritmo aprende depende del tipo de datos que tenga.



Metodología de aprendizaje

¿Qué significan los datos etiquetados?

¿Tienes datos etiquetados?

Si

La característica de salida (Y) que te interesa predecir se encuentra registrada en los datos. Si tienes Y etiquetada, puede usar métodos de aprendizaje supervisado.

Y = Número de personas con malaria

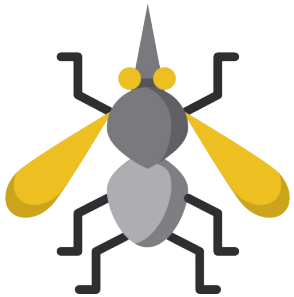
2007: 80
2008: 40
2009: 42
2010: 35

No

La característica de salida (Y) no se encuentra registrada en los datos. No tienes Y etiquetado.

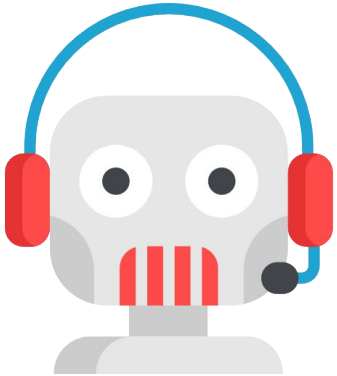
Y = Número de personas con malaria

2007:
2008:
2009:
2010:



Metodología de aprendizaje

Si tienes o no datos etiquetados, determina si se trata de un problema de aprendizaje supervisado o no supervisado



¿Tienes datos etiquetados?

Y está en tus datos

Si

No

Y no está en tus datos

$f(x) + e = Y^*$

Aprendizaje Supervisado

- Para cada x, existe un Y
- El objetivo es **predecir** Y usando x

Aprendizaje No Supervisado

- Para cada x, no existe un Y
- El objetivo no es predecir, sino **investigar** x

La mayoría de los problemas con los que te encontrarás inicialmente son algoritmos supervisados.

¿Cómo aprenden los algoritmos supervisados?

aprendizaje supervisado

Explicación intuitiva de cómo los algoritmos supervisados aprenden:

Y Número de personas con malaria

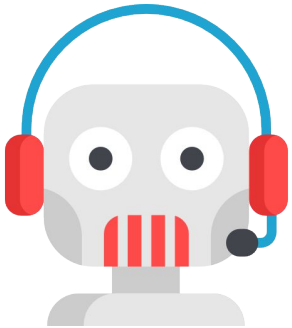
2007: 80
2008: 40
2009: 42
2010: 35

Imagina que eres profesor y le haces una pregunta a tus alumnos.

Las etiquetas Y proporcionan la respuesta correcta al problema que los estudiantes intentan resolver. **Ya que sabes la respuesta correcta, puedes recompensar el buen desempeño de los alumnos y castigar el mal desempeño.** Esto fomenta el aprendizaje continuo

aprendizaje supervisado

Extendiendo este ejemplo, usted (el investigador) es el maestro y el Modelo es el estudiante.



Modelo

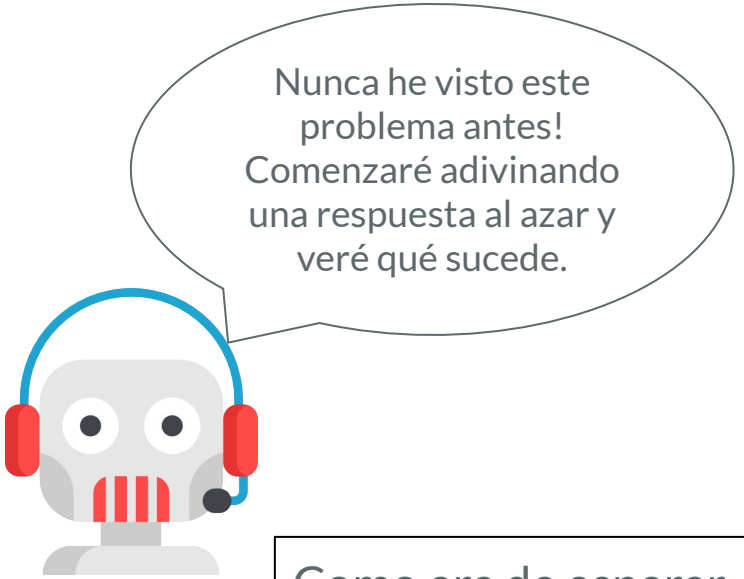
Quiero obtener la respuesta correcta para predecir Y y ser el mejor estudiante de la clase.

¡Muy bien Sr. Modelo! Una vez que me des tu respuesta, te haré saber la respuesta correcta.

Cada vez que el Sr. Modelo predice Y *, comparas Y * con la Y verdadera para ver qué tan bien lo hizo.

aprendizaje supervisado

Nuestro modelo comienza a tratar de proporcionar un Y * estimado adivinando.



Modelo

Y*	Y
Número previsto de personas con malaria	Número real de personas con malaria
2007: 1	2007: 80
2008: 2000	2008: 40
2009: 300	2009: 42
2010: 40	2010: 35

