

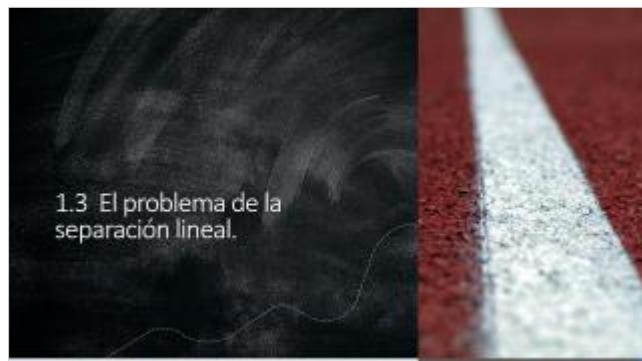
Redes Neuronales (y Bayesianas)

LDS1081

Juan Manuel Ahuactzin Larios
juan.ahuactzin@udlap.mx

Todas las imágenes de este curso fueron obtenidas de Pexels, Pxhere y Microsoft Powerpoint:
<https://www.pexels.com/> <https://pxhere.com/>

Contenido: Da clic en la sección que quieras consultar.





Antes de
empezar



Protección Civil

Video

Lineamientos para la evaluación

Tipo de Evaluación	Porcentaje
Asistencia y participación	5%
Ejercicios	10%
Exámenes parciales (2)	30%
Examen final	20%
Proyecto 1	15%
Proyecto 2 (final)	20%
Total	100%

Sobre el plagio

Cualquier tipo de fraude académico o plagio en tareas/trabajos, considerados como “**falta grave**”, será sancionado con una calificación de **CERO**.

asistencia

- Es importante la asistencia del estudiante a clases.
- Una calificación reprobatoria en asistencias impide al estudiante aprobar el curso.

Puntualidad

Es responsabilidad del estudiante llegar puntualmente a clases.

Utilización racional de tecnología

No está permitido el uso de gadgets (teléfonos celulares, tabletas, computadoras, etc.) en las sesiones de clases para cuestiones ajenas a la clase.

Entregas oportunas.

Las tareas y trabajos serán entregados en tiempo y forma, por lo que no se recibirán después de la fecha y hora establecida. No se aplicarán trabajos o tareas de reposición debido a inasistencias.

Aplicación de exámenes.

- Los exámenes deberán presentarse en la **fecha y hora establecida** por el profesor.
- En caso de existir alguna causa de fuerza mayor, ésta será revisada por el profesor y deberá quedar plenamente justificada **por escrito**.

Honestidad en exámenes.

- Los exámenes serán en Blackboard, por ninguna razón se podrá navegar en otras páginas o abrir aplicaciones o archivos durante el examen, de lo contrario éste será calificado con CERO.

Honestidad en tareas y trabajos.

- Cualquier tipo de fraude académico o plagio en tareas/trabajo y en exámenes, considerados como “**falta grave**” será sancionado con una calificación de **CERO**.
- Si se considera necesario se llevará a la Comisión Disciplinaria.
- No hay reposiciones de las tareas o trabajos calificados con CERO por PLAGIO.

Reporte de comportamientos no éticos.

Cualquier comportamiento deshonesto que el profesor(a) identifique en un examen o trabajo de investigación o actividad asignada significará 0 (cero).

Calificación final.

La revisión final sólo se enfocará **al examen y/o trabajo final** y se realizará en el día y horario indicado por el profesor. Es responsabilidad del estudiante monitorear su desempeño a lo largo del periodo académico.

Respeto para la clase.

- Generar un ambiente de participación, reflexión y respeto que facilite la creatividad y el aprendizaje.
- Utilizar un lenguaje apropiado y de respeto en clases.

Interacción alumno-profesor.

La interacción entre el alumno-profesor, a través de cualquier medio debe realizarse de manera cordial y respetuosa.

Fuentes de consulta.

Sólo se aceptarán **fuentes confiables de referencia**. Ejemplos de fuentes de baja o dudosa calidad (monografías.com, elrincondelvago.com, unamosapuntes.com, buenastareas.com, ilustrados.com, gestiopolis.com, geocities.com, elprisma.com, clubensayos.com, entre otras).

Comunicación.

Es responsabilidad del estudiante estar pendiente de los comunicados generados por el profesor relativos al curso, de acuerdo a los medios utilizados (e-mail institucional, Blackboard y/o Portafolios).



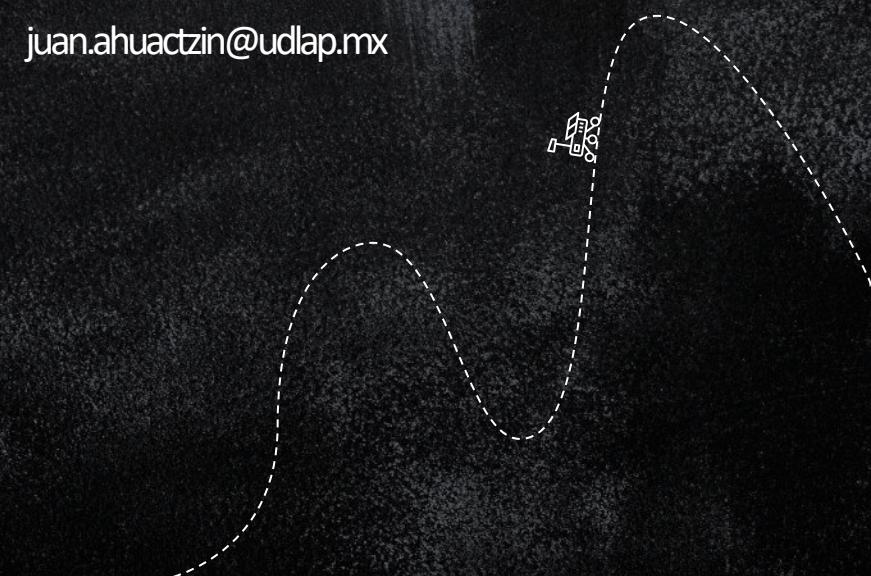
Equipo



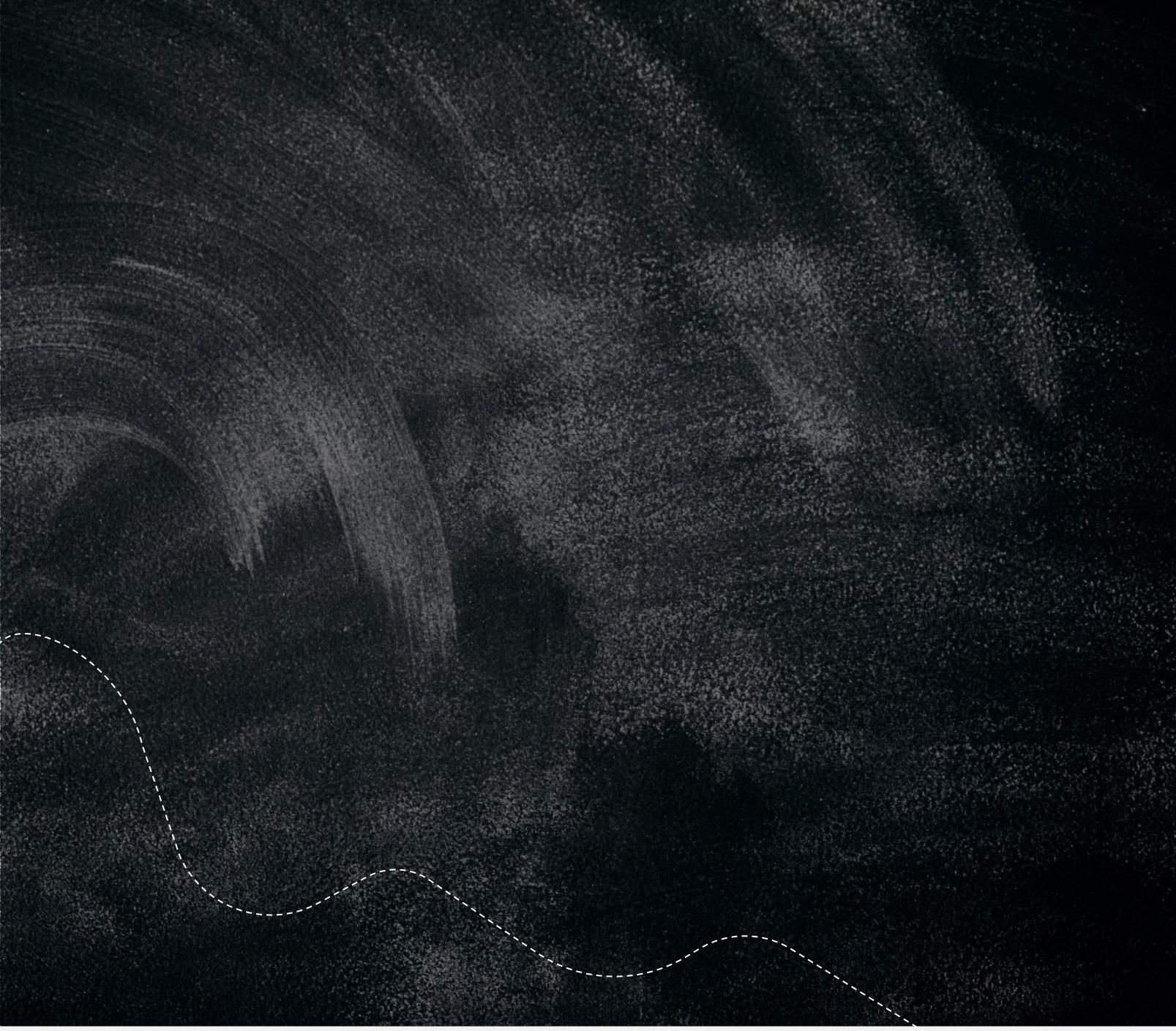
1. Introducción a las redes neuronales.

Juan Manuel Ahuactzin Larios

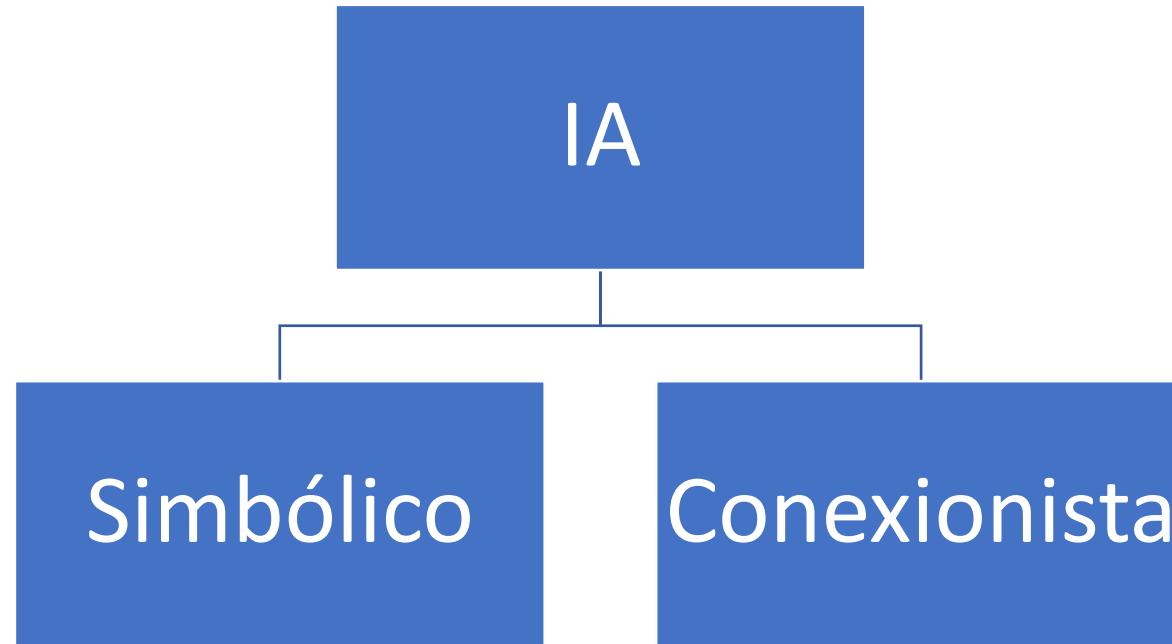
juan.ahuactzin@udlap.mx



1.1 Modelos de procesamiento paralelo distribuido.



Inteligencia Artificial: Simbólico o conexiónista



Inteligencia Artificial: Simbólico o conexiónista

Simbólicos:

- uso de “lenguajes formales” , lógicas proposicionales, sistemas lógicos de primer orden, lenguajes como Prolog.
- **GOFAI: Good Old-Fashioned IA**
- **Sistemas Expertos**

Inteligencia Artificial: Simbólico o conexionista

Simbólicos

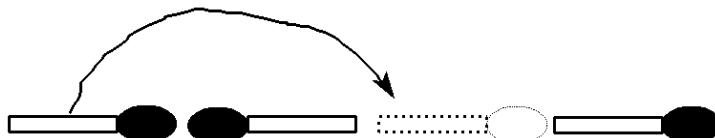
Seis cerillos se alinean de la forma siguiente:



a) en el sentido que apuntan sus cabezas siempre y cuando al frente exista un espacio vacío :



b) saltando a un y solo un cerillo que venga en dirección contraria :



El problema consiste en pasar los cerillos de la izquierda a la derecha y los de la derecha a la izquierda. Es decir a la siguiente posición:



Inteligencia Artificial: Simbólico o conexionista

Simbólicos



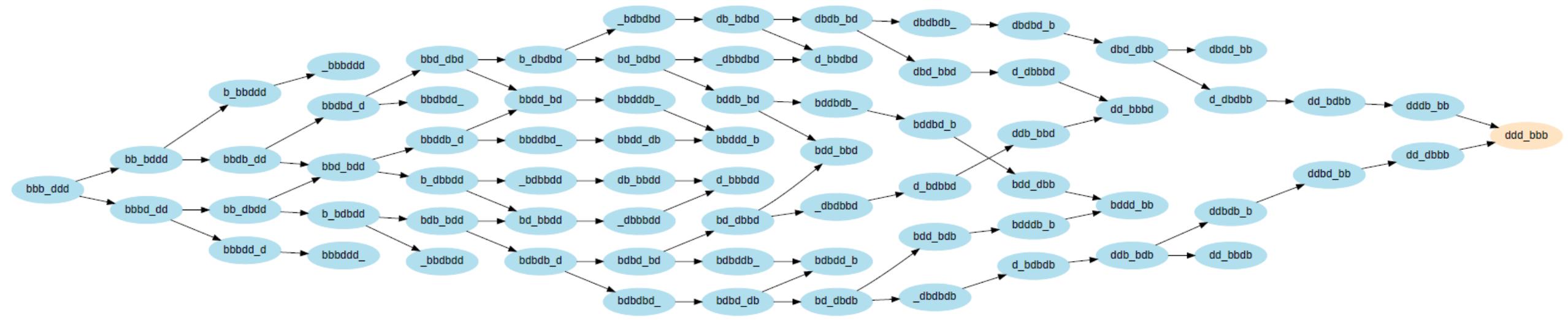
bbb _ ddd



ddd _ bbb

Inteligencia Artificial: Simbólico o conexionista

Simbólicos



Inteligencia Artificial: Simbólico o conexionista

Simbólicos

```
regla(d*e*A*B*C*D*E,e*d*A*B*C*D*E).  
regla(A*d*e*B*C*D*E,A*e*d*B*C*D*E).  
regla(A*B*d*e*C*D*E,A*B*e*d*C*D*E).  
regla(A*B*C*d*e*D*E,A*B*C*e*d*D*E).  
regla(A*B*C*D*d*e*E,A*B*C*D*e*d*E).  
regla(A*B*C*D*E*d*e,A*B*C*D*E*e*d).
```

```
1  regla(d*e*A*B*C*D*E,e*d*A*B*C*D*E).
2  regla(A*d*e*B*C*D*E,A*e*d*B*C*D*E).
3  regla(A*B*d*e*C*D*E,A*B*e*d*C*D*E).
4  regla(A*B*C*d*e*D*E,A*B*C*e*d*D*E).
5  regla(A*B*C*D*d*e*E,A*B*C*D*e*d*E).
6  regla(A*B*C*D*E*d*,A*B*C*D*E*e*d).
7
8  regla(e*d*i*A*B*C*D,i*d*e*A*B*C*D).
9  regla(A*e*d*i*B*C*D,A*i*d*e*B*C*D).
10 regla(A*B*e*d*i*C*D,A*B*i*d*e*C*D).
11 regla(A*B*C*e*d*i*D,A*B*C*i*d*e*d).
12 regla(A*B*C*D*e*d*i,A*B*C*D*i*d*e).
13
14 regla(A*B*C*D*E*e*i,A*B*C*D*E*i*e).
15 regla(A*B*C*D*e*i*E,A*B*C*D*i*e*E).
16 regla(A*B*C*e*i*D*E,A*B*C*i*e*D*E).
17 regla(A*B*e*i*C*D*E,A*B*i*e*C*D*E).
18 regla(A*e*i*B*C*D*E,A*i*e*B*C*D*E).
19 regla(e*i*A*B*C*D*E,i*e*A*B*C*D*E).
20
21 regla(d*i*e*A*B*C*D,e*i*d*A*B*C*D).
22 regla(A*d*i*e*B*C*D,A*e*i*d*B*C*D).
23 regla(A*B*d*i*e*C*D,A*B*e*i*d*C*D).
24 regla(A*B*C*d*i*e*D,A*B*C*e*i*d*D).
25 regla(A*B*C*D*d*i*e,A*B*C*D*e*i*d).
26
27 camino(X,0,X,L,L).
28 camino(X,N,Y,L,LB):-N>0, N1 is N-1,regla(X,XP),camino(XP,N1,Y,[XP|L],LB).
29
30 elcamino(X,N,Y,L,LB):-N>=0, camino(X,N,Y,L,LB); NP is N+1, elcamino(X,np,Y,L,LB).
31
32 encuentra_camino(X,Y,LB):-elcamino(X,0,Y,[X],LB).
33
34 escribe_camino([]).
35 escribe_camino([H|T]):- escribe_camino(T), writeln(H).
```

Inteligencia Artificial: Simbólico o conexionista

Simbólicos

swipl cerillos.pl

Inteligencia Artificial: Simbólico o conexionista

Simbólicos

```
dot -Tpdf sol.dot > sol.pdf
```

Inteligencia Artificial: Simbólico o conexionista

Simbólicos

El enfoque simbólico transforma la realidad por medio de un lenguaje simbólico, establece y modela las relaciones entre los símbolos. Por medio de un sistema basado en la lógica proposicional trata de emular el razonamiento humano.

Conexionista

Inteligencia Artificial: Simbólico o conexionista

Conexionista

Conexionistas:

- Espera que las relaciones surjan de forma automática.
- Cooperación de elementos simples que forman parte de una red.
- En esta categoría encontramos las Redes Neuronales y las Redes Bayesianas.

Inteligencia Artificial: Simbólico o conexionista

Conexionista

“En general los seres humanos son más inteligentes que las computadoras de hoy en día por que el cerebro utiliza una arquitectura que está más adaptada para tratar con un aspecto central del procesamiento de la información natural para la cual las personas son buenas.”

Inteligencia Artificial: Simbólico o conexionista

Conexionista

Las tareas requieren de consideraciones de muchas piezas de información y restricciones simultáneas.

Inteligencia Artificial: Simbólico o conexionista

Conexionista

El aprendizaje puede ocurrir espontáneamente, como subproducto de la actividad de procesamiento .