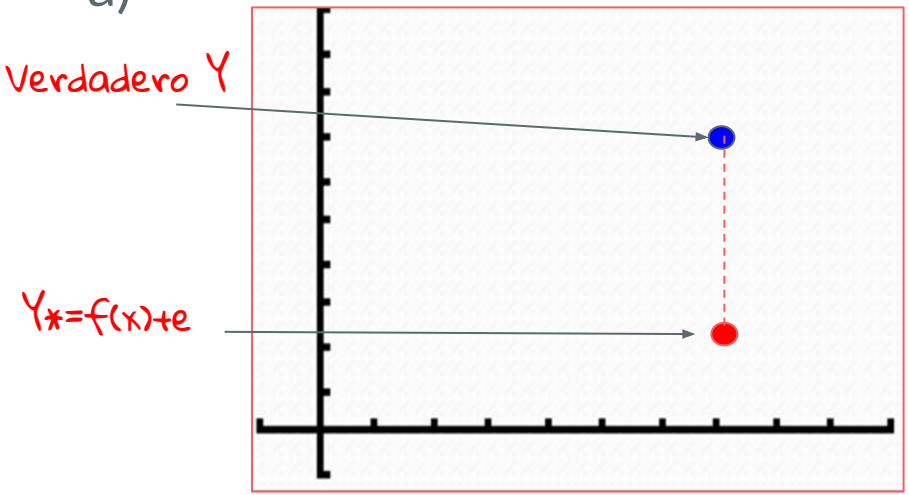
Como era de esperar, los resultados parecen terribles, a juzgar por el hecho de que los números reales son muy diferentes de los números predichos. **Para cuantificar qué tan malos o buenos son los resultados, utilizamos Y-Y *.**

Metodología de aprendizaje

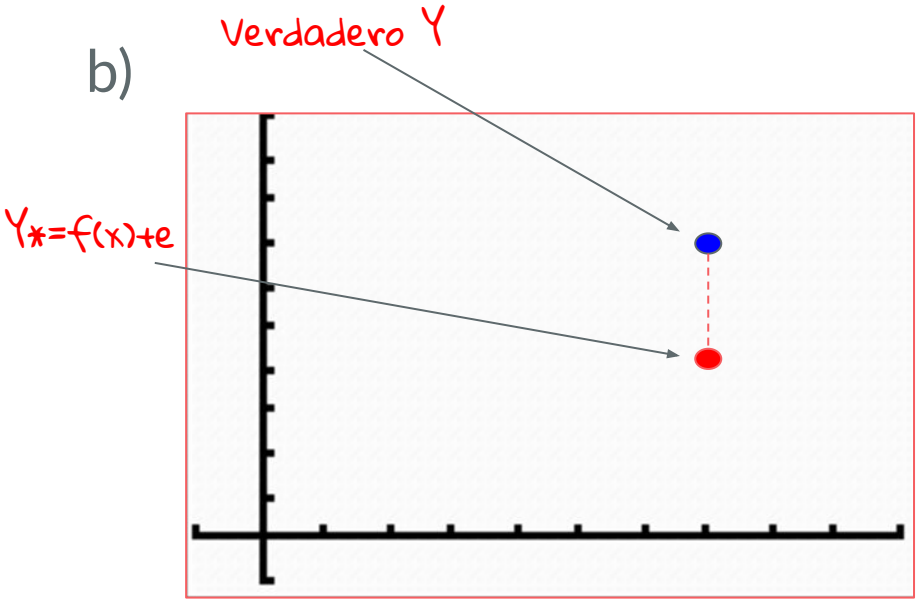
Cuál es nuestra función de pérdida?

¿Qué modelo es más útil para mapear x cerca del Y verdadero?

a)



b)

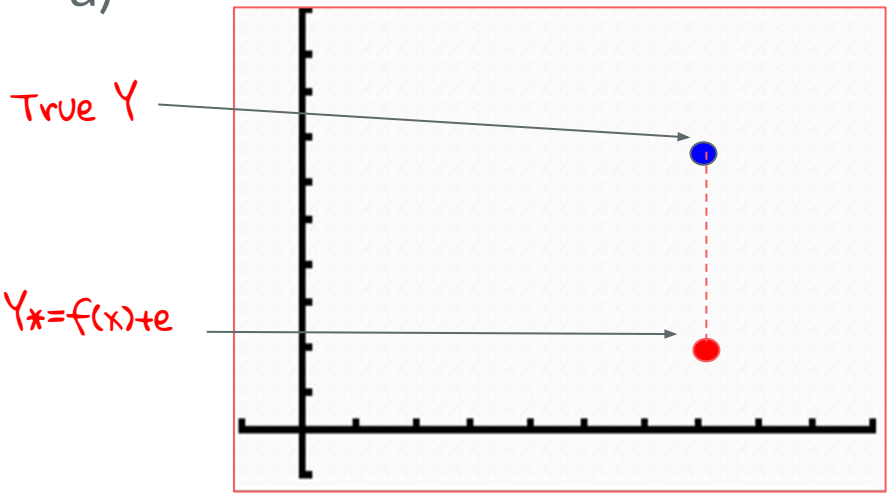


¿Qué predicción fue peor, a) o b)?

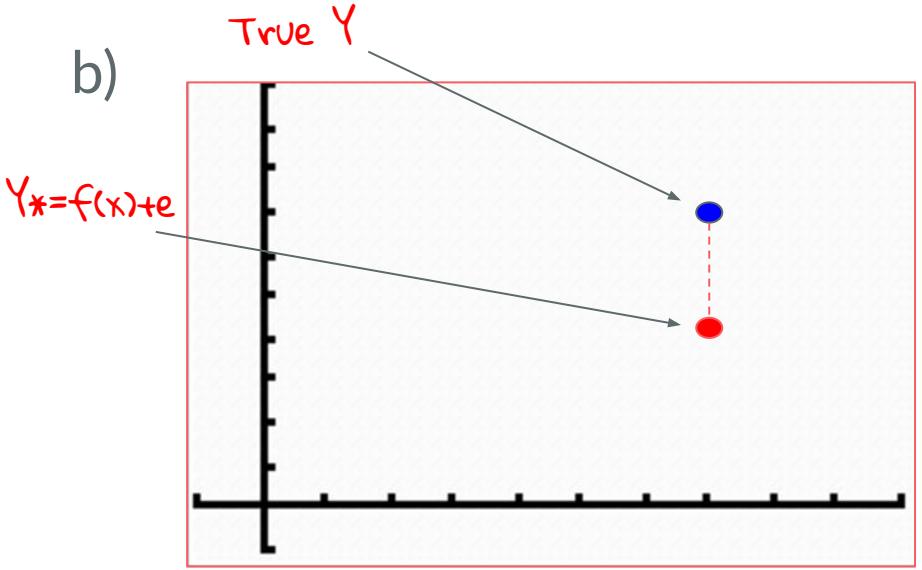
¡Podemos decir de inmediato que b es mejor!

Podemos ver que f (x) en b mapea x a una Y * mucho más cerca de la verdadera Y. Una función de pérdida nos permite cuantificar esta diferencia.

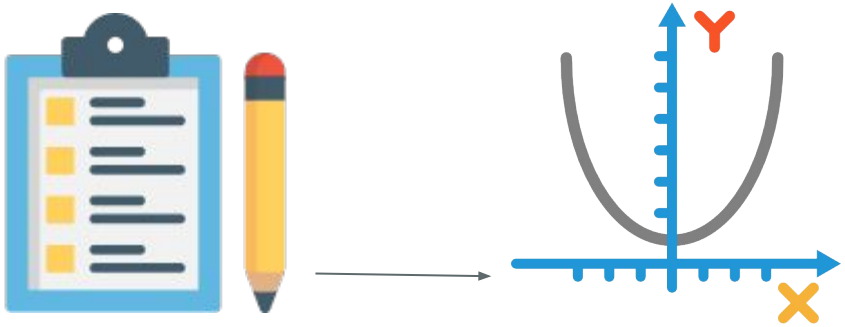
a)



b)



El objetivo de un modelo es minimizar la función de pérdida.



La Tarea

La Función de Pérdida

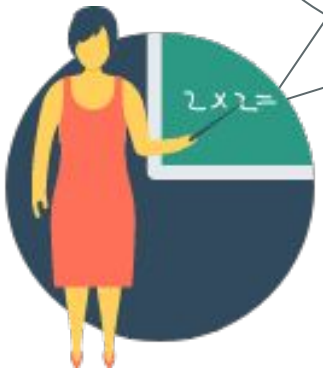
Una función de pérdida cuantifica qué tan insatisfecho estarías si usaras f (x) para predecir Y * cuando la salida correcta es Y. Es lo que queremos minimizar.

Otra forma de pensarlo es que una función de pérdida cuantifica qué tan bien nuestra f (x) se ajusta a nuestros datos.

Ya hemos visto un ejemplo simple: Y-Y *, o la diferencia entre la Y predicha y la Y real. Más adelante, veremos funciones de pérdida más sofisticadas.

Fuente: Stanford ML Lecture 1

aprendizaje supervisado



Ya que sé la respuesta correcta, puedo comparar la predicción de Y * con la verdadera Y para ayudar a guiar al Sr. Modelo.

Y*	Y
Número previsto de personas con malaria	Número real de personas con malaria
2007: 1	2007: 80
2008: 2000	2008: 40
2009: 300	2009: 42
2010: 40	2010: 35

Primero decidimos cómo medir cuán insatisfechos estamos con estos resultados. A esto le llamamos nuestra **función de pérdida**. En la siguiente diapositiva, mostraremos algunas funciones de pérdida posibles diferentes que podemos usar para evaluar al Sr. Modelo.

Tarea de aprendizaje supervisado

Recuerda que hay dos tipos diferentes de tareas:

La Tarea

Regresión

Variable continua

Un problema de regresión es cuando estamos tratando de predecir un **valor numérico**, como "costo" o "peso".

Clasificación

Variable categórica

Un problema de clasificación es cuando estamos tratando de predecir si algo **pertenece a una categoría**, como "rojo" o "azul" o "enfermedad" y "no enfermedad".

Fuente: Andrew Ng, Stanford CS229 Machine Learning Course

Metodología de aprendizaje

¿Cuál es nuestra función de pérdida?

La elección de la función de pérdida depende del tipo de tarea. Discutiremos las funciones de pérdida para ambos tipos de tareas.

Regresión

Variable continua

- Error absoluto (L1)
- Error de mínimos cuadrados (L2)

Incluyendo error cuadrático medio (MSE), error cuadrático medio (RMSE)

Clasificación

Variable categórica

- log loss
- hinge loss

Fuente: Andrew Ng, Stanford CS229 Machine Learning Course

Metodología de aprendizaje

¿Cuál es nuestra función de pérdida?

Hay algunas funciones de pérdida diferentes que podríamos elegir, dependiendo del problema que estamos tratando de resolver.

Regresión

Variable continua

error absoluto (L1)	raíz cuadrada del error cuadrático medio (RMSE)
error de mínimos cuadrados (L2)	error cuadrático medio (MSE)

Clasificación

Variable continua

log loss
hinge loss

Funciones de pérdida de la regresión.

1. L1 norm (error absoluto medio)
2. L2 norm (error de mínimos cuadrados)
    - Error cuadrático medio

Metodología de aprendizaje

¿Cuál es nuestra función de pérdida?

Nuestra característica de salida es continua: la cantidad de personas que tienen malaria.

La característica de salida en los datos es continua → Tarea de regresión → Función de pérdida L1 o L2

Y

Número de personas con malaria

2007: 80
2008: 40
2009: 42
2010: 35

Metodología de aprendizaje

¿Cuál es nuestra función de pérdida?

L1 y L2 son dos opciones posibles para evaluar cuán descontentos estamos con la elección de f (x) del Sr. Modelo.

error absoluto (L1)

También llamada pérdida L1, ésta minimiza la **suma** de errores absolutos entre la Y verdadera y la Y* predicha.

$$S = \sum_{i=1}^{n} |y_i - f(x_i)|.$$

error de mínimos cuadrados (L2)

También llamada pérdida L2, ésta minimiza el **cuadrado** del error entre la Y verdadera y la Y * predicha.

$$S = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

error cuadrático medio

Toma la media de la pérdida L2 sobre todas las observaciones.

Fuente: L1 and L2

Metodología de aprendizaje

¿Cuál es nuestra función de pérdida?