Inteligencia Artificial: Simbólico o conexionista

Conexionista

Vi el gran cañón volando hacia Nueva York.
Vi a las ovejas pastando en el campo.

Inteligencia Artificial: Simbólico o conexionista

Conexionista

Las influencias parecen ir en ambos sentidos, desde la sintaxis a la semántica y de la semántica a la sintaxis.

Inteligencia Artificial: Simbólico o conexionista

Conexionista

I like the joke.
I like the drive.
I like to joke.
I like to drive.

Inteligencia Artificial: Simbólico o conexionista

Conexionista

Varios teóricos sugirieron que almacenamos este conocimiento en términos de estructuras denominadas de diversas maneras:

- Guiones (Schank , 1976) ,
- marcos (Minsky , 1975) , o
- esquemas (Norman & Bobrow , 1976; Rumelhart , 1975).

Inteligencia Artificial: Simbólico o conexionista

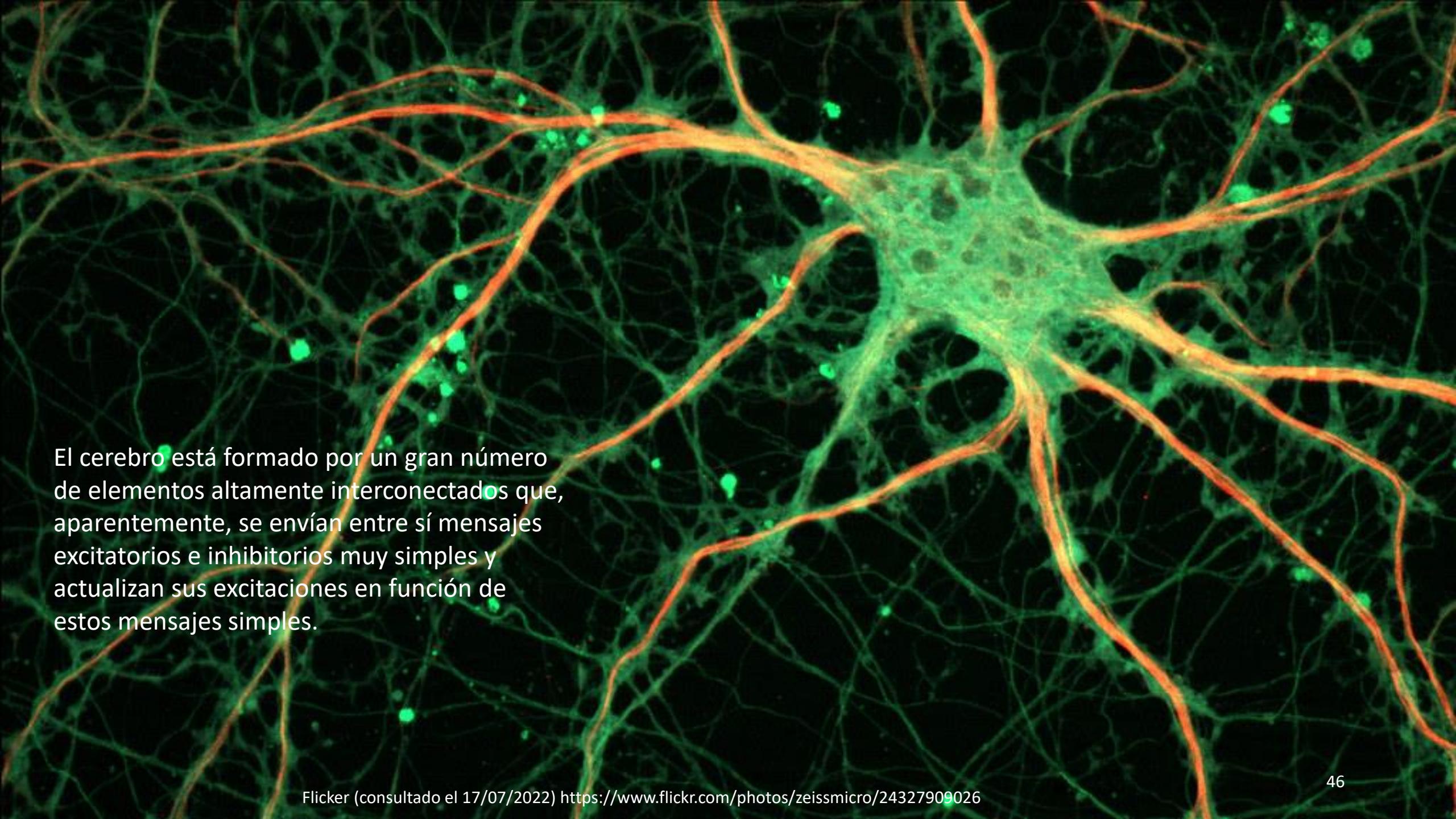
Conexionista

El cerebro es un ordenador altamente complejo,
no lineal y paralelo.

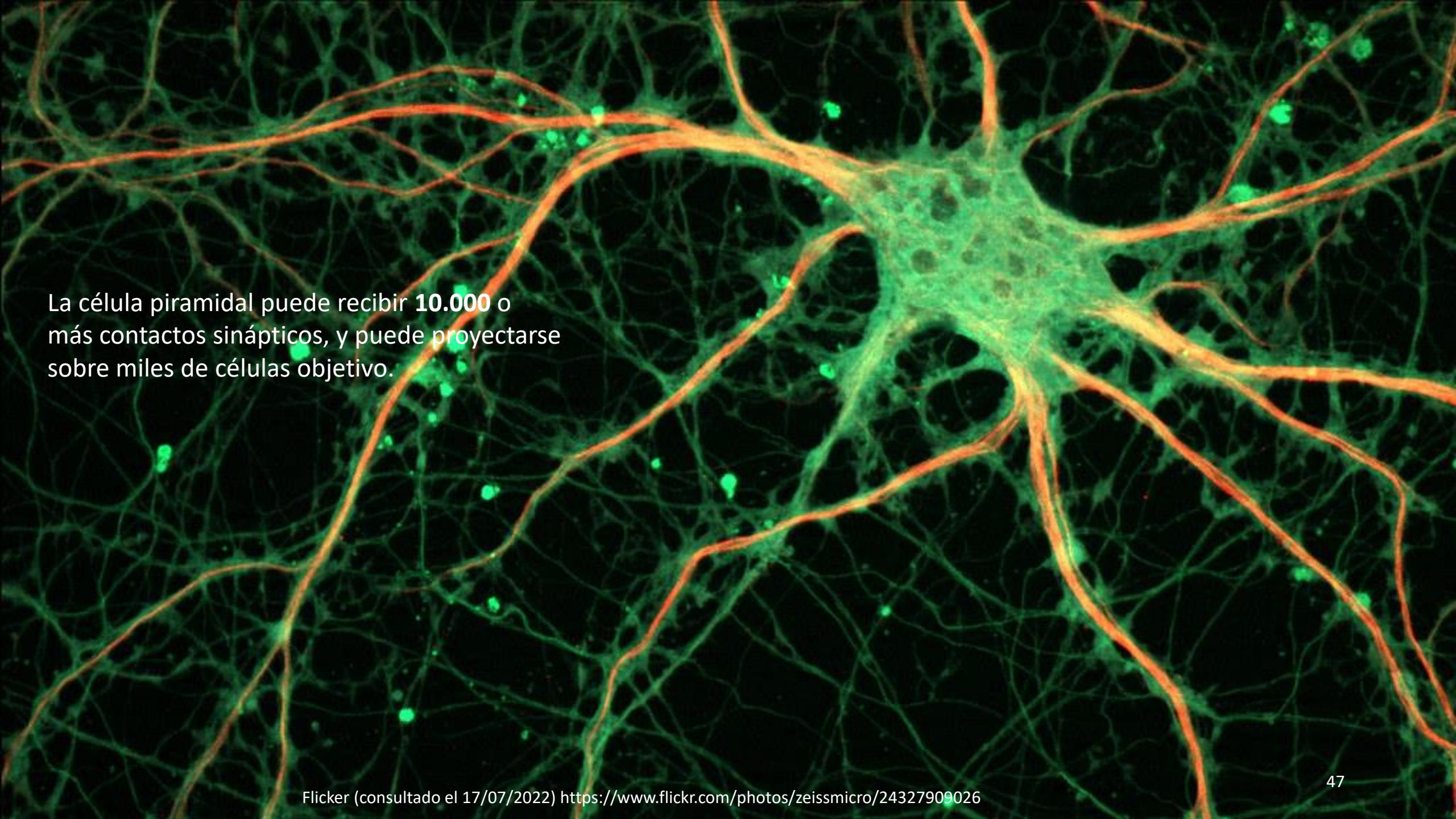
Inteligencia Artificial: Simbólico o conexionista

Conexionista

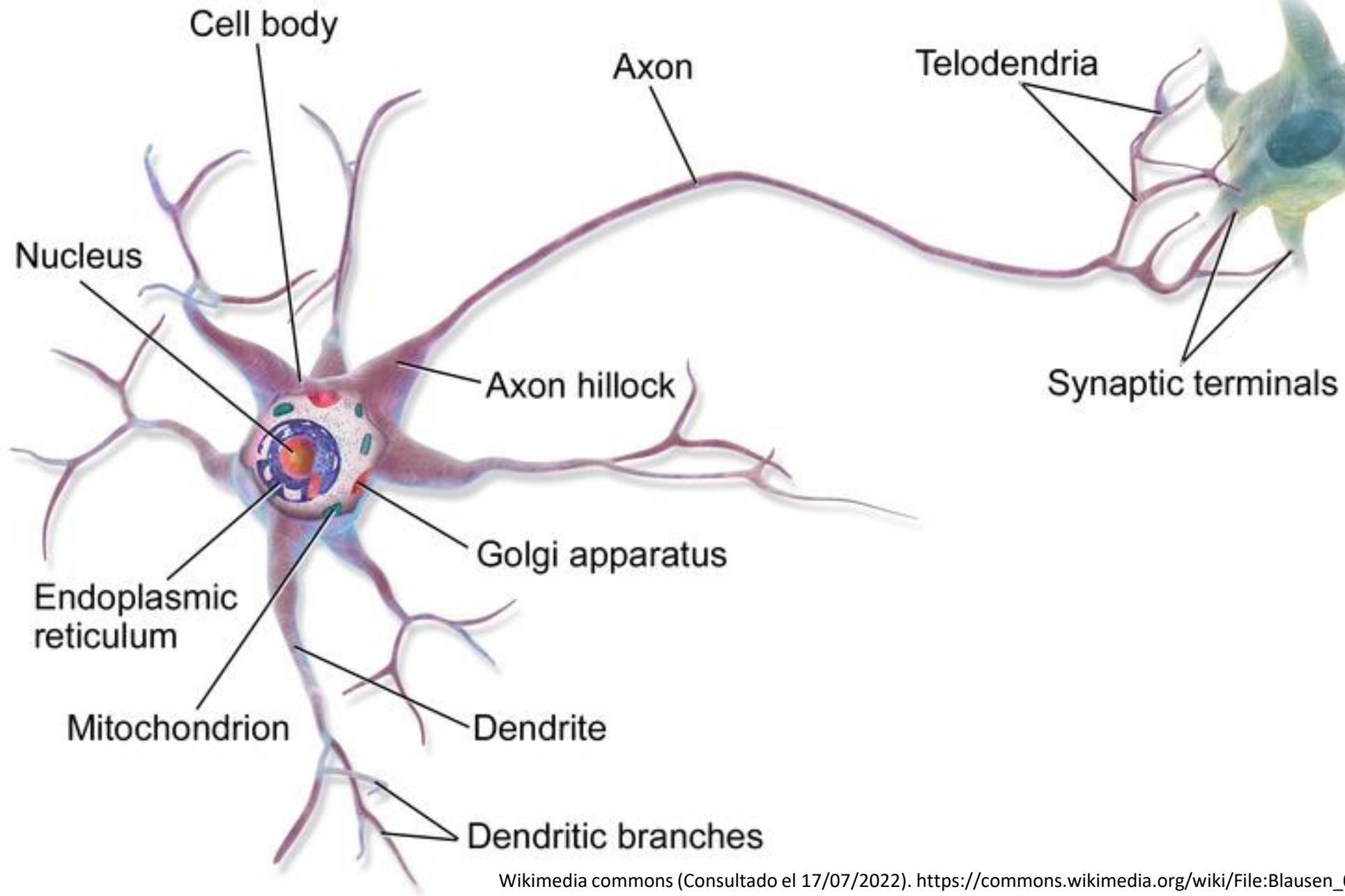
Modelos de Procesamiento Distribuido Paralelo (POP), asumen que el procesamiento de la información tiene lugar a través de interacciones de un gran número de elementos de procesamiento simples llamados **unidades**, cada una de las cuales envía **señales excitatorias e inhibidoras** a otras unidades.

A microscopic image showing a dense network of neurons. The cell bodies, or soma, are stained green and appear as small, bright spots against a dark background. The long, thin processes of the neurons, called axons, are stained orange and form a complex web-like structure that connects the different cells.

El cerebro está formado por un gran número de elementos altamente interconectados que, aparentemente, se envían entre sí mensajes excitatorios e inhibitorios muy simples y actualizan sus excitaciones en función de estos mensajes simples.



La célula piramidal puede recibir **10.000** o más contactos sinápticos, y puede proyectarse sobre miles de células objetivo.



Inteligencia Artificial: Simbólico o conexionista

Conexionista

En un cerebro adulto, la plasticidad puede explicarse por dos mecanismos: la creación de nuevas conexiones sinápticas entre neuronas y la modificación de las sinapsis existentes.

Inteligencia Artificial: Simbólico o conexionista

Conexionista

Hay ocho aspectos principales de un modelo de procesamiento distribuido paralelo:

1. Un conjunto de unidades de procesamiento.
2. Un estado de activación.
3. Una función de salida para cada unidad.
4. Un patrón de conectividad entre unidades.
5. Una regla de propagación para propagar patrones de actividades a través de la red de conectividades.
6. Una regla de activación para combinar las entradas que inciden en una unidad con el estado actual de esa unidad para producir un nuevo nivel de activación de la unidad.
7. Una regla de aprendizaje por la que los patrones de conectividad se modifican por experiencia.
8. Un entorno en el que el sistema debe funcionar.

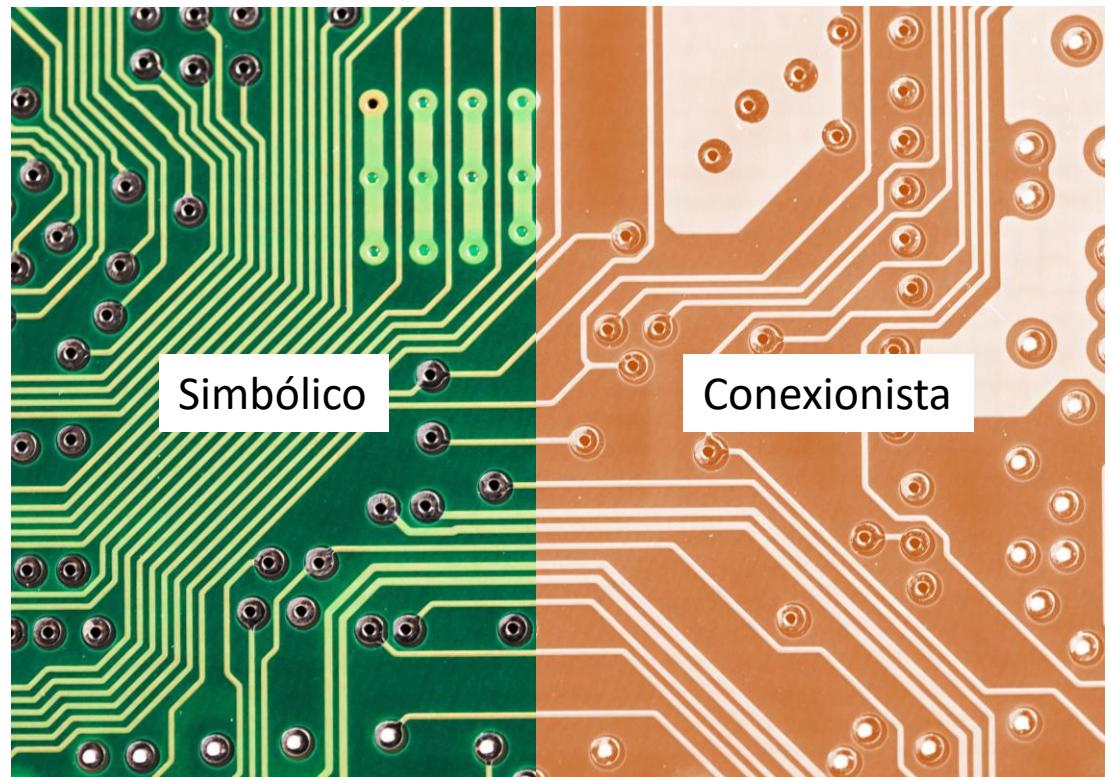
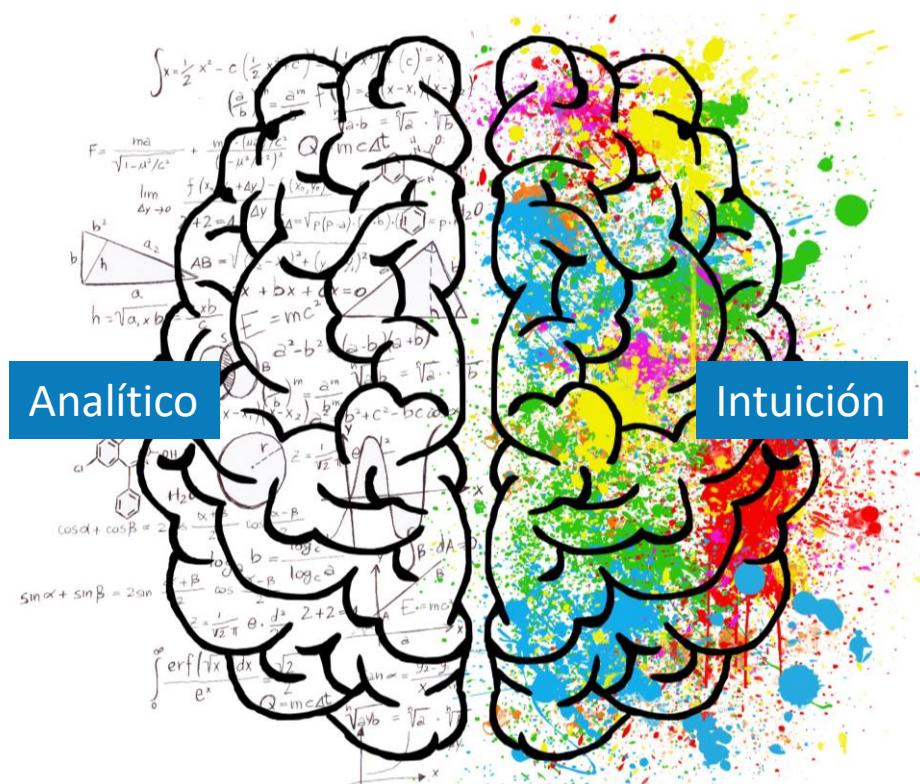
Modelos de procesamiento paralelo distribuido.

Una red neuronal artificial es un procesador distribuido, masivamente paralelo, formado por unidades de procesamiento sencillas que tiene una propensión natural a almacenar el conocimiento de la experiencia y ponerlo a disposición para su uso. Se asemeja al cerebro en dos aspectos:

1. El conocimiento lo adquiere la red de su entorno a través de un proceso de aprendizaje.
2. La fuerza de las conexiones entre las neuronas, conocida como peso sináptico, se utiliza para almacenar los conocimientos adquiridos.

Inteligencia Artificial: Simbólico o conexiónista

- Humanos
- Máquinas





Caracol: 0.047 km/h

1974: Procesador 4040
0.062 MIPS

75,097

566,467

Blackbird: 3,529.6 km/h

2020: Procesador AMD Ryzen 5 3600X
35,121 MIPS

7.54 Veces aún más rápido que el blackbird



Herramientas



Google colab

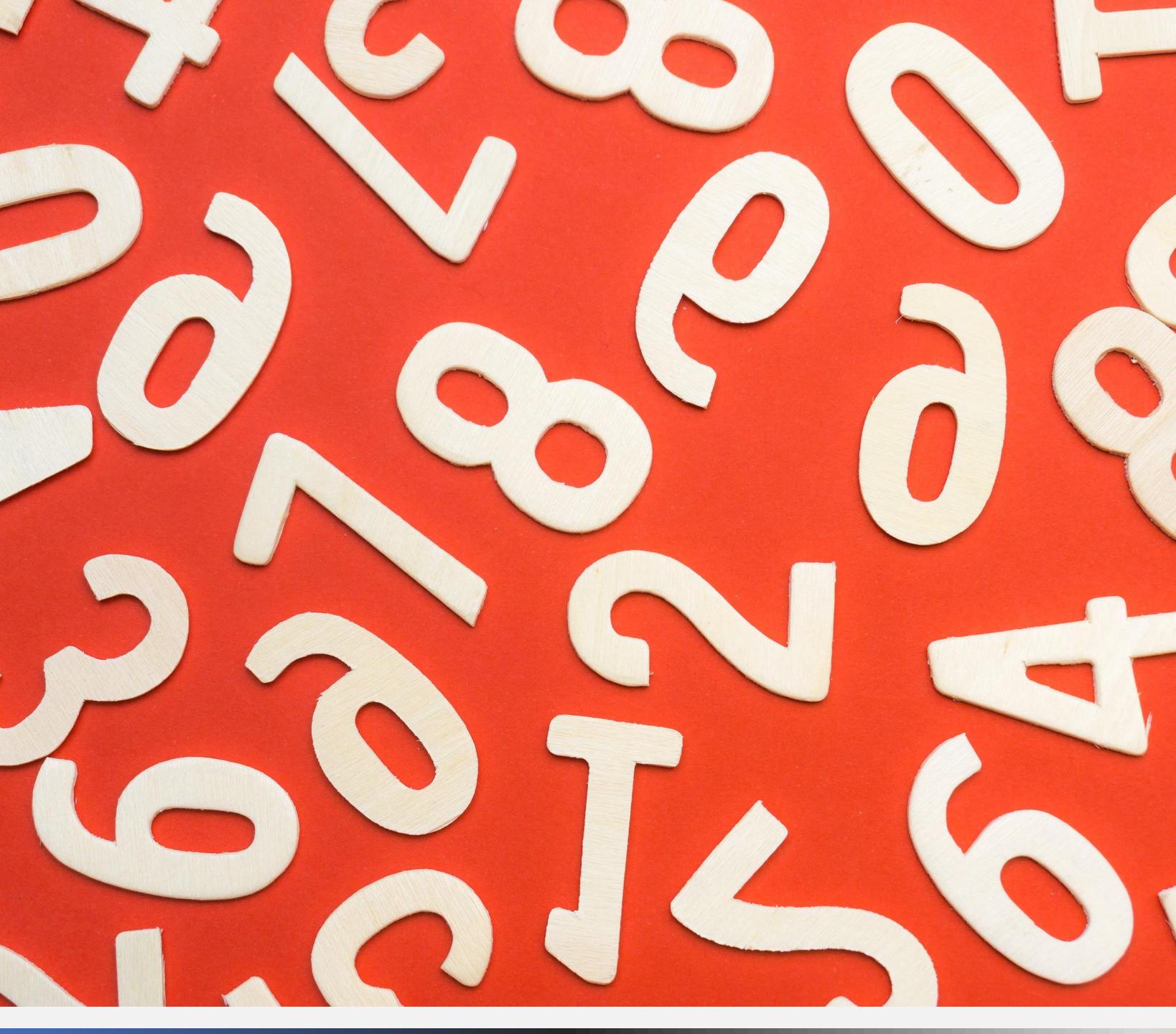


Anaconda / Jupyter Notebook

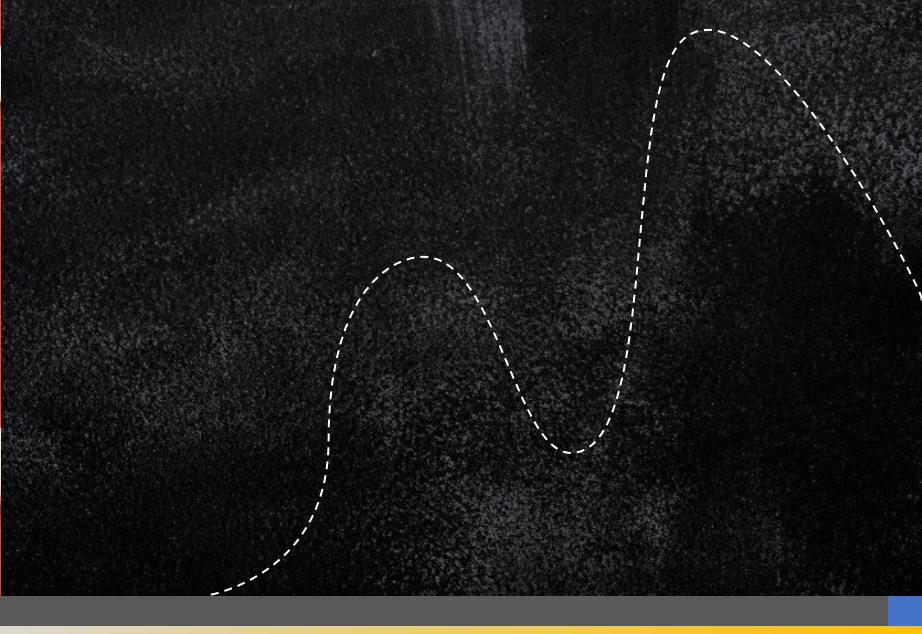


ProBT





1.2 Clasificación y reconocimiento de patrones.

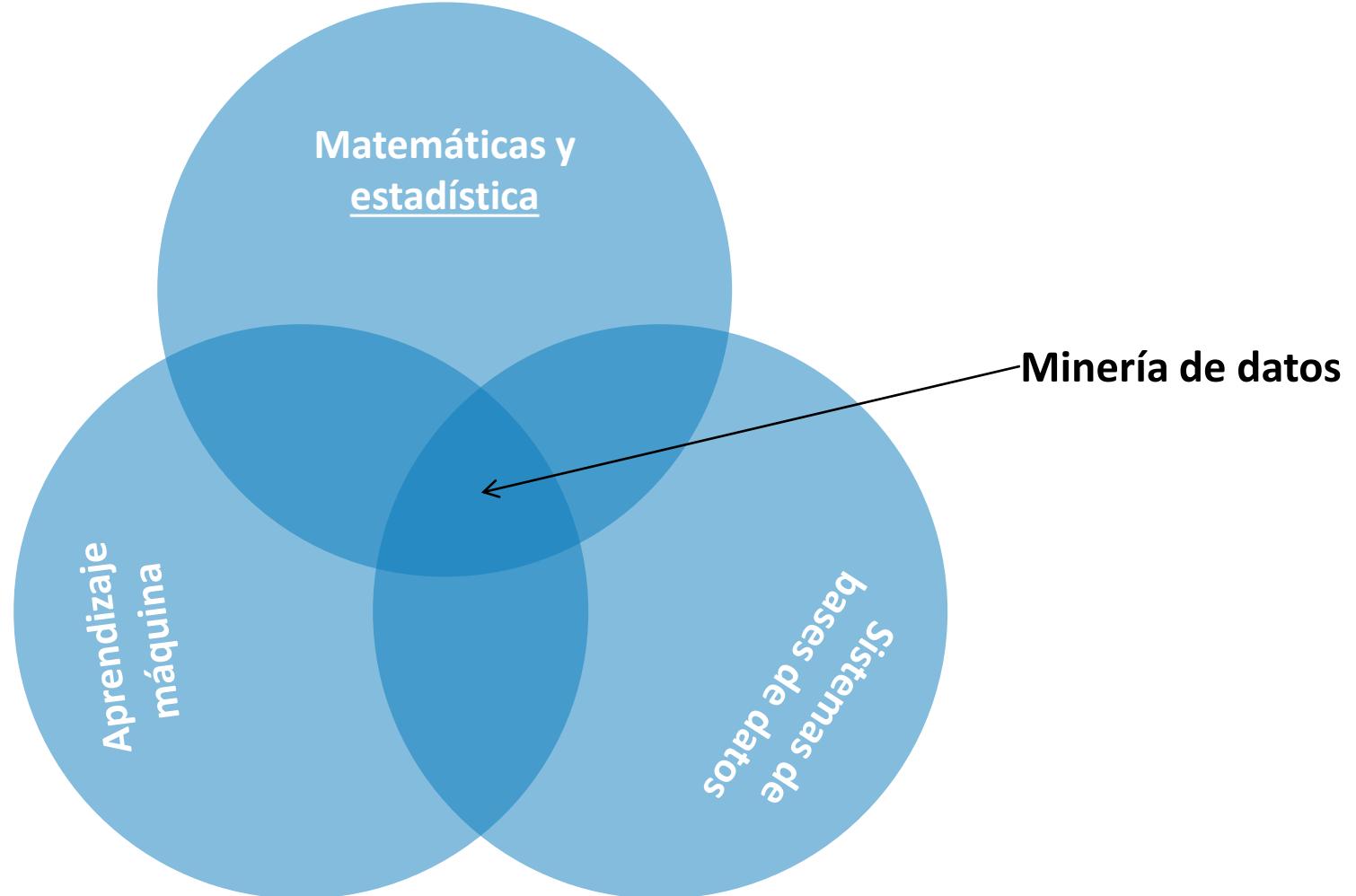


¿Qué es la minería de datos?

Minería de datos

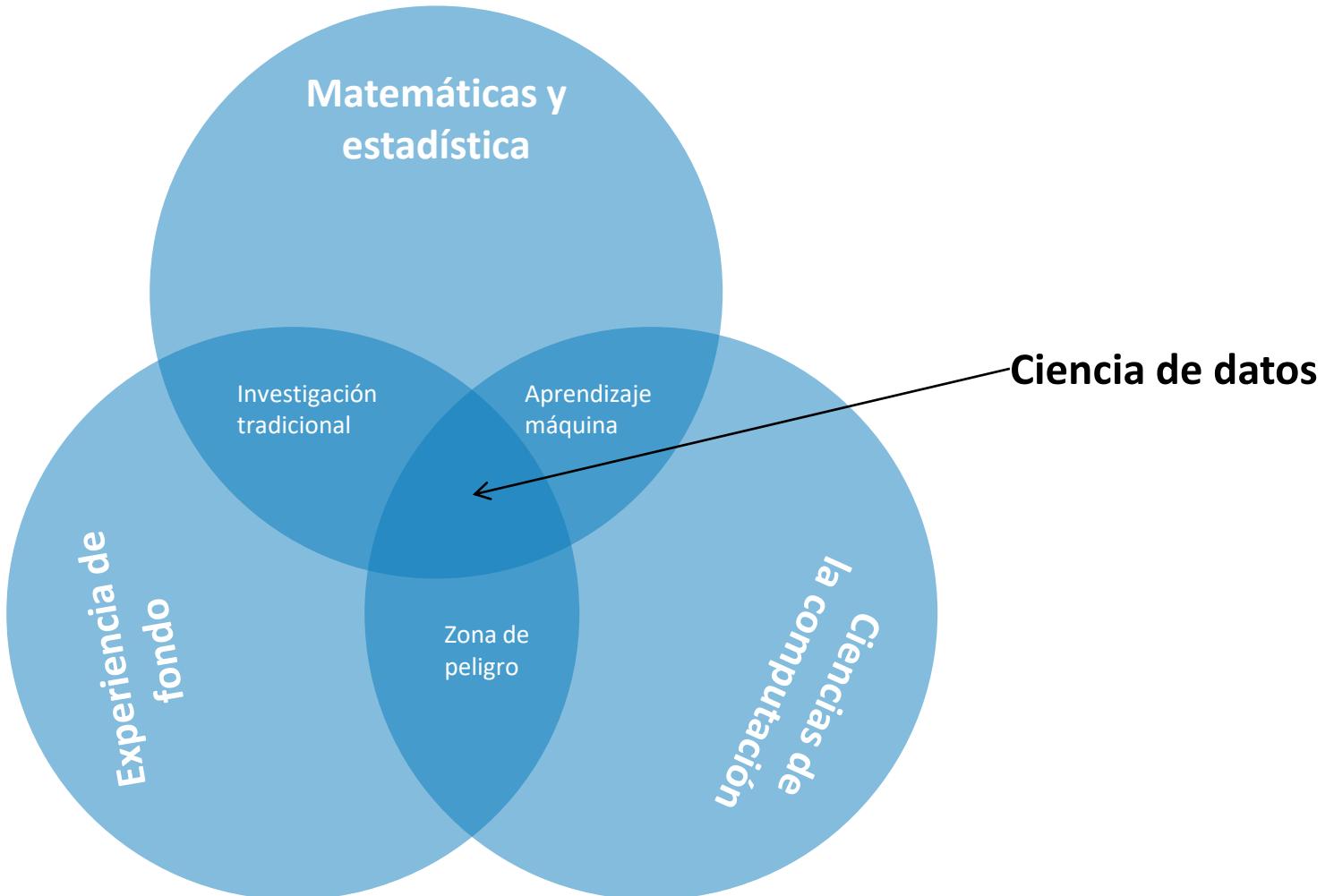
Consiste en encontrar tendencias en una colección de datos con el objetivo de encontrar patrones futuros. Es uno de los componentes esenciales para el descubrimiento de conocimientos. Puede o no incluir el análisis de Big-Data.

Minería de datos

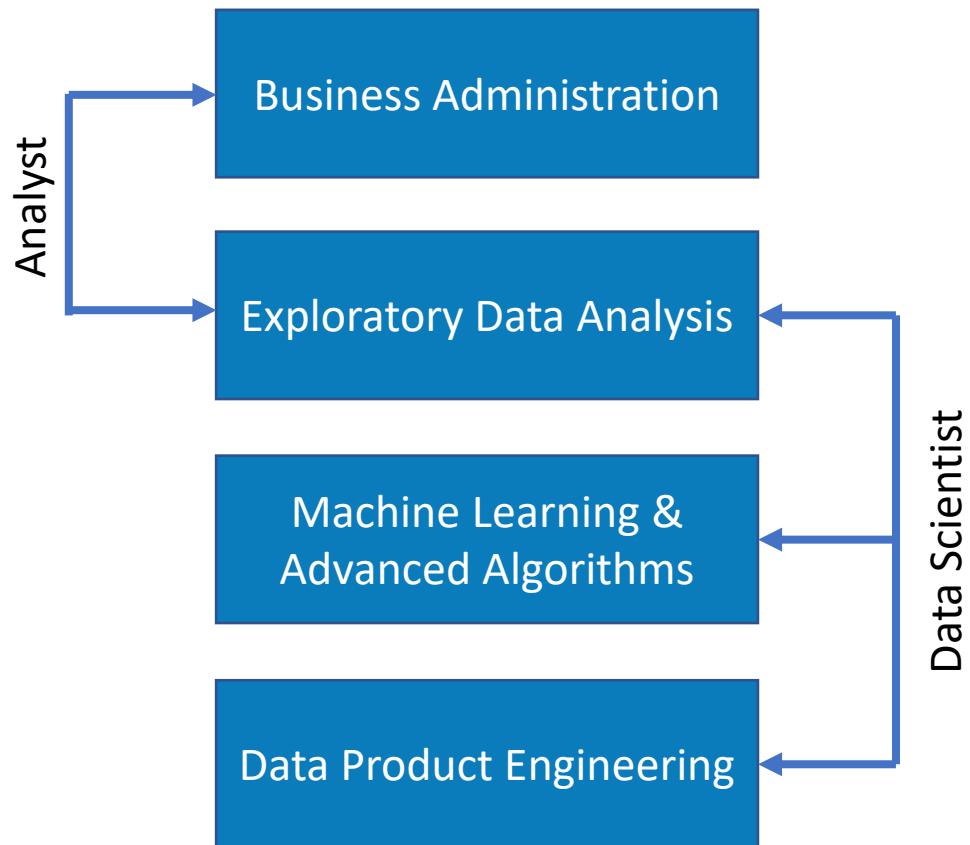


¿Qué es la ciencia de datos?