¿Qué tan malos fueron los resultados iniciales del Sr. Modelo? Vamos a calcular la norma L1.

¿Qué tan bien funcionó la suposición aleatoria del Sr. Modelo?

$$S = \sum_{i=1}^{n} |y_i - f(x_i)|.$$

error absoluto (L1)

Y*
Número previsto de personas con malaria

2007: 1
2008: 2000
2009: 300
2010: 40

Y
Número real de personas con malaria

2007: 80
2008: 40
2009: 42
2010: 35

(|1-80|+|2000-40|+|300-42|+|40-35|) = 2,302

Fuente: Intro to Stat - Introduction to Linear Regression

Metodología de aprendizaje

¿Cuál es nuestra función de pérdida?

¿Qué tan malos fueron los resultados iniciales del Sr. Modelo? Vamos a calcular la norma L2.

¿Cómo estuvo la estimación aleatoria inicial del Sr. Model?

$$S = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

error de mínimos cuadrados (L2)

Y* Número previsto de personas con malaria	Y Número real de personas con malaria
2007: 1	2007: 80
2008: 2000	2008: 40
2009: 300	2009: 42
2010: 40	2010: 35

(80-1)^2+(40-2000)^2+(42-300)^2 +(35-40)^2=3,914,430

Fuente: Intro to Stat -Introduction to Linear Regression

Metodología de aprendizaje

¿Cuál es nuestra función de pérdida?

Podemos normalizar nuestra pérdida de L2 calculando el error cuadrático medio o la raíz del error cuadrático medio.

error de mínimos cuadrados (L2)

También llamada pérdida L2, ésta minimiza el cuadrado del error entre la Y verdadera y la Y * predicha.

$$S = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

error cuadrático medio

Toma la media de la pérdida L2 sobre todas las observaciones.

MSE = mean(S)

raíz del error cuadrático medio

Toma la raíz cuadrada de la media de la pérdida de L2.

RMSE = sqrt(mean(S))

Fuente: L1 and L2,

Metodología de aprendizaje

¿Cuál es nuestra función de pérdida?

El error cuadrático medio toma el error L2 promedio por observación.

¿Cómo estuvo la estimación aleatoria inicial del Sr. Model?

$$MSE = mean(S)$$

error cuadrático medio

Y*	Y
Número previsto de personas con malaria	Número real de personas con malaria
2007: 1	2007: 80
2008: 2000	2008: 40
2009: 300	2009: 42
2010: 40	2010: 35

((80-1)^2+(40-2000)^2+(42-300)^2
+(35-40)^2)/4
= 978,607.5

Fuente: Intro to Stat -Introduction to Linear Regression

Metodología de aprendizaje

¿Cuál es nuestra función de pérdida?

La raíz del error cuadrático medio toma la raíz cuadrada del error L2 promedio por observación.

¿Cómo estuvieron los modelos aleatorios iniciales?

RMSE = sqrt(mean(S))

raíz del error cuadrático medio

Y*	Y
Número previsto de personas con malaria	Número real de personas con malaria
2007: 1	2007: 80
2008: 2000	2008: 40
2009: 300	2009: 42
2010: 40	2010: 35

(((80-1)^2+(40-2000)^2+(42-300)^2+(35-40)^2)/4)^(½)
= 989.25

Fuente: Intro to Stat -Introduction to Linear Regression

Metodología de aprendizaje

Podemos calcular para cada una de las funciones de pérdida lo insatisfechos que estamos con los modelos de estimación aleatoria inicial.

No te preocupes por estos números. Lo importante es que entiendas cómo los estamos transformando paso a paso.

2,302	3,914,430	978,608	989
error absoluto (L1)	error de mínimos cuadrados (L2)	error cuadrático medio	raíz del error cuadrático medio
También llamada pérdida L1, ésta minimiza la **suma** de errores absolutos entre la Y verdadera y la Y* predicha.	También llamada pérdida L2, ésta minimiza el cuadrado del error entre la Y verdadera y la Y * predicha.	Toma a media de la pérdida L2 por observación en los datos.	Toma la raíz cuadrada de la pérdida de L2 promedio por observación en los datos.
$S = \sum_{i=1}^{n} \lvert y_i - f(x_i) \rvert.$	$S = \sum_{i=1}^{n} (y_i - f(x_i))^2$	MSE = mean(S)	RMSE = sqrt(mean(S))

Fuente: L1 and L2

Metodología de aprendizaje

RMSE es la raíz cuadrada de la pérdida media de L2 por observación.

Modelo

Hay cinco pasos para RMSE:

Y-Y*	Para cada observación en nuestro conjunto de datos, mide la diferencia entre la Y verdadera y la Y predicha.
^2	Cuadrar cada Y-Y * para obtener la distancia absoluta, así los valores positivos no cancelan los negativos cuando sumamos.
Sum	Suma todas las observaciones para obtener el error total.
mean	Divide la suma por el número de observaciones que tenemos
root	Tome la raíz cuadrada de la media calculada anteriormente.

MSE

Fuente: Intro to Stat -Introduction to Linear Regression

¿Qué función de pérdida deberíamos usar?

1. L1 norm (error absoluto medio)
2. L2 norm (error de mínimos cuadrados)

Cada función de pérdida tiene ventajas y desventajas importantes.

error absoluto (L1) vs. error de mínimos cuadrados (L2)

	¿Robusto?	Stable Solution?	How many solutions?
L1	Robusto	No estable	Múltiples soluciones posibles
L2	No muy robusto	Estable	Una splución posible

MSE y RMSE son versiones normalizadas del error L2. Si decidimos utilizar L2, elegiremos MSE o RMSE.

Fuente: L1 and L2

Si decidimos utilizar el error de mínimos cuadrados (L2), podemos usar RMSE o MSE

MSE vs. RMSE

La diferencia clave entre RMSE y MSE es que tomar la raíz en RMSE normaliza el error a las mismas unidades de medida.

Esto hace que el término de error sea más interpretable.