Ciencia de datos



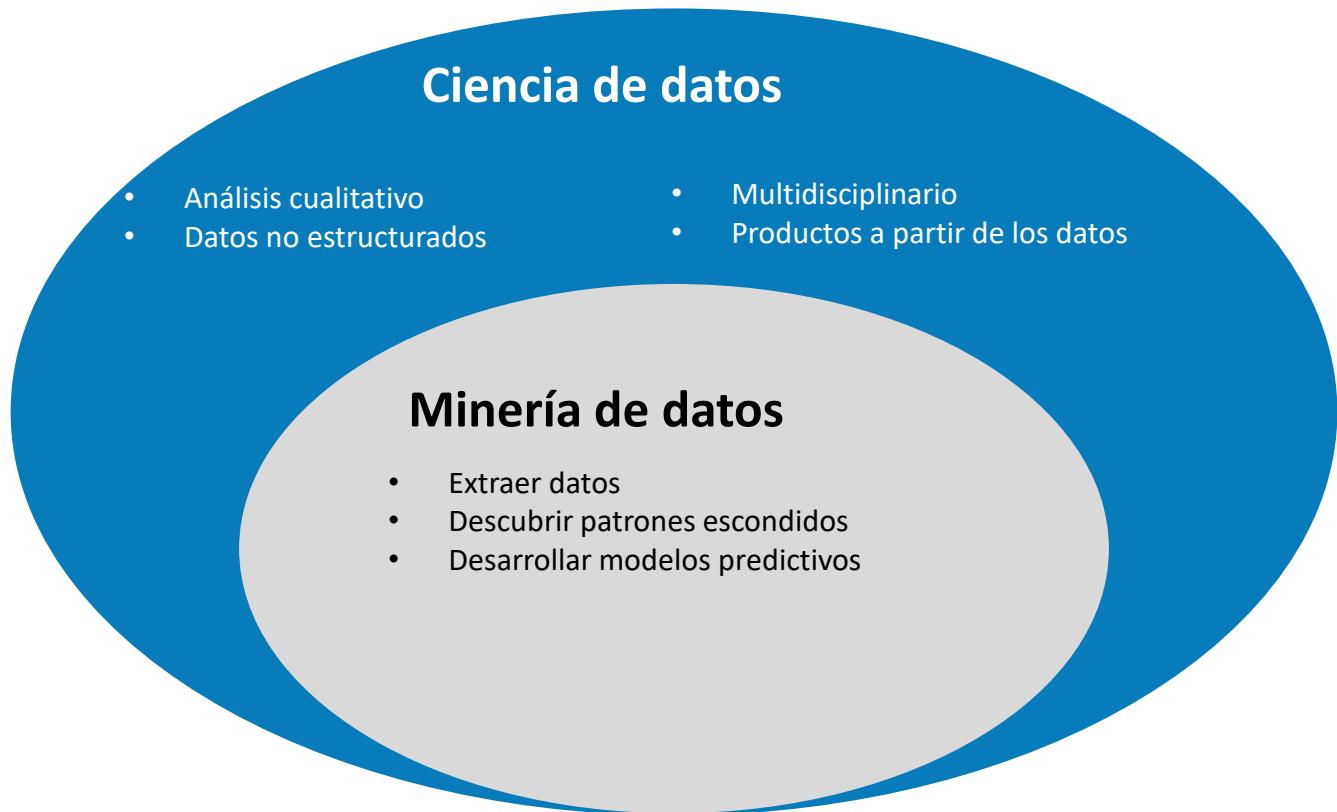
Ciencia de datos



“Es una combinación de varias herramientas, principios de aprendizaje automático y algoritmos cuyo propósito es descubrir patrones ocultos a partir de datos sin procesar”

“Un científico de datos analiza los datos desde diferentes perspectivas y ángulos”

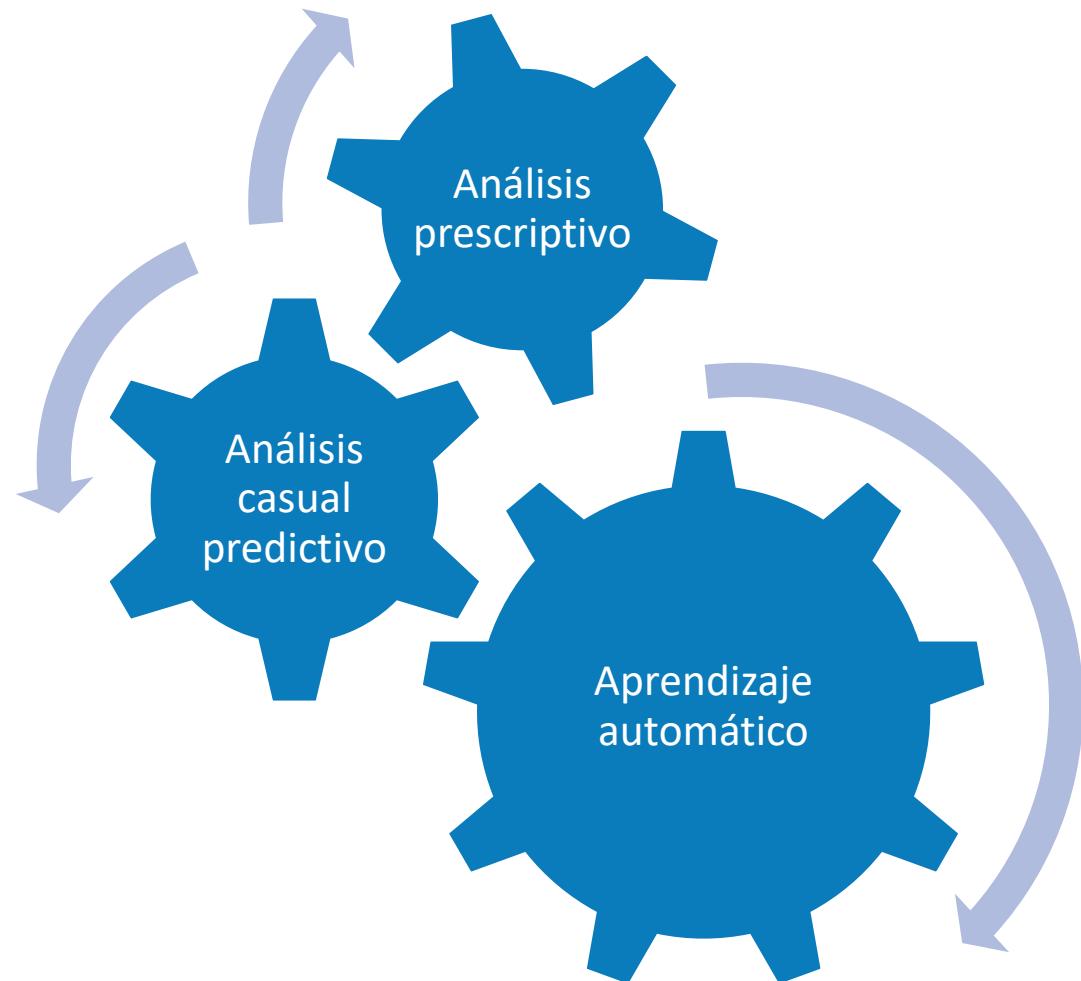
Minería de datos vs ciencia de datos



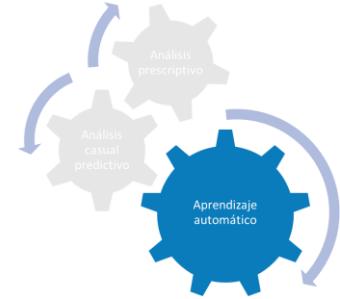
Minería de datos vs ciencia de datos

Comparación	Minería de datos	Ciencia de Datos
¿Qué es?	Conjunto de técnicas	Un área
Enfoque	Proceso de negocios	Estudio científico
Objetivo	Explorar los datos	Construir productos centrados en los datos para una organización.
Salida	Patrones	Diversas
Propósito	Encontrar tendencias desconocidas	Análisis social, construir modelos predictivos, desenterrar hechos desconocidos y más
Perspectiva vocacional	Manejo de datos y entendimiento estadístico.	Aprendizaje máquina, programación, técnicas infográficas, conocimiento de dominios específicos.
Alcance	Subconjunto de la ciencia de datos	Multidisciplinario, visualización, computación, estadística, minería de datos, procesamiento de lenguaje natural, imágenes, etc.
Tipo de datos	Estructurados	No estructurados, semiestructurados, estructurados.
Otros nombres	<i>Data archeology, Information Harvesting, Information Discovery, Knowledge Extraction.</i>	Ciencia basada en los datos (<i>Data-driven Science</i>)

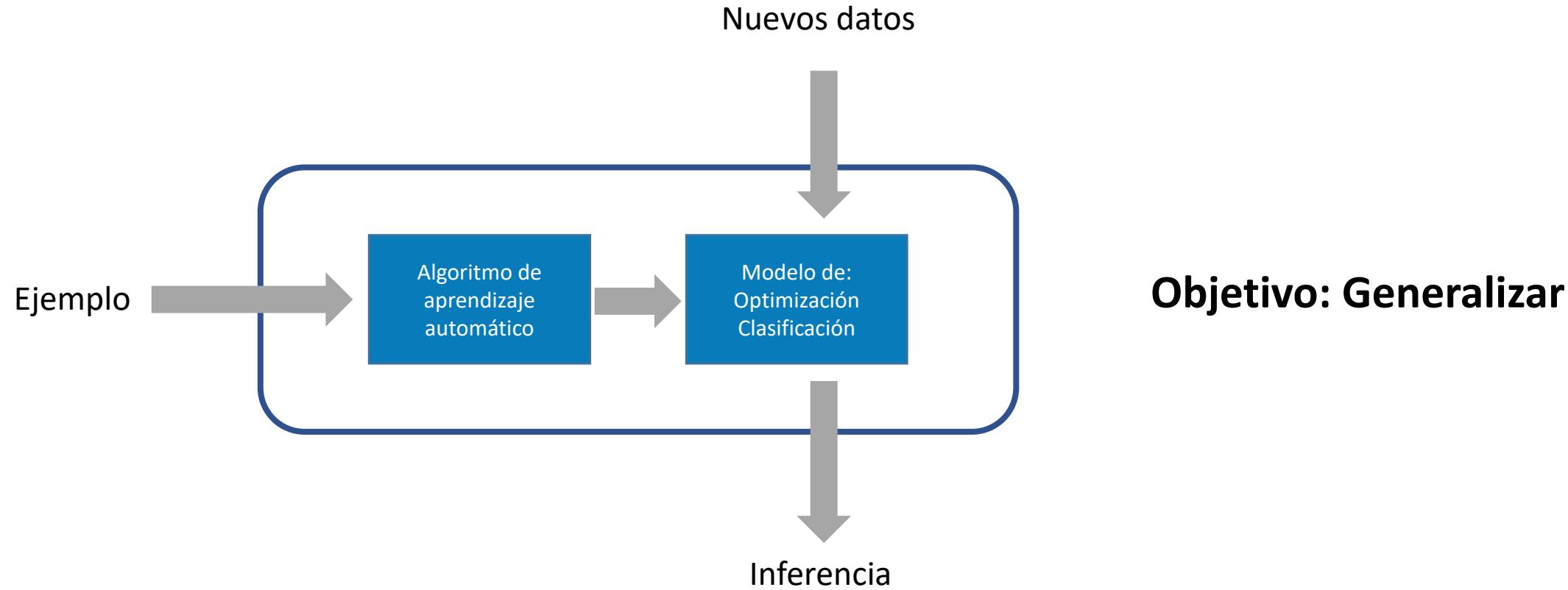
Predecir y tomar decisiones



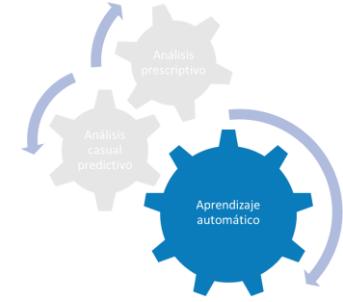
Aprendizaje automático o máquina vs reconocimiento de patrones



Búsqueda de algoritmos y heurísticas para convertir muestras de datos en programas, estos últimos sin que hayan sido programados por el humano.



Aprendizaje automático o máquina vs reconocimiento de patrones



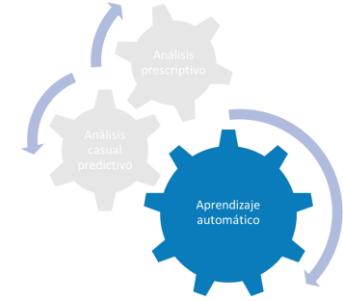
Reconocimiento de patrones:

- Tiene sus orígenes en ingeniería.
- Es un tipo de problema: es el reconocimiento de patrones y regularidades en los datos.

Aprendizaje Máquina

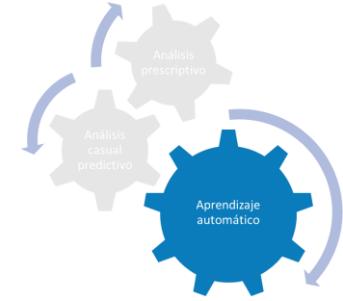
- Creció en las ciencias computacionales
- Es un tipo de solución: se ocupa de la construcción y el estudio de sistemas que pueden aprender de los datos, en lugar de seguir únicamente instrucciones programadas explícitamente.

Aprendizaje automático o máquina vs reconocimiento de patrones



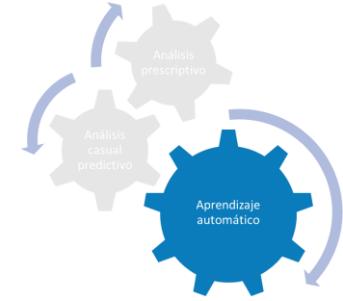
El campo del **reconocimiento de patrones** se ocupa del descubrimiento automático de regularidades en los datos mediante el uso de algoritmos informáticos y de la utilización de estas regularidades para tomar decisiones como la clasificación de los datos en diferentes categorías.

Aprendizaje automático o máquina vs reconocimiento de patrones



Aprendizaje máquina es el uso y desarrollo de sistemas informáticos capaces de aprender y adaptarse sin seguir instrucciones explícitas, mediante el uso de algoritmos y modelos estadísticos para analizar y sacar conclusiones de los patrones de los datos.

Aprendizaje automático o máquina vs reconocimiento de patrones

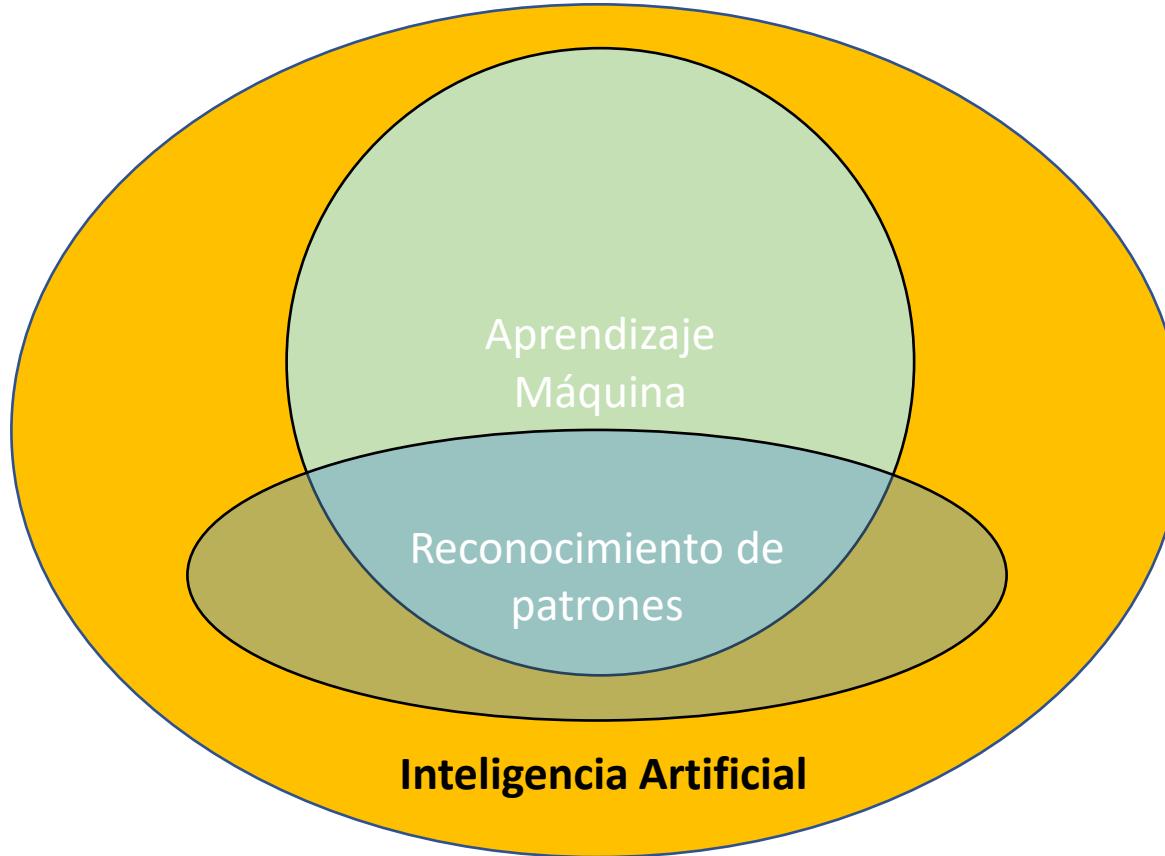
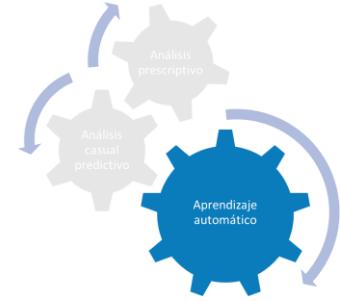


Aprendizaje automático	Reconocimiento de patrones
El aprendizaje automático es un método de análisis de datos que automatiza la construcción de modelos analíticos.	El reconocimiento de patrones es la aplicación de ingeniería de varios algoritmos con el fin de reconocer patrones en los datos.
El aprendizaje automático está más del lado práctico.	El reconocimiento de patrones es más teórico.
Puede ser una solución de un problema en tiempo real.	Puede ser un problema en tiempo real.
Necesitamos máquinas / ordenadores para aplicar los algoritmos de Aprendizaje Máquina.	El reconocimiento de patrones puede estar fuera de la máquina.

Machine Learning and Pattern Recognition. What's the difference between ML and pattern recognition? (Consultado el 18/07/2022).

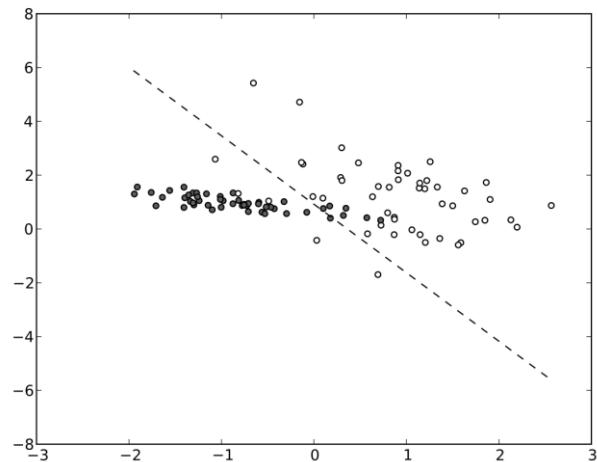
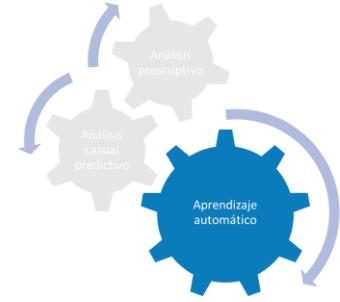
<https://dzone.com/articles/machine-learning-and-pattern-recognition#:~:text=Pattern%20Recognition%20is%20an%20engineering,patterns%20and%20regularities%20in%20data>.

Aprendizaje automático o máquina vs reconocimiento de patrones

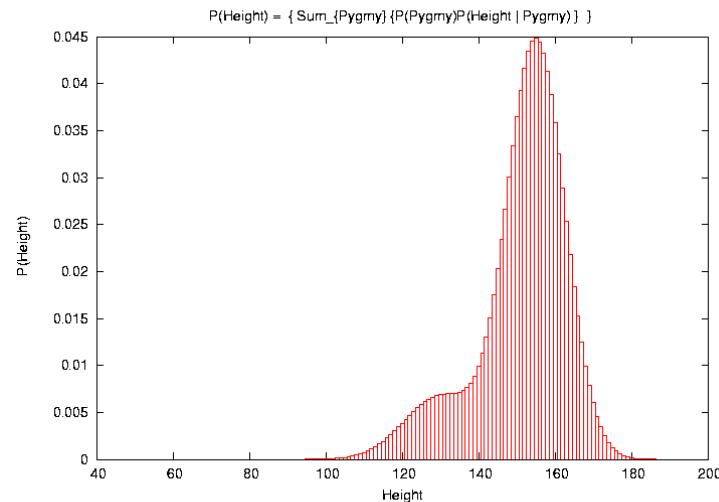


Aprendizaje automático o máquina vs reconocimiento de patrones

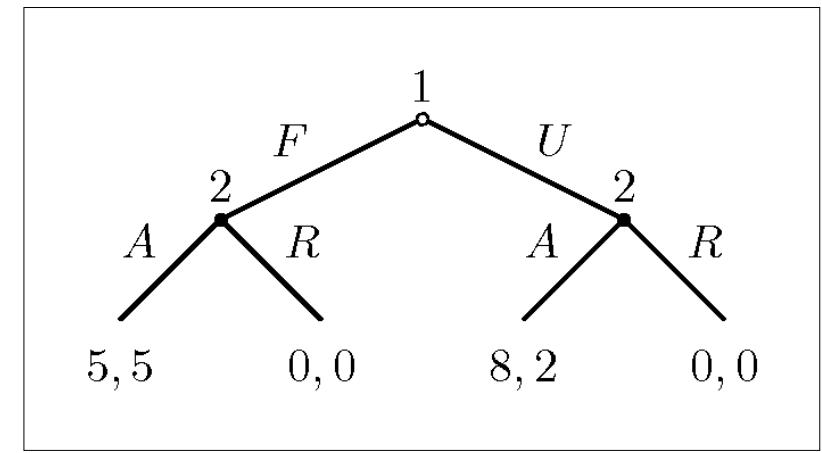
Una vista global



Geométricos

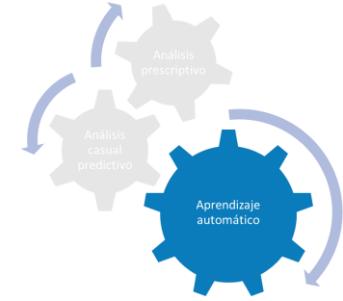


Probabilísticos



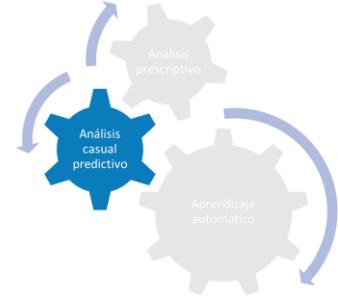
Lógicos

Aprendizaje automático o máquina vs reconocimiento de patrones



Modelo de acuerdo a la técnicas

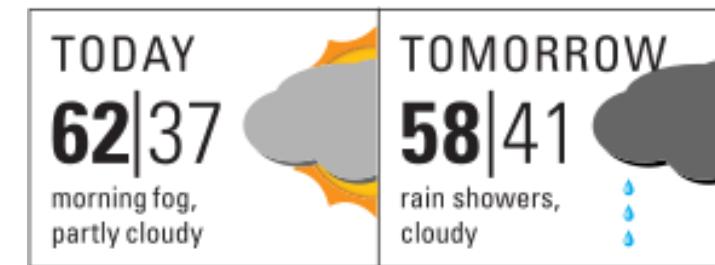
Modelo	Geométrico	Probabilístico	Lógico
Árboles de decisiones		X	X
Algoritmos genéticos	X	X	
Reglas de asociación		X	X
Redes neuronales artificiales	X	X	
Máquinas de vectores de soporte	X		
Algoritmos de agrupamiento (<i>Clustering</i>)	X	X	
Redes bayesianas		X	

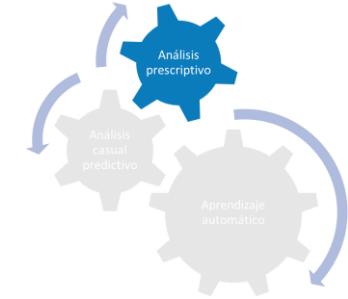


Analítica causal predictiva

Predecir las posibilidades de que ocurra un evento en el futuro. Basado en datos históricos. Realizar un análisis de lo que pasará en el futuro.

- ¿Qué probabilidad hay de que le guste la película?
- ¿Cuánto venderemos la próxima semana?
- ¿Qué posibilidades hay de que llueva hoy por la tarde?





Analítica prescriptiva

“Un modelo que tiene la inteligencia y la capacidad para tomar sus propias decisiones. ...no sólo predice sino que también recomienda diferentes acciones prescritas y resultados relacionados”.



Habilidades de un científico de datos

$\Omega\omega$

Matemáticas



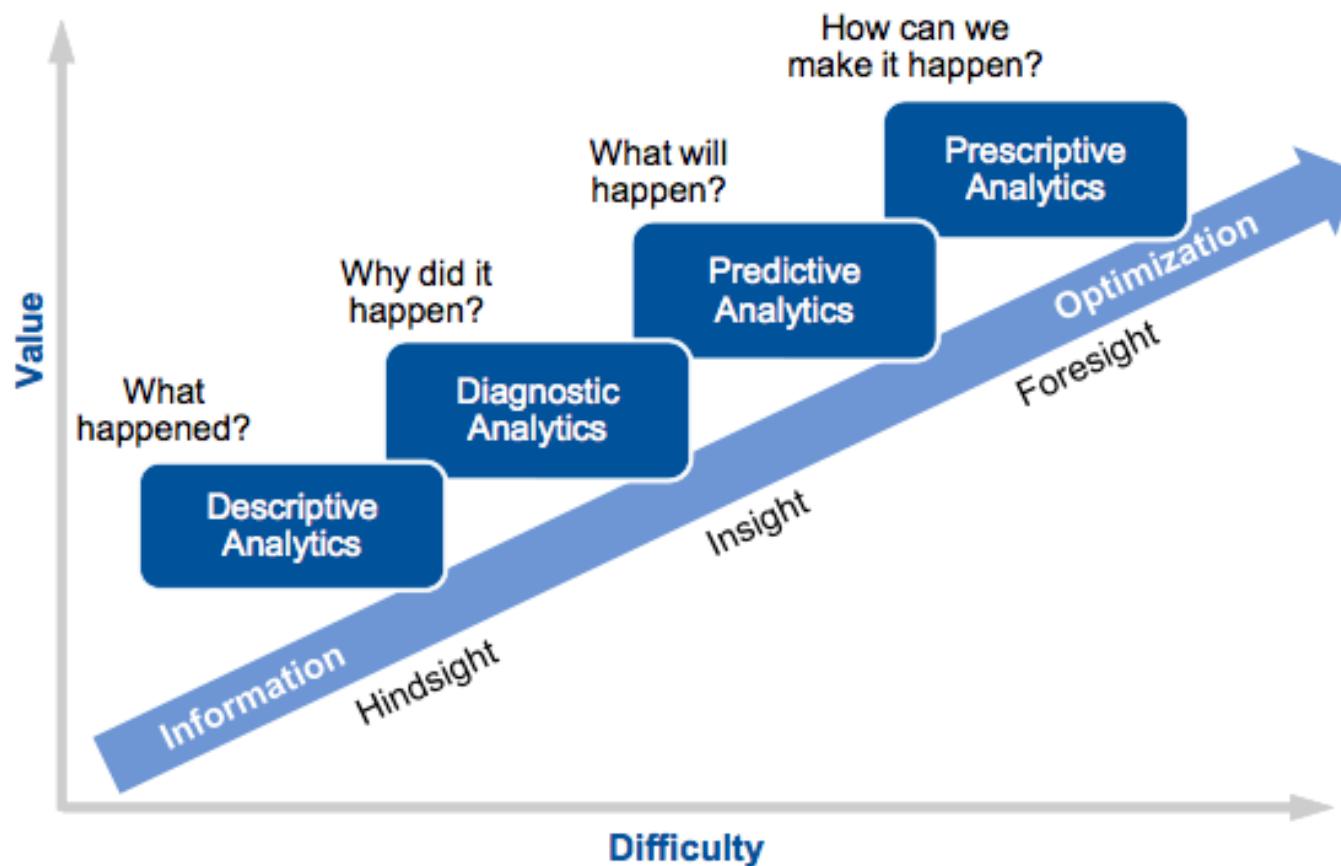
Ingenio en los negocios



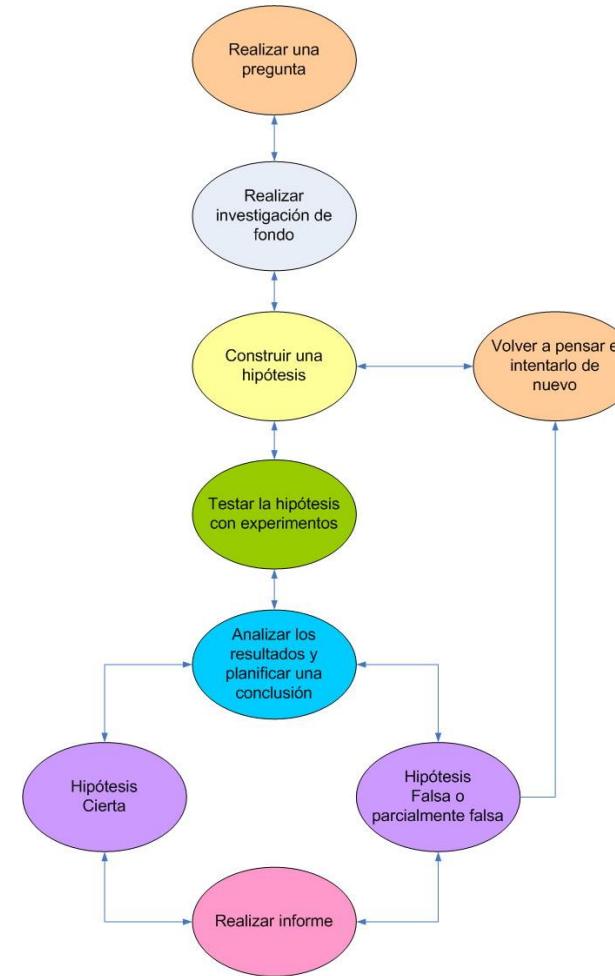
Tecnología

Aun que en menor grado, la minería de datos requiere también de estas habilidades.

Los 4 niveles del análisis de datos

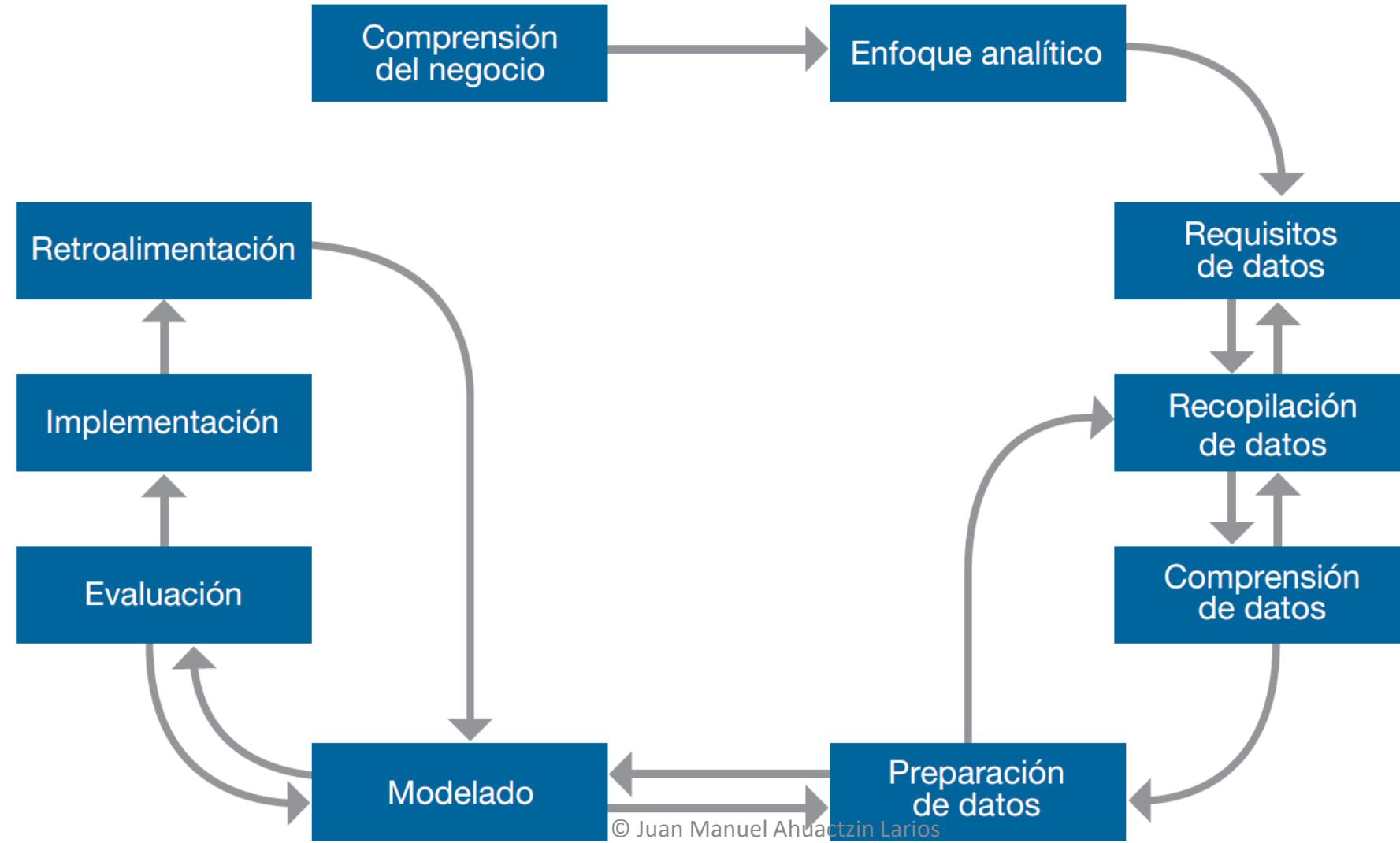


Source: Gartner (March 2012)



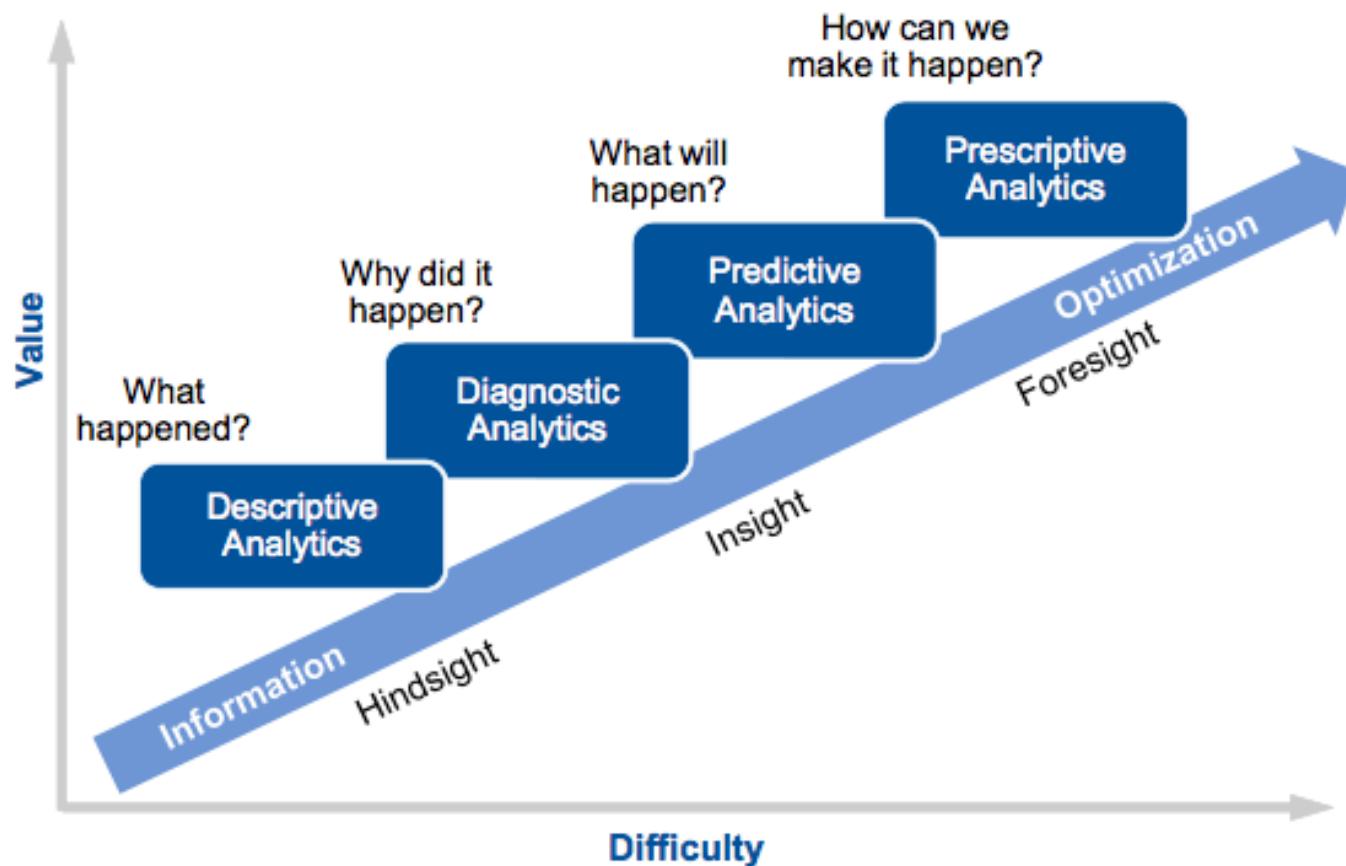
Modelo simplificado de las etapas del método científico

Metodología



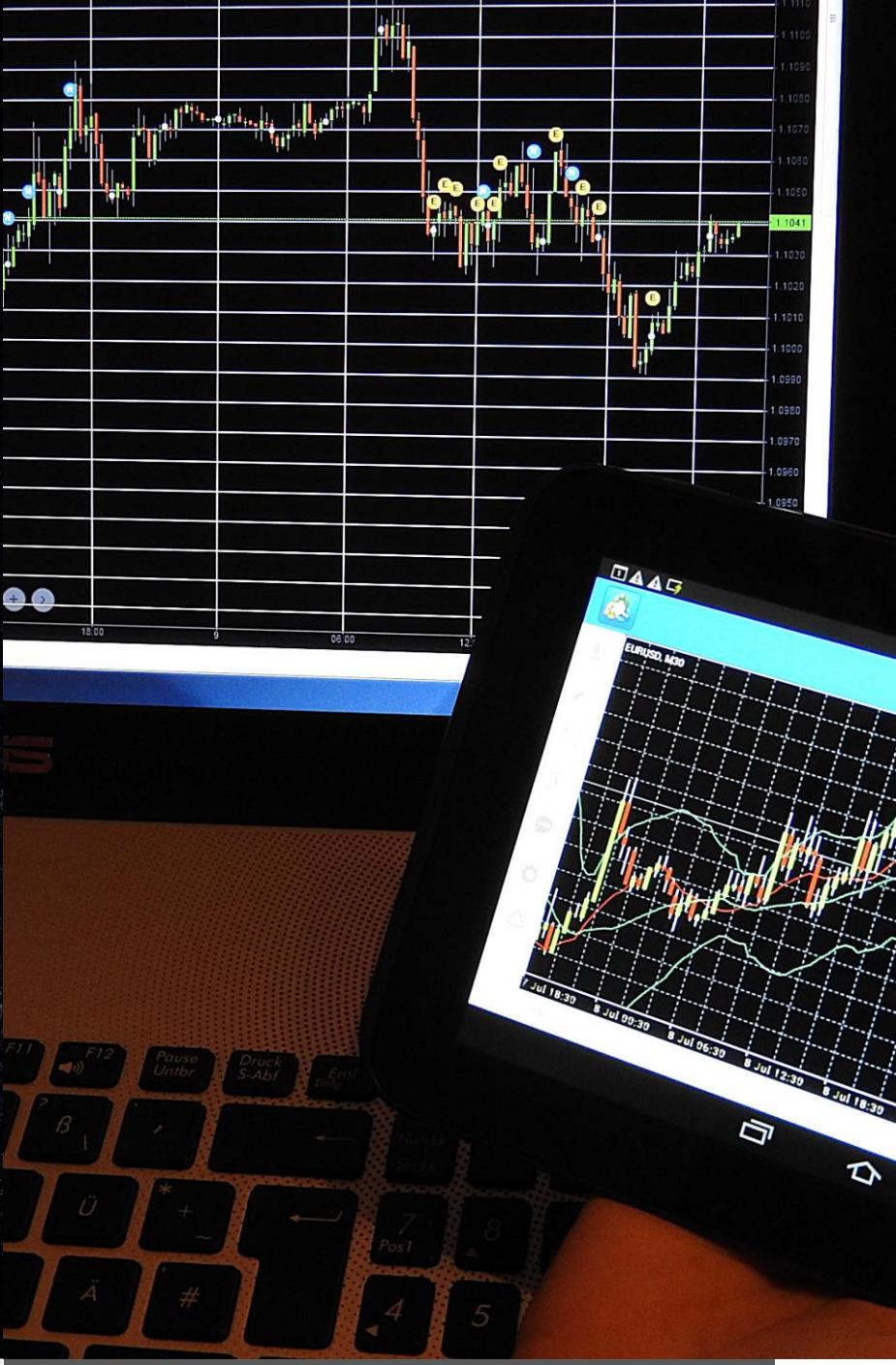
Avisos

Los 4 niveles del análisis de datos



Source: Gartner (March 2012)

Modelos predictivos



Modelos predictivos

Son técnicas **estadísticas** que permiten predecir salidas o comportamiento futuros.