Tanto MSE como RMSE amplifican y penalizan severamente los errores grandes más que los pequeños al elevar al cuadrado el error.

Funciones de pérdida en clasificación

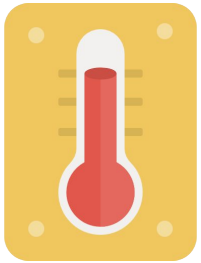
1. Log loss
2. Hinge loss

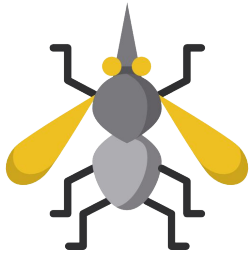
Definir característica explicativa y de resultado.

Definamos una tarea ligeramente diferente para poder discutir las pérdidas hinge y log.

Tarea: Queremos predecir si un paciente tiene malaria usando su temperatura.

X	Y
Temperatura del paciente	¿El paciente tiene malaria?
39.5°C	No
37.8°C	Yes
37.2°C	Yes
37.2°C	No





Metodología de aprendizaje

¿Cuál es nuestra función de pérdida?

Nuestra característica de salida es categórica: queremos predecir si alguien tiene malaria o no. Este es un problema de clasificación binario.

Esta es una tarea de clasificación, por lo que podemos usar la pérdida log o hinge.

Pero primero, ¿qué es una tarea de clasificación?

Clasificación basada en el umbral de probabilidad

Las tareas de clasificación generan la **probabilidad de pertenecer a una clase**. Normalmente, en función de un umbral del 50%, asignamos la clase predicha.

característica de salida		probabilidad predicha	resultado predicho	
Y	¿El paciente tiene malaria?	¿Cuál es la probabilidad de que el paciente tenga malaria?	Y*	¿Predice el modelo que el paciente tiene malaria?
No		0.55	Yes	
Yes		0.80	Yes	
Yes		0.85	Yes	
No		0.2	No	

Clasificación basada en el umbral de probabilidad

Podemos evaluar la precisión observando solo el resultado predicho frente al resultado real. Aquí, la precisión es del 75%!

característica de salida	probabilidad predicha	resultado predicho
Y: ¿El paciente tiene malaria?	¿Cuál es la probabilidad de que el paciente tenga malaria?	Y*: ¿Predice el modelo que el paciente tiene malaria?
No	0.55	Yes
Yes	0.80	Yes
Yes	0.85	Yes
No	0.2	No

Sin embargo, **nos estamos perdiendo en el uso de la probabilidad**, que es información importante acerca de cuán cierto es el modelo acerca de su predicción. Veamos algunas funciones de pérdida que utilizan esta métrica.

Log loss

Para cada predicción que hace el modelo, podemos medir la pérdida logarítmica. **¿Qué es la pérdida logarítmica?**

Modelo

Log loss evalúa la probabilidad de pertenecer a una clase

$$L = -\frac{1}{n}\sum_{i=1}^{n}[y_i log(\hat{y}_i) + (1-y_i) log(1-\hat{y}_i)]$$

- Cuanto menor sea la pérdida logarítmica, menor será la incertidumbre, mejor será el modelo
  - Un clasificador perfecto tendría un log loss = 0
- Log loss penaliza en gran medida a los clasificadores que confían en una clasificación incorrecta

- Maneras de mejorar la pérdida de registro:
  - ¿Hay errores problemáticos en los datos?
  - ¿Queremos suavizar las probabilidades?

Fuente: https://www.r-bloggers.com/making-sense-of-logarithmic-loss/

Hinge loss

Para cada predicción de nuestro modelo, también podemos medir el hinge loss. **¿Qué es hinge loss?**

Hinge loss es la extensión lógica de la función de pérdida de regresión, **pérdida absoluta**.

Pérdida absoluta: Y- Y*, donde Y e Y* son enteros.

Hinge loss: max(0,1-(Y*)(Y))

Donde Y puede ser igual a -1 (no) o 1 (yes) para casa clase.

Para cada observación, si Y* == Y (ambos son 1 o ambos son -1), hinge loss = 0. If Y =/= Y*, hinge loss **se incrementa**.

La hinge loss acumulada es, por lo tanto, el límite superior del número de errores cometidos por el clasificador.

Fuentes: https://en.wikipedia.org/wiki/Hinge_loss;
http://scikit-learn.org/stable/modules/generated/sklearn.metrics.hinge_loss.html

Metodología de aprendizaje

¿Cuál es nuestra función de pérdida?

¿Cómo elegimos una función de pérdida para un problema de clasificación?

Log Loss

Conduce a **probabilidades más exactas**, pero a costa de la precisión

VS.

Hinge Loss

Conduce a una **mayor precisión**, pero a costa de probabilidades exactas

¿Cómo elegimos una función de pérdida para un problema de clasificación?

Depende de la pregunta que quieras responder!

Por ejemplo, para un problema en el que estamos tratando de evaluar la salud del paciente, sabemos que los falsos positivos (el modelo predice que usted tiene malaria, pero en realidad no la tiene) son más seguros y generalmente más preferibles que los falsos negativos (el modelo predice que no tienes malaria, pero en realidad la tienes.)

Por lo tanto, probablemente sea más seguro evaluar nuestra producción como una **probabilidad** de que usted tenga malaria o no. Vamos a utilizar el **log loss**.

Una nota sobre las funciones de pérdida categóricas ...

Proporcionamos sólo una amplia descripción conceptual de las pérdidas log loss e hinge ya que no los utilizaremos en nuestro laboratorio de codificación. Sin embargo, te animamos a explorarlas más a fondo.

Se pueden encontrar más recursos al final de este módulo..

Nuestro modelo calculó una aproximación inicial utilizando RMSE. **¿Cómo puede mejorar en su aproximación inicial?**

Metodología de aprendizaje

¿Cuál es nuestra función de pérdida?

Nuestro RMSE inicial es muy alto. Nuestro modelo intenta un f (x) diferente y compara RMSE.