El significado de “**Futuro**” puede estar relacionado u una línea del tiempo o no, esto es se puede referir también a un evento desconocido.

Modelos predictivos (los problemas)

1. Clasificación (*Classification*)
2. Agrupación en clústers (*Clustering*)
3. Predicción (*Forecast*)
4. Valores atípicos (*Outlier*)
5. Series temporales(*Time series*)

1.- Modelos de clasificación

Categorizar datos en un número dado de clase.

¿El cliente dejará la subscripción?

¿Nuestro cliente pagará a tiempo?

¿Qué tipo de auto comprará el prospecto?

¿Este e-mail es un spam?

¿El producto está dañado?

¿Qué carácter es este?



Binary Classification

Multi-Class Classification

Multi-Label Classification

Imbalanced Classification

2.- Agrupación en clústers

Ordenar individuos, objetos, animales, plantas,... entidades, en grupos separados basados en atributos similares.

Hard clustering



Clúster 1

Clúster 2

Clúster 3

Clúster 4

Clúster 5

Clúster 6

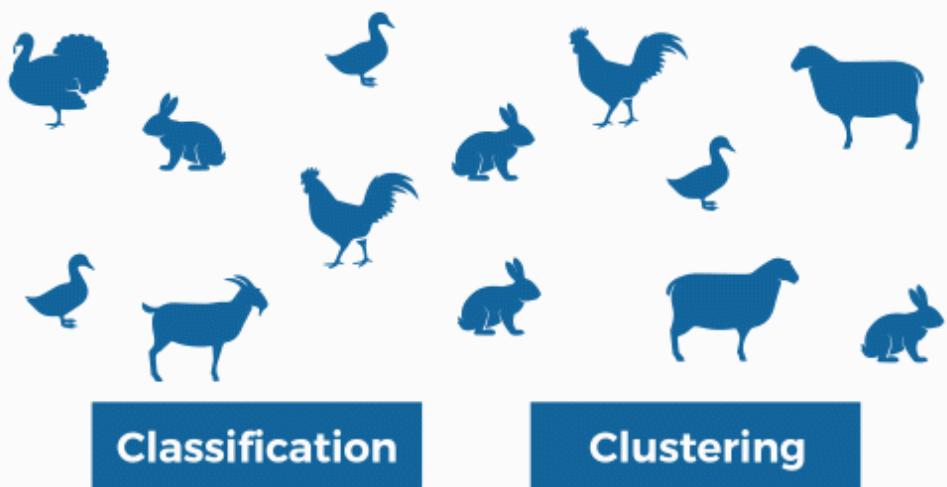
Soft clustering



2.- Agrupación en clústers

- Posibles áreas "buenas" para un negocio
- Diferentes tipos de clientes
- Separar las clases de animales
- Clasificar a los clientes para un préstamo
- Predecir de qué idioma es un nombre

2.- Clasificación vs agrupación en clústers



Clasificación: clases predefinidas
(aprendizaje supervisado)

Agrupación en clases: identificar las
similaridades (aprendizaje no supervisado)

3.- Predicción

Estimar el valor de las nuevas entradas de datos basándose en lo aprendido con los datos históricos.

Predecir:

- La cantidad de clientes que recibiremos en el día
- La cantidad de energía que se necesitará en el día
- Las necesidades de inventario
- Quién ganará la copa del mundo

4.- Valores atípicos

Identificar datos anómalos: transacciones, salidas, entidades, etc.

- Transacciones o reclamaciones fraudulentas
- Ciberataques
- Consumo inusual de energía
- Problemas de producción

5.- Series temporales

Predecir las salidas o tendencias a través del tiempo.

- Ventas para los próximos 3 días.
- Temperaturas para la próxima semana.
- Valor de los productos en el mercado.

Los modelos predictivos en resumen

Problema	Objetivo	Ejemplo
Clasificación	Categorizar datos en un número dado de clase.	Spam o no spam.
Agrupación en clústers	Ordenar individuos, objetos, animales, plantas,... entidades, en grupos separados basados en atributos similares.	Posibles áreas "buenas" para un negocio
Predictión	Estimar el valor de las nuevas entradas de datos basándose en lo aprendido con los datos históricos.	La cantidad de energía que se necesitará en el día de hoy.
Valores atípicos	Identificar datos anómalos: transacciones, salidas, entidades, etc.	Transacciones o reclamaciones fraudulentas
Series temporales	Predecir las salidas o tendencias a través del tiempo.	Ventas para los próximos 3 días.

1.3 El problema de la separación lineal.



Recordatorio

Esperanza matemática

Para una variable discreta

$$E(X) = \sum_x xP(x).$$

Para una variable continua

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

Donde $f(x)$ es la función de densidad de X.

Covarianza y correlación

La **covarianza** entre dos variables aleatorias X_1 y X_2 se define como

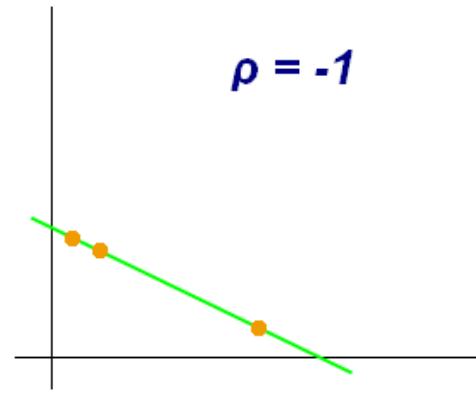
$$\text{Cov}(X_1, X_2) = E[(X_1 - E[X_1])(X_2 - E[X_2])]$$

El **coeficiente de correlación lineal** se define como:

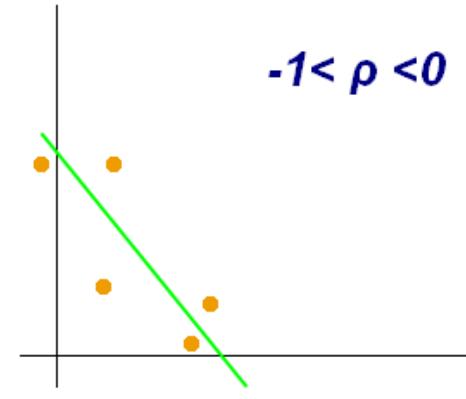
$$\rho = \frac{\text{Cov}(X_1, X_2)}{\sigma_1 \sigma_2}$$

También llamado coeficiente de correlación de Pearson

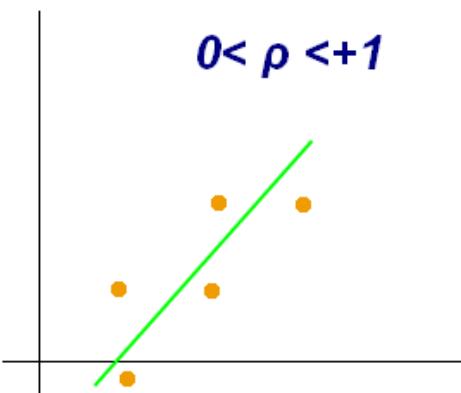
Covarianza y correlación



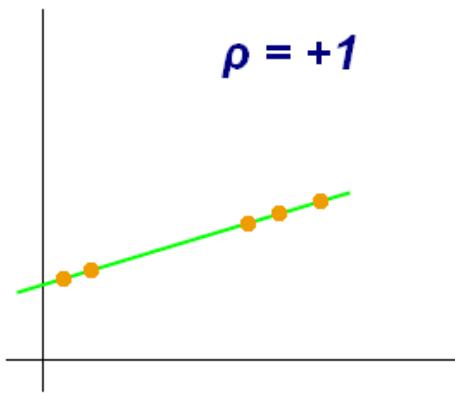
$$\rho = -1$$



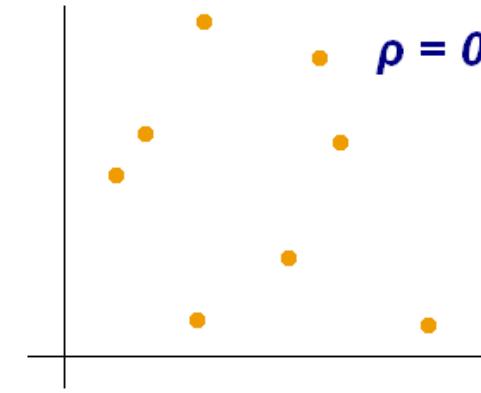
$$-1 < \rho < 0$$



$$0 < \rho < +1$$

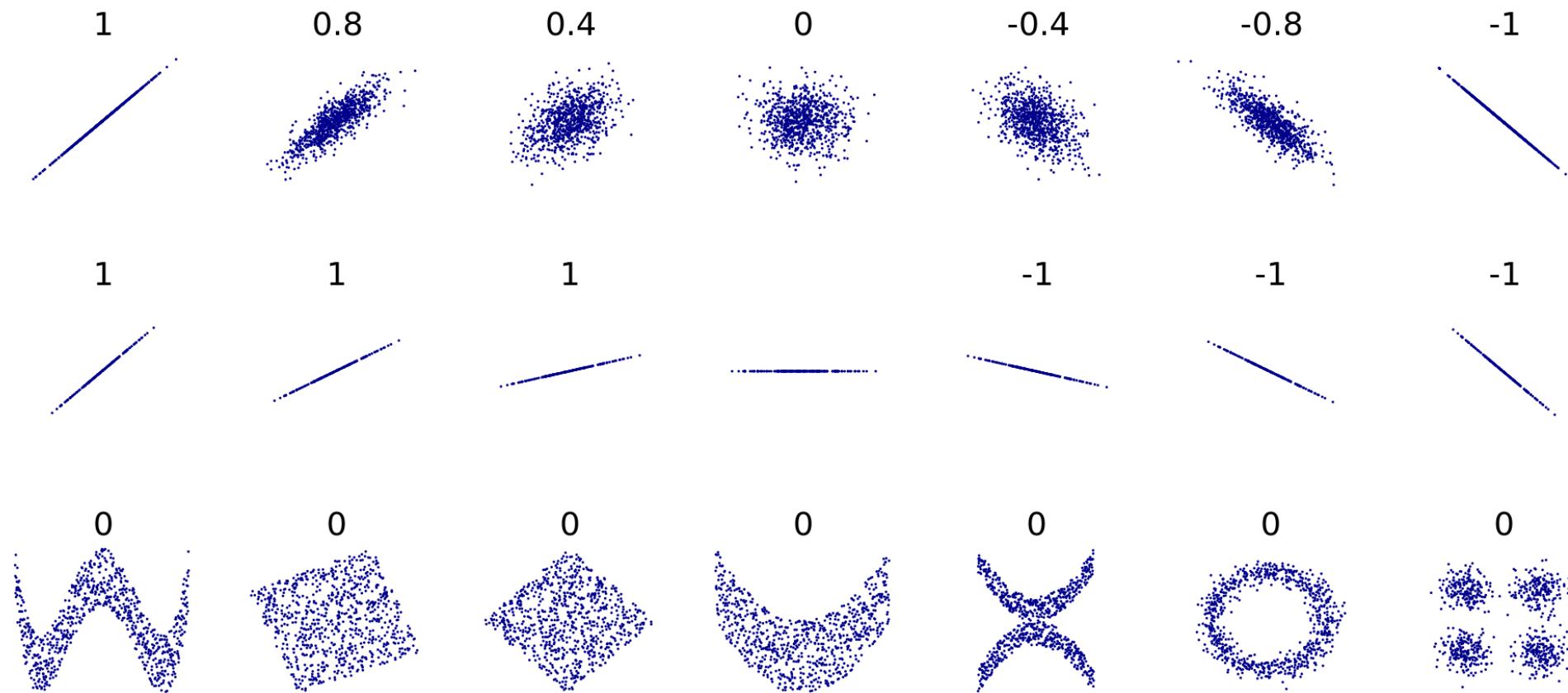


$$\rho = +1$$



$$\rho = 0$$

Covarianza y correlación



Covarianza y correlación

Para un conjunto de n variables aleatorias la **matriz de covarianza** está dada por

$$S = \begin{pmatrix} \text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Cov}(X_2, X_2) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \cdots & \text{Cov}(X_n, X_n) \end{pmatrix}$$

Covarianza y correlación

Para un conjunto de n variables aleatorias la **matriz de correlación muestral** está dada por

$$S_{\rho} = \begin{pmatrix} \rho_{1,1} & \rho_{1,2} & \cdots & \rho_{1,n} \\ \rho_{2,1} & \rho_{2,2} & \cdots & \rho_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n,1} & \rho_{n,2} & \cdots & \rho_{n,n} \end{pmatrix}$$

Ejemplo 1: Covarianza y correlación

Covarianza y correlación

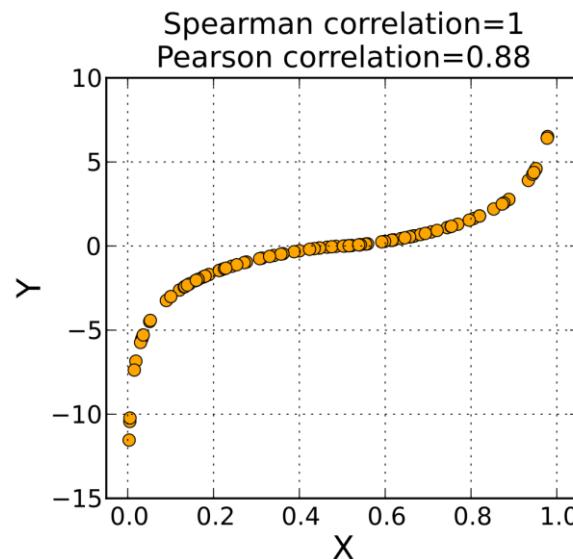
Otras medidas de correlación:

- Coeficiente de correlación de rango de Kendall

$$\tau = \frac{(\text{número de pares concordantes}) - (\text{número de pares discordantes})}{\binom{n}{2}}$$

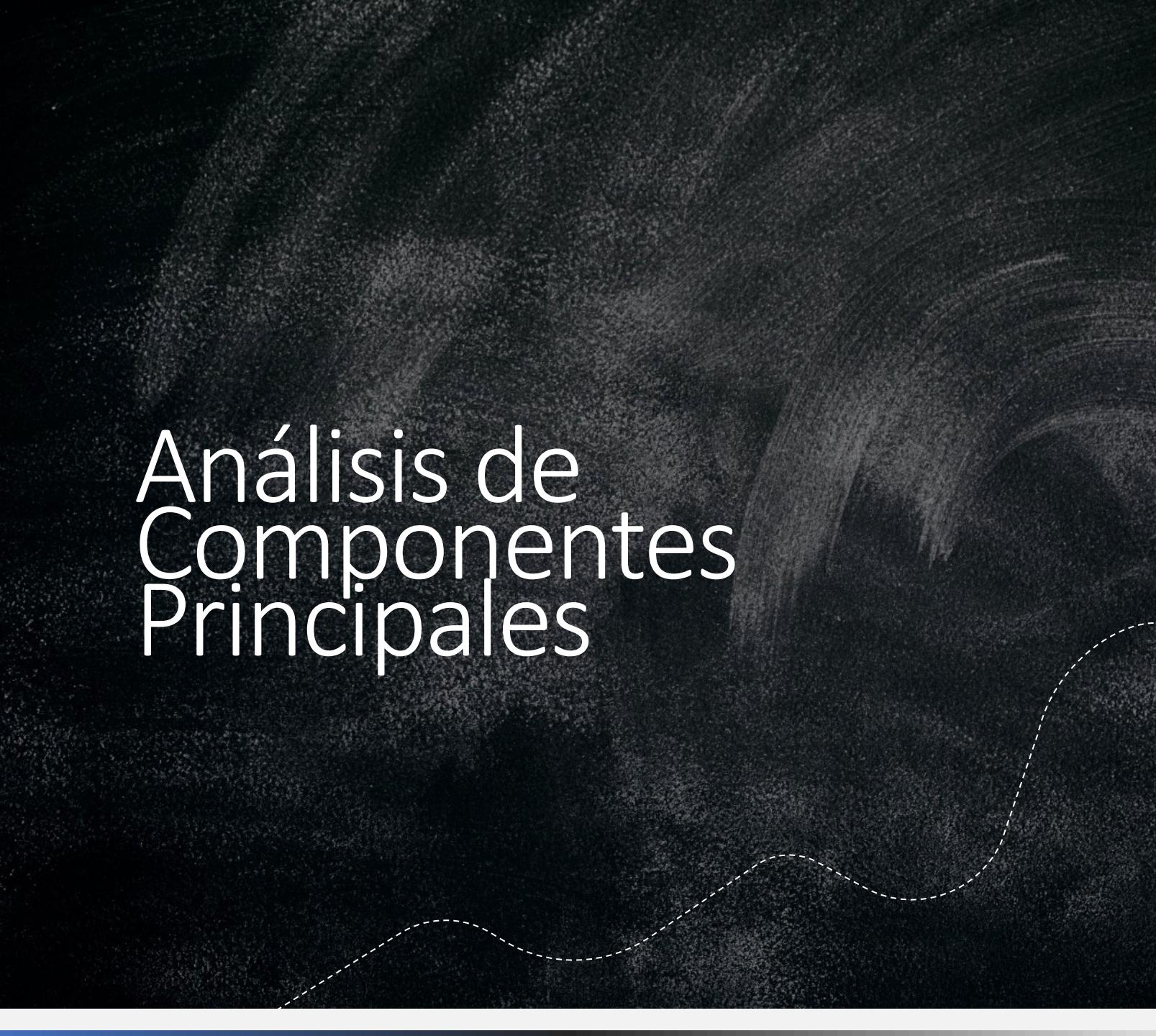
- Coeficiente de correlación de Spearman