	Suposición inicial	2nd actualización	3rd actualización
RMSE	1,000	1,300	800
# redes	300	100	400

Si el nuevo f(x) reduce la pérdida, **nuestro modelo cambia constantemente la f (x) en esa dirección.** Después de cada cambio, el modelo mide si la pérdida ha aumentado, disminuido o se ha mantenido igual.

A medida que el modelo actualiza la predicción en cada paso, vemos que el modelo está **aprendiendo**.

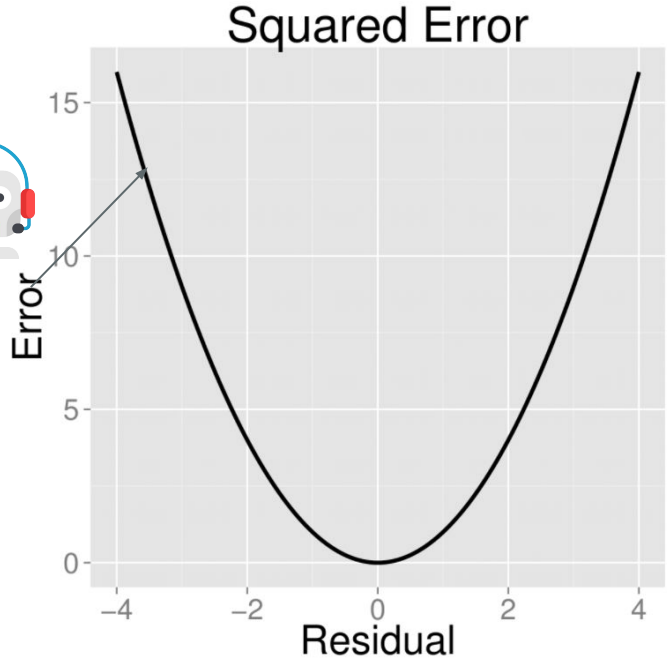
Fuente: Intro to Stat -Introduction to Linear Regression

Metodología de aprendizaje

¿Cómo aprende nuestro modelo ML?

El proceso de cambiar f (x) para reducir la función de pérdida se denomina **aprendizaje**. Es lo que hace que la regresión de mínimos cuadrados ordinarios (OLS) sea un algoritmo de aprendizaje automático..

El f (x) inicial **aleatorio** de nuestro modelo nos da un alto error inicial

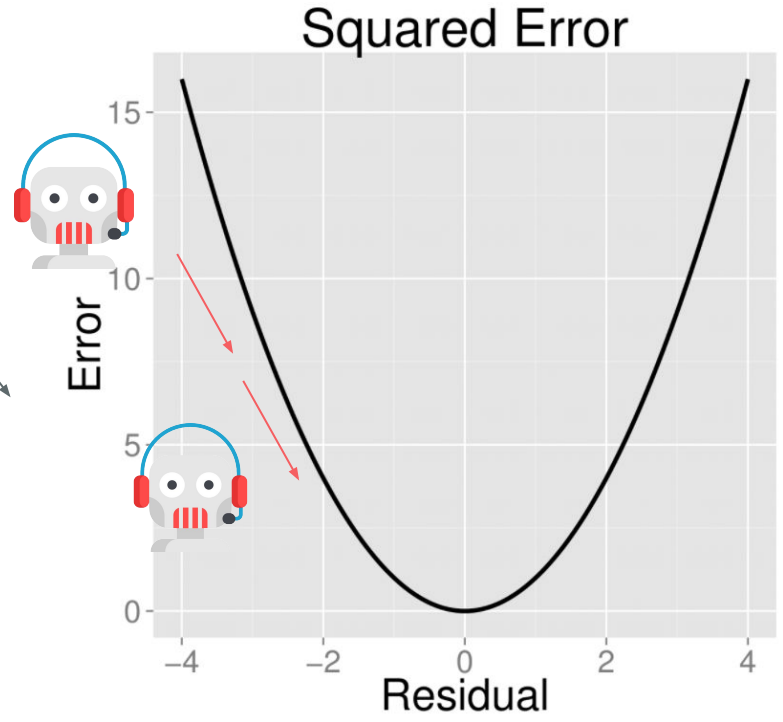


Para cada f (x) elegimos hay una pérdida asociada.

El proceso de aprendizaje implica la actualización de f (x) para alcanzar la pérdida mínima global.

Nuestro modelo comienza con un f (x) aleatorio y una actualización de f (x) para que nuestra pérdida sea lo más pequeña posible.

El trabajo de nuestro modelo es cambiar los parámetros para que cada vez que cambie f(x), la pérdida disminuya.

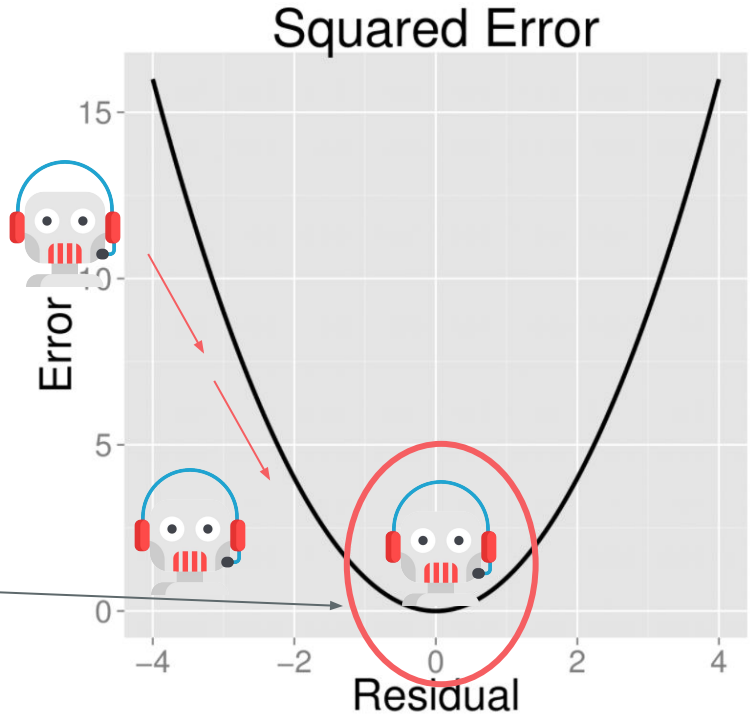


Nuestro modelo tiene éxito cuando reduce el error a su **mínimo**.

¿Cuándo se detiene nuestro modelo?

El trabajo de nuestro modelo es cambiar los parámetros para que cada vez que cambie f(x), la pérdida disminuya.

Nuestro modelo tiene éxito cuando reduce el error a su **mínimo**.



Metodología de aprendizaje

¿Cómo aprende nuestro modelo ML?

¿Qué pasa si no tenemos datos etiquetados?

¿Tienes datos etiquetados?

Si

La característica de salida (Y) que te interesa predecir está registrada en los datos. Tienes Y etiquetada, puedes usar métodos de aprendizaje supervisado.

Y=Número de personas con malaria

2007: 80
2008: 40
2009: 42
2010: 35

No

La característica de resultado (Y) no está registrada en los datos. No tienes Y etiquetada.

Y=Número de personas con malaria

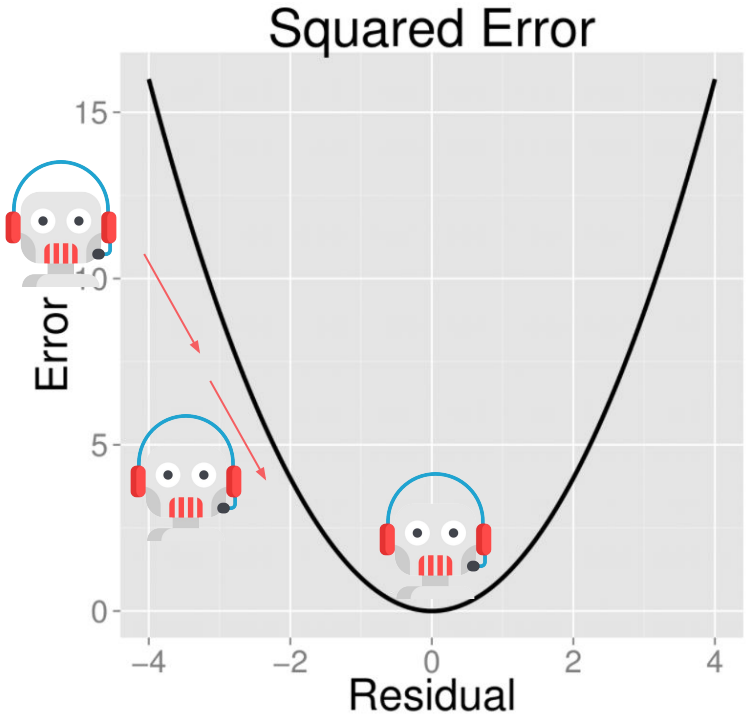
2007:
2008:
2009:
2010:

¿Cuándo se detiene nuestro modelo si no tenemos datos etiquetados?

Nuestro modelo puede actualizar los parámetros **sólo si tenemos datos etiquetados.**

¿Qué sucede si no tenemos una Y en nuestros datos?

Nos dirigimos a las técnicas de **aprendizaje no supervisado**.



Metodología de aprendizaje

El aprendizaje no supervisado no tiene datos etiquetados. Sin embargo, es el área de investigación actual más prometedora en aprendizaje automático. Desbloquear el aprendizaje sin supervisión cambiará fundamentalmente nuestro mundo.

Algoritmos no supervisados

- Para cada x, no hay Y.
- No sabemos las respuestas correctas, por lo que no podemos actuar como profesor.
- En su lugar, tratamos de comprender la distribución de x para obtener una inferencia sobre Y.

Fuente: Machine Learning Mastery - Supervised and Unsupervised Learning

Metodología de aprendizaje

¿Por qué es importante el aprendizaje no supervisado?

Yann Lecun, un investigador de aprendizaje profundo, hizo la analogía de que si la inteligencia fuera una torta, el aprendizaje no supervisada sería la torta y el aprendizaje supervisado sería la guinda del pastel.

Sabemos cómo hacer el granizado, pero no sabemos cómo hacer el pastel. **El aprendizaje no supervisado es el santo grial del aprendizaje automático.**

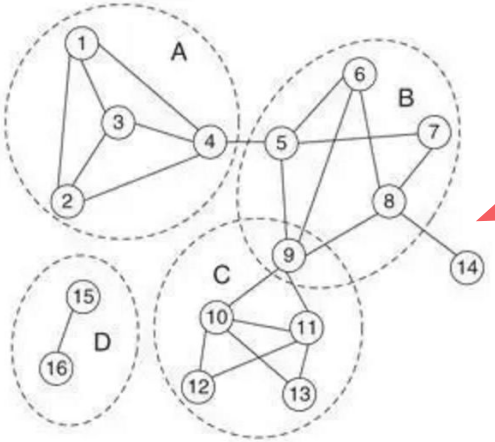
Para alcanzar la verdadera inteligencia de la máquina, ML necesita mejorar en el aprendizaje no supervisado.

Los humanos aprenden principalmente a través de un aprendizaje no supervisado: absorbemos grandes cantidades de datos de nuestro entorno sin necesidad de una etiqueta.

Metodología de aprendizaje

Ejemplos de aprendizaje no supervisado: este algoritmo de agrupamiento (clustering) predice los amigos de un usuario en función de su actividad en las redes sociales.

Social Network Analysis: In a social network, clustering can be used to find users that interact a lot with each other (say, via e-mails). This is shown in the figure below where the users have been clustered into four clusters - A,B,C and D.