$$\tau_s = \frac{\text{Cov}(rg_{x1}, rg_{x2})}{\sigma_{rgx1} \sigma_{rgx2}}$$

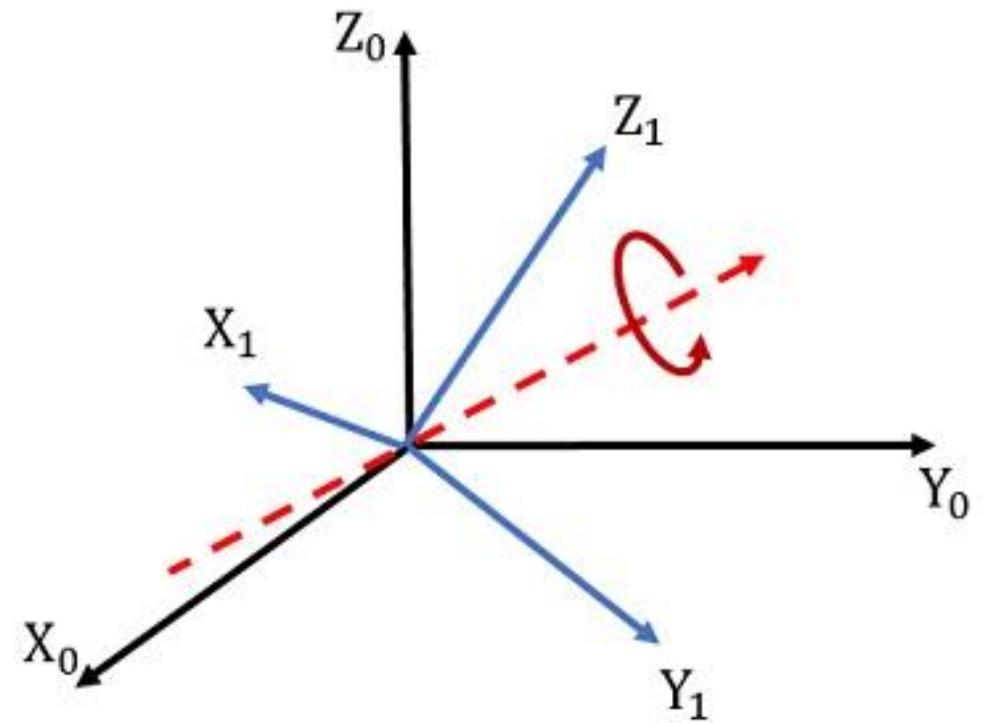


Ejemplo 01: Análisis de correlación invariable

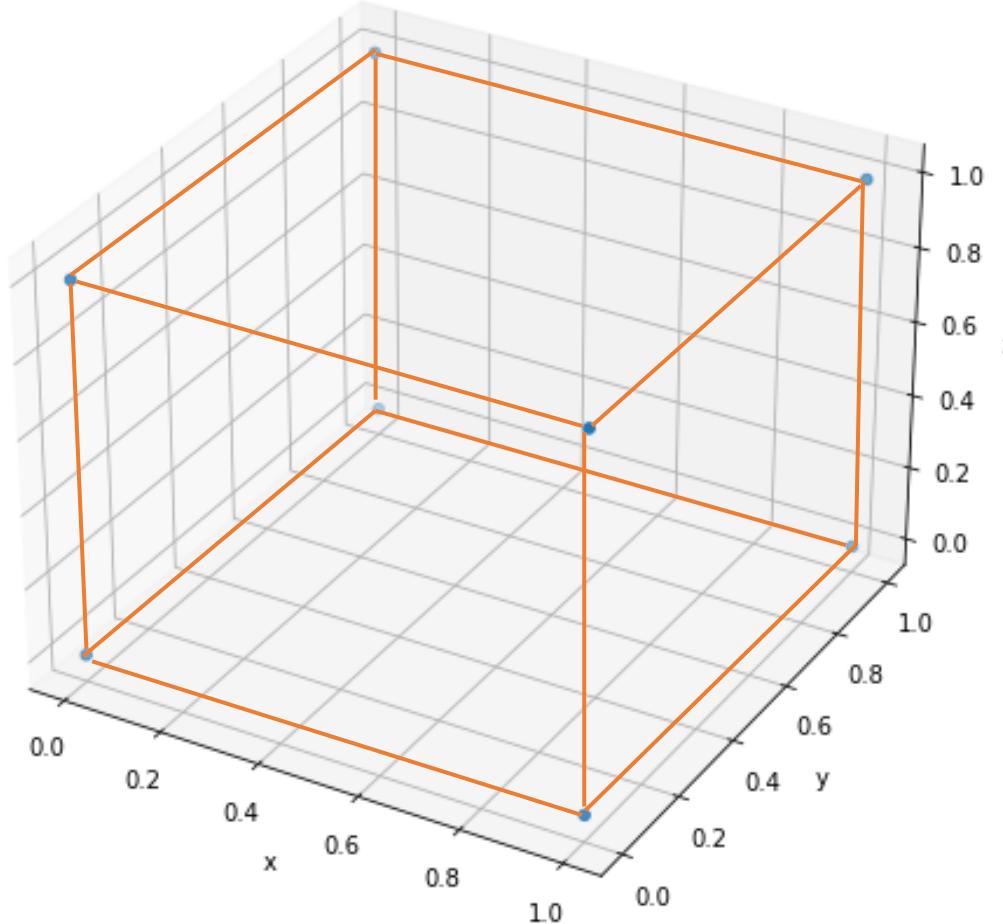
Análisis de Componentes Principales



Covarianza y correlación

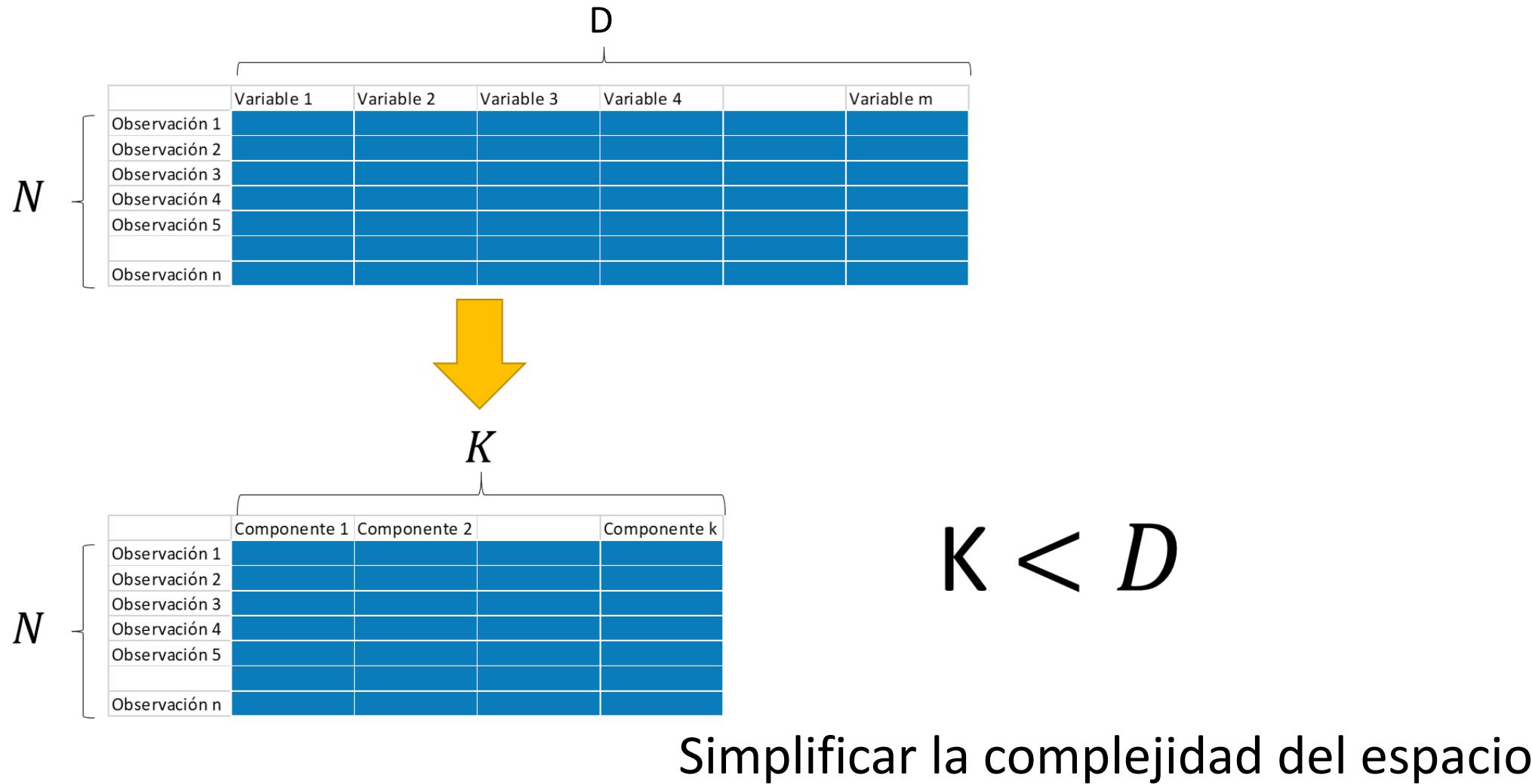


Covarianza y correlación



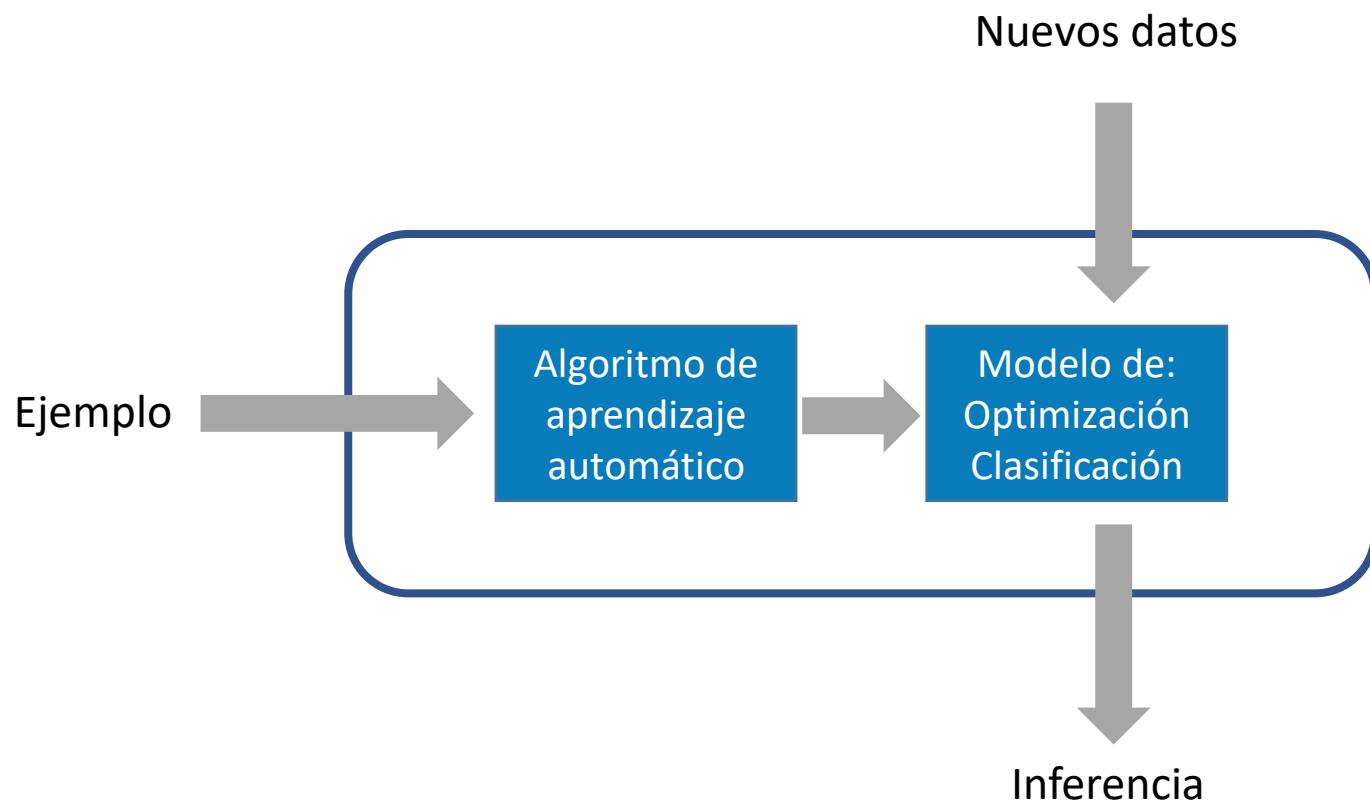
Ejemplo 02: Análisis de cambio de correlación

Análisis de Componentes Principales (ACP o PCA): Reducción de la dimensionalidad



ACP: Reducción de la dimensionalidad

Aprendizaje
automático o
máquina



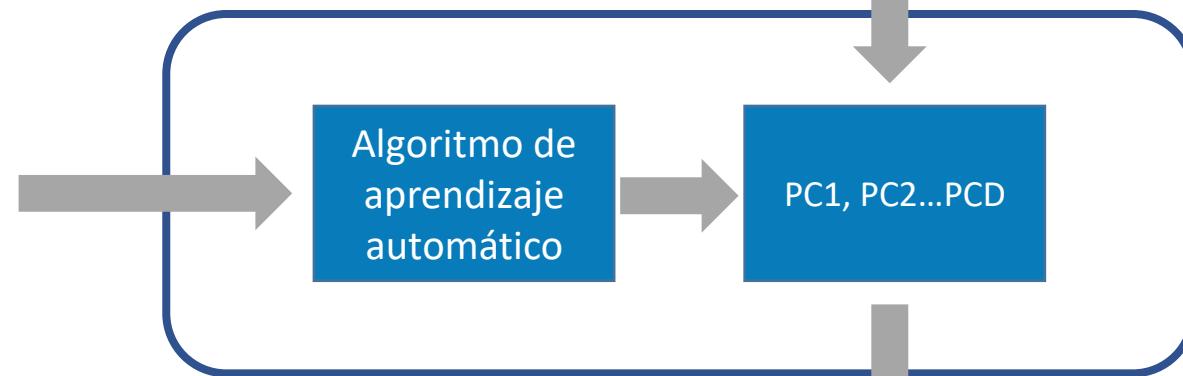
ACP: Reducción de la dimensionalidad

	Variable 1	Variable 2	Variable 3	Variable 4		Variable m
Observación 1						
Observación 2						
Observación 3						
Observación 4						
Observación 5						
Observación n						

Ejemplo

	Variable 1	Variable 2	Variable 3
Observación 1			
Observación 2			
Observación 3			
Observación 4			
Observación 5			
Observación n			

Nuevos datos

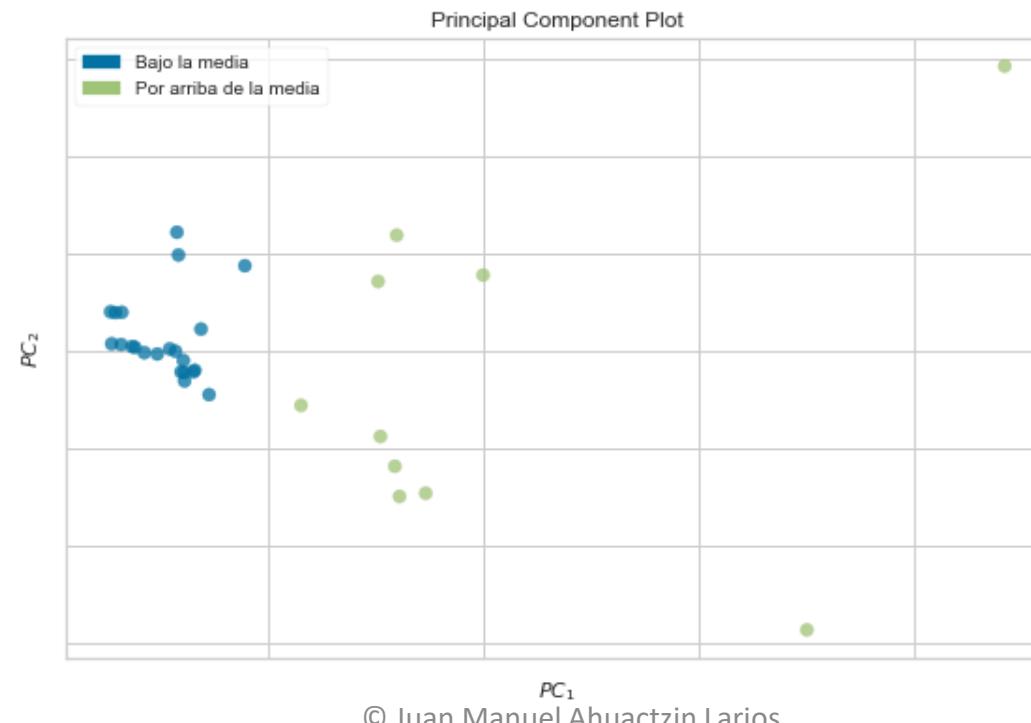


	Variable 4	Variable m

PCA: Aplicaciones

El análisis de componentes principales tiene dos aplicaciones primordiales:

1. La visualización
2. El preprocessado de predictores con el objetivo de realizar ajustes en los modelos supervisados.



ACP: Conceptos matemáticos

Vectores y valores propios de una matriz*

Sea A una matriz de $n \times n$ con componentes reales, λ un número real y v un vector distinto de cero, entonces si se cumple que

$$Av = \lambda v.$$

Entonces

- v es el vector característico de A y
- λ el valor característico de A

* También llamados vectores y valores característicos o eigenvector y eigenvalues.