No tenemos ningún dato etiquetado que nos diga que cualquier nodo es amigo de otro nodo.

En su lugar, podemos utilizar las interacciones del usuario para proporcionar las etiquetas. La suposición es que si estás interactuando fuertemente con alguien, es más probable que sea tu amigo.

Fuente: Quora - What is unsupervised learning with an example.

Metodología de aprendizaje

El algoritmo de clstering utiliza patrones de correo electrónico para predecir la jerarquía de una organización empresarial.

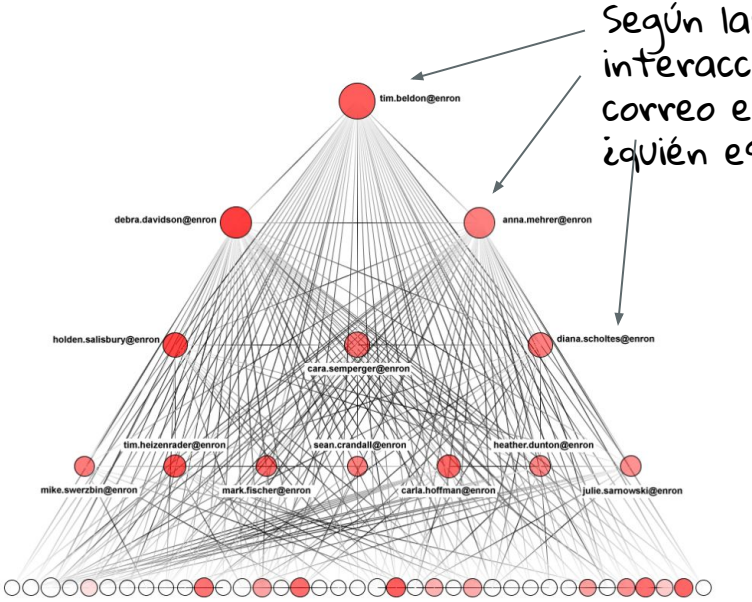


Figure 1: Enron North American West Power Traders Extracted Social Network

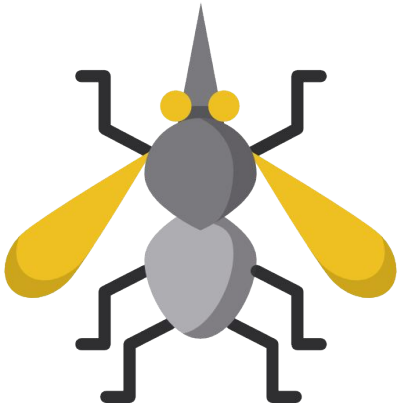
Este algoritmo no solo toma en cuenta a las personas involucradas en una interacción sino también la direccionalidad de la interacción.

Fuente: Automated Social Hierarchy Detection Through Email Network Analysis - Academic Paper

En la próxima clase, presentaremos la selección y evaluación de modelos.

Vamos a resumir rápidamente lo que hemos aprendido en este módulo.

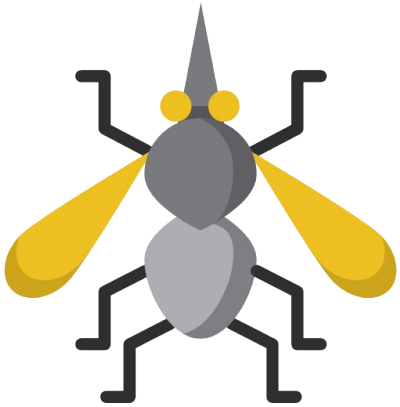
¿Cuál es la tarea y la medida de rendimiento para nuestro ejemplo de red de malaria?



Tarea

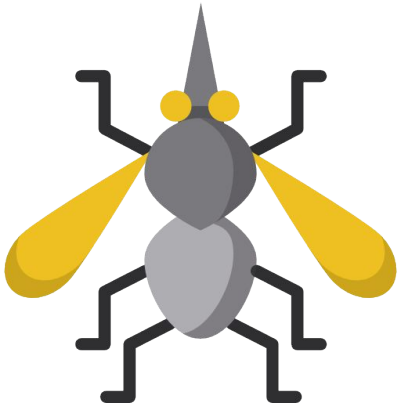
Medida de rendimiento

¿Cuál es la tarea y la medida de rendimiento para nuestro ejemplo de red de malaria?



Tarea	Predecir con precisión la disminución de casos de malaria para un aumento dado de mosquiteros
Medida de rendimiento	Error cuadrático medio

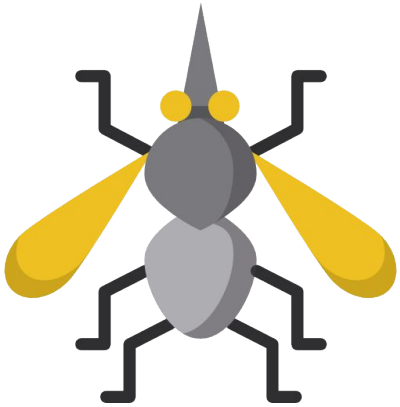
¿Cuál es la tarea y la medida de rendimiento para nuestro ejemplo de paciente con malaria?



Tarea

Medida de rendimiento

¿Cuál es la tarea y la medida de rendimiento para nuestro ejemplo de paciente con malaria?



Tarea	Predice con precisión si es probable que un paciente dado tenga malaria, dada su temperatura.
Medida de rendimiento	Log Loss

Obtén más información sobre el machine learning de Delta para una buena misión aquí.

Recursos adicionales

Recursos adicionales

Rosasco, Lorenzo, et al. "Are loss functions all the same?." Neural Computation 16.5 (2004): 1063-1076. http://web.mit.edu/lrosasco/www/publications/loss.pdf

“Loss Functions for Regression and Classification”, David Rosenberg, NYU: https://davidrosenberg.github.io/ml2015/docs/3a.loss-functions.pdf