ACP: Conceptos matemáticos

Vectores y valores propios de una matriz

$$\begin{pmatrix} 10 & -18 \\ 6 & -11 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 10 * 2 - 18 * 1 \\ 6 * 2 - 11 * 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

Por lo tanto

$$v = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \text{ y } \lambda = 1$$

ACP: Conceptos matemáticos

Vectores y valores propios de una matriz

$$\begin{pmatrix} 10 & -18 \\ 6 & -11 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \begin{pmatrix} -6 \\ -4 \end{pmatrix}$$

$$\begin{pmatrix} 10 * 3 - 18 * 2 \\ 6 * 3 - 11 * 2 \end{pmatrix} = -2 \begin{pmatrix} 3 \\ 2 \end{pmatrix}$$

Por lo tanto

$$v = \begin{pmatrix} 3 \\ 2 \end{pmatrix} \text{ y } \lambda = -2$$

ACP: Conceptos matemáticos

Vectores y valores propios de una matriz

$$v_1 = \begin{pmatrix} 3/\sqrt[2]{13} \\ 2/\sqrt[2]{13} \end{pmatrix}$$

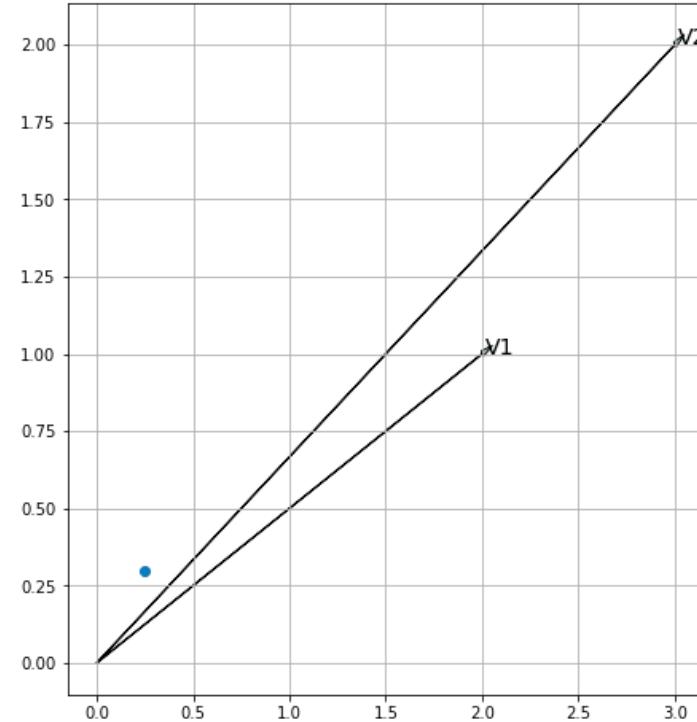
$$v_2 = \begin{pmatrix} 2/\sqrt[2]{5} \\ 1/\sqrt[2]{5} \end{pmatrix}$$

ACP: Conceptos matemáticos

Vectores y valores propios de una matriz

$$\begin{pmatrix} 10 & -18 \\ 6 & -11 \end{pmatrix} =$$

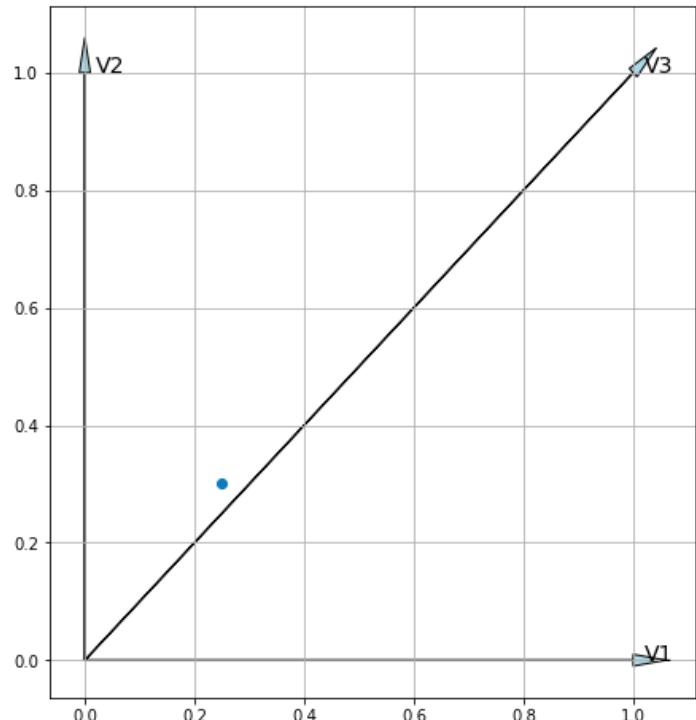
La matriz es representada por los **dos vectores**.



ACP: Conceptos matemáticos

Vectores y valores propios de una matriz

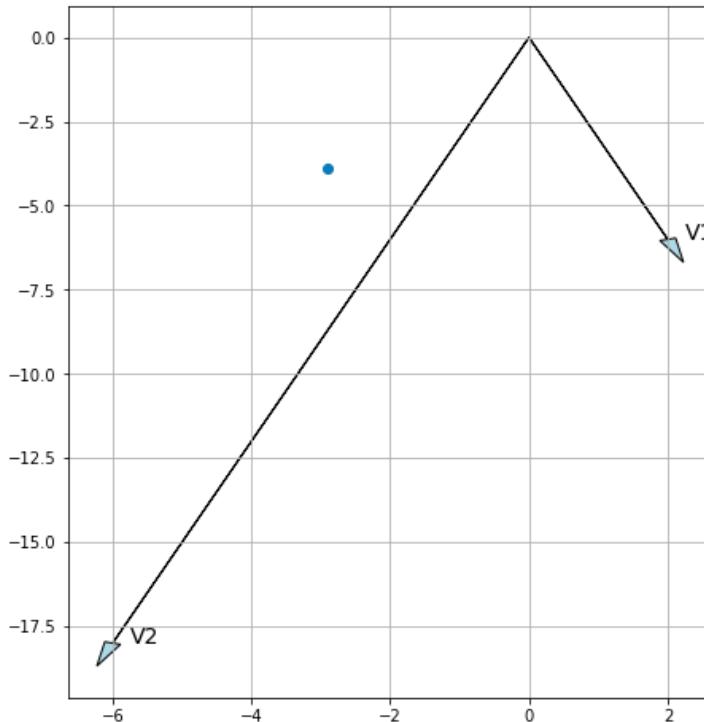
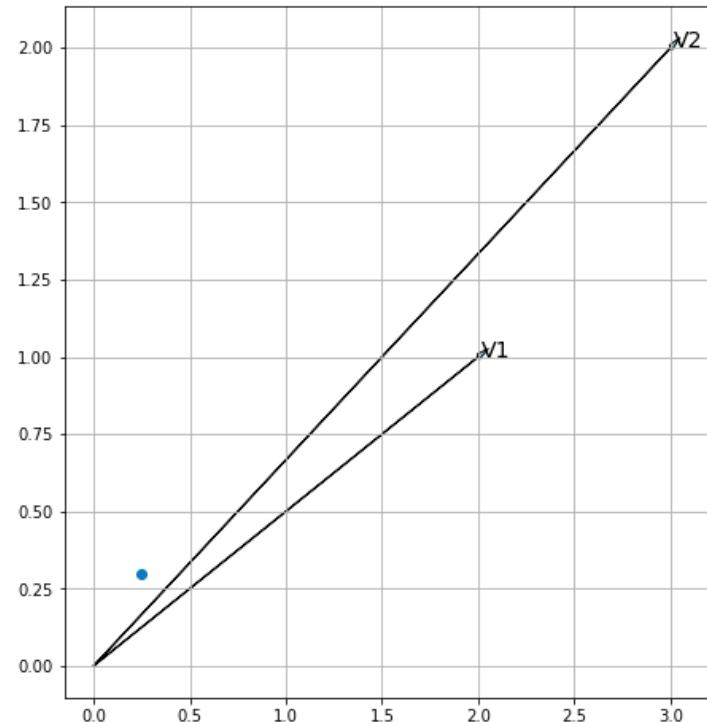
Primero veamos
la transformación
Con ejes
conocidos.



ACP: Conceptos matemáticos

Vectores y valores propios de una matriz

Ahora veamos la transformación con otros vectores.



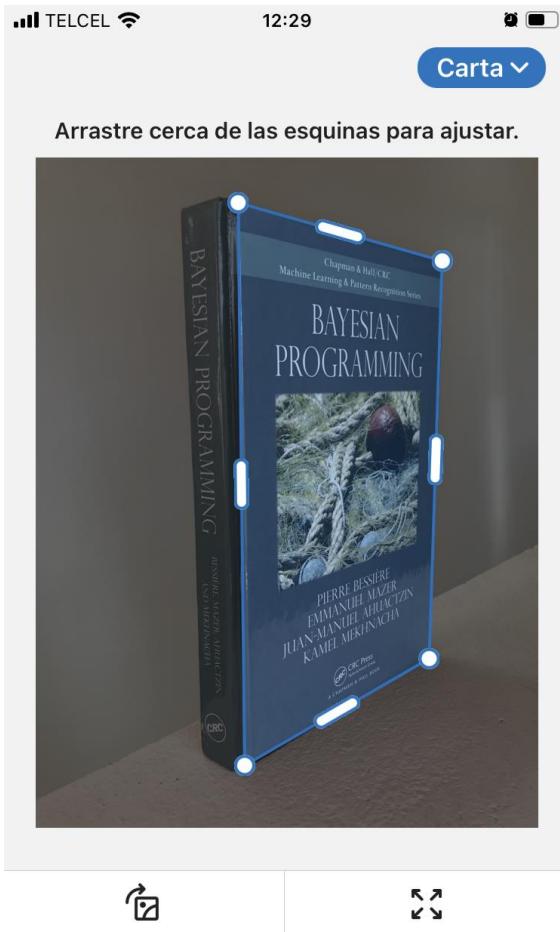
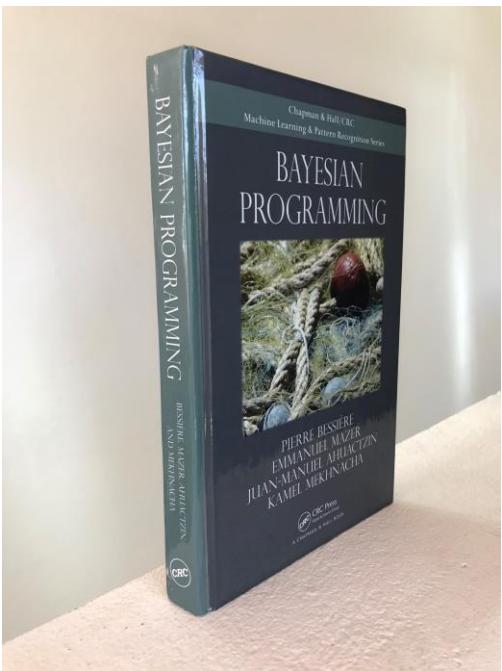
ACP: Conceptos matemáticos

Analogía con una homografía.

Una homografía es una transformación proyectiva 2D que mapea puntos de un plano a otro.

$$\begin{bmatrix} x' \\ y' \\ w' \end{bmatrix} = \begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{bmatrix} \begin{bmatrix} x \\ y \\ w \end{bmatrix}$$

ACP: Conceptos matemáticos



Cancelar

Hecho



Tiny Scanner - PDF Scanner App

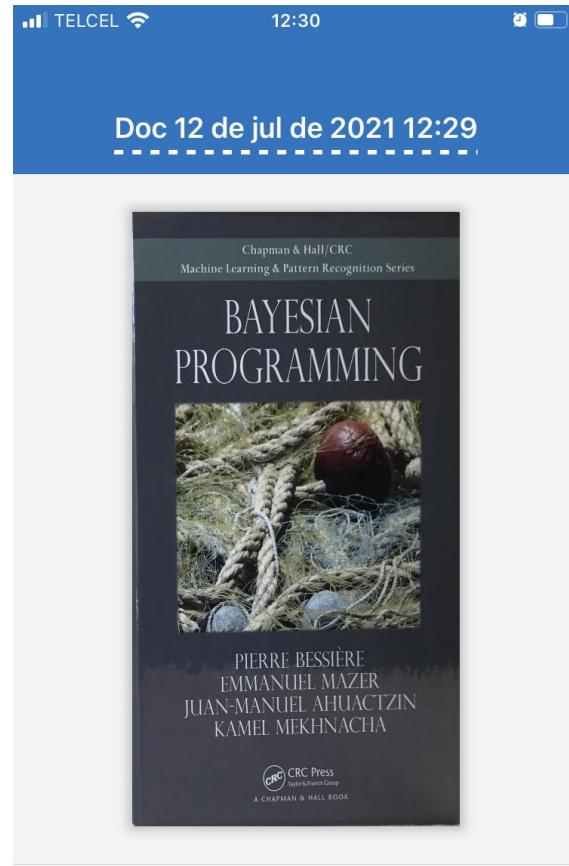
Beesoft Apps Business

Everyone

Contains Ads · Offers in-app purchases
⚠ You don't have any devices

Add to Wishlist

Install



Cancelar

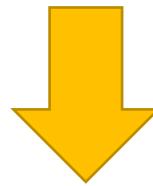
Guardar

ACP: Conceptos matemáticos

A large blue matrix representing data. The columns are labeled "Variable 1", "Variable 2", "Variable 3", "Variable 4", ..., "Variable m". The rows are labeled "Observación 1", "Observación 2", "Observación 3", "Observación 4", "Observación 5", ..., "Observación n". A bracket on the left indicates there are N observations, and a bracket at the top indicates there are D dimensions.

	Variable 1	Variable 2	Variable 3	Variable 4		Variable m
Observación 1						
Observación 2						
Observación 3						
Observación 4						
Observación 5						
Observación n						

1. Consideraremos un conjunto de N observaciones con D dimensiones.



K

A smaller blue matrix representing projections. The columns are labeled "Componente 1", "Componente 2", ..., "Componente k". The rows are labeled "Observación 1", "Observación 2", "Observación 3", "Observación 4", "Observación 5", ..., "Observación n". A bracket on the left indicates there are N observations, and a bracket at the top indicates there are K components.

	Componente 1	Componente 2		Componente k
Observación 1				
Observación 2				
Observación 3				
Observación 4				
Observación 5				
Observación n				

2. Proyectemos las N observaciones en un espacio de dimensión $K < D$.

ACP: Conceptos matemáticos

Primero proyectemos las observaciones **con $k = 1$** por medio de un vector unitario u_1

$$u_1^T u_1 = 1$$

Cada punto es entonces proyectado en un valor escalar dado por

$$s_{1,n} = u_1^T X_n$$

La media de los datos proyectados es entonces

$$m_1 = u_1^T \bar{x}$$

ACP: Conceptos matemáticos

Con un dato y un componente

$$s_{1,n} = \mathbf{u}_1^T \mathbf{X}_n$$

Estado	Poblacion	Confirmados	Defunciones	Ficticio
AGUASCALIENTES	1,434,635	26,770	2,469	87,719



Normalización

Estado	PC1	Poblacion	Confirmados	Defunciones	Ficticio	
AGUASCALIENTES	-1.151600	PC1	0.461464	0.479641	0.526949	0.528507

=

Estado	Población	Confirmados	Defunciones	Ficticio
AGUASCALIENTES	-0.778576	-0.438388	-0.595642	-0.507415

$$s_{1,n}$$

$$\mathbf{u}_1^T$$

$$\mathbf{X}_n$$

ACP: Conceptos matemáticos

La **varianza** de los datos proyectados esta dada por:

$$\frac{1}{n} \sum_{i=1}^n \{\mathbf{u}_1^T \mathbf{x}_i - \mathbf{u}_1^T \bar{\mathbf{x}}\}^2 = \mathbf{u}_1^T S \mathbf{u}_1$$

Donde S es la matriz de covarianza de los datos

$$S = \begin{pmatrix} \text{Cov}(X_1, X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Cov}(X_2, X_2) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \cdots & \text{Cov}(X_n, X_n) \end{pmatrix}$$

ACP: Conceptos matemáticos

Es posible mostrar que la varianza $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$ se maximiza cuando \mathbf{u}_1 es el vector propio de \mathbf{S} que **maximiza** su valor propio λ_1

$$\mathbf{S} \mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

$$\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1 = \lambda_1$$

A \mathbf{u}_1 se le conoce como la primera componente principal.

ACP: Conceptos matemáticos

Entre mayor varianza más dispersión de los datos.

ACP: Conceptos matemáticos

Es la matriz S es una matriz de $n \times n$ y por lo tanto tiene n vectores propios y n valores propios.

El orden de los vectores propios está dado por los valores de los valores propios, definiéndose u_1, u_2, \dots, u_n como las componentes principales.

ACP: Conceptos matemáticos

Con un dato y varios componentes

Estado	PC1	PC2	PC3	PC4	Poblacion	Confirmados	Defunciones	Ficticio
	0.461464	0.479641	0.526949	0.528507				
AGUASCALIENTES	-1.151600	0.271993	-0.081147	0.062276	PC1	-0.740218	0.642553	-0.104998 0.167865
					PC2	0.474771	0.434366	-0.764010 -0.046991
					PC3	-0.117135	-0.410362	-0.357196 0.830839
					PC4			

=

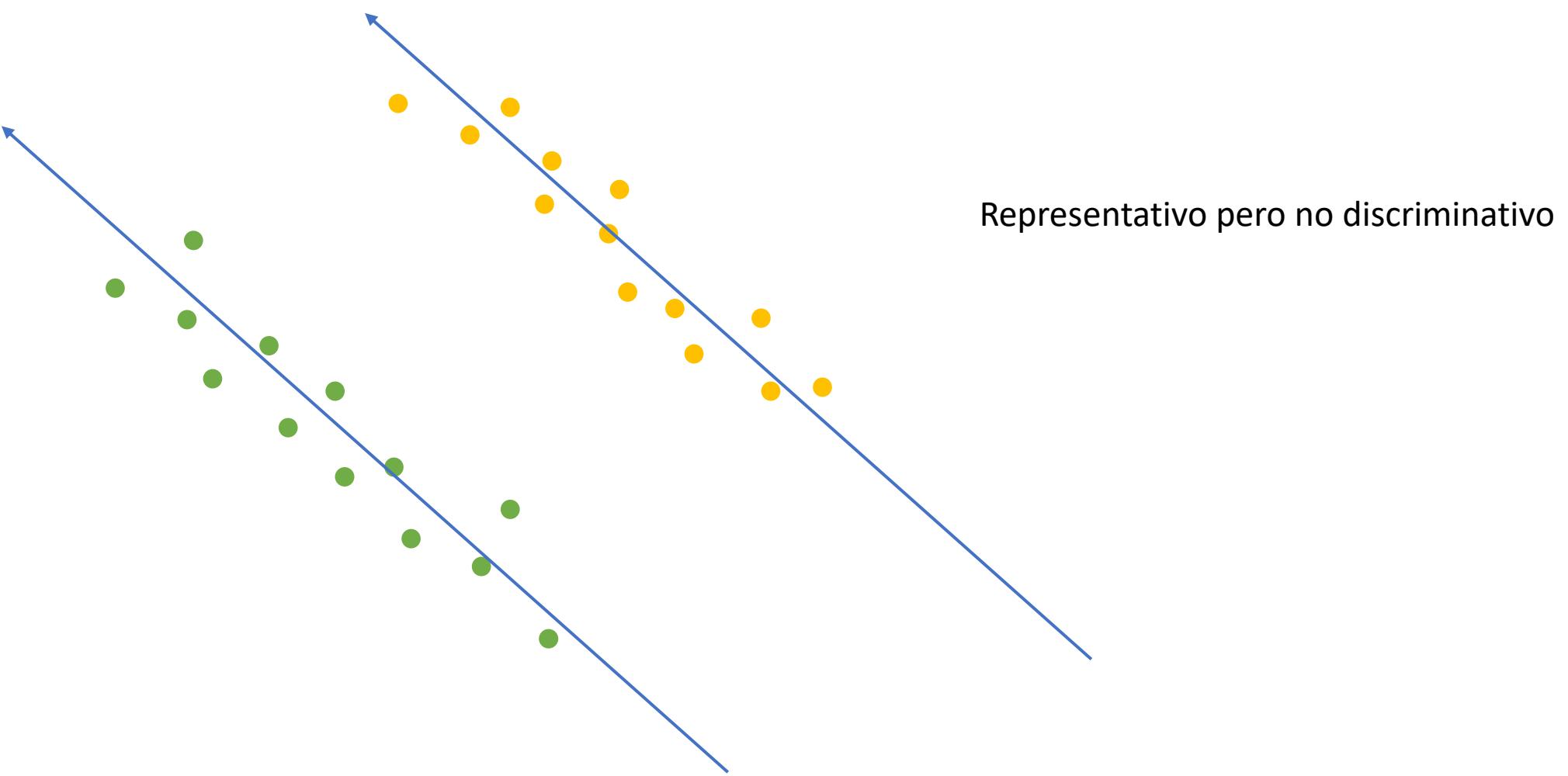
Estado	Población	Confirmados	Defunciones	Ficticio
	AGUASCALIENTES	-0.778576	-0.438388	-0.595642

ACP: Conclusiones

ACP encuentra la representación de los datos más precisa en un espacio de menor dimensión maximizando la varianza de las direcciones.

A pesar de ello tales direcciones podrían no funcionar bien cuando queremos realizar una clasificación.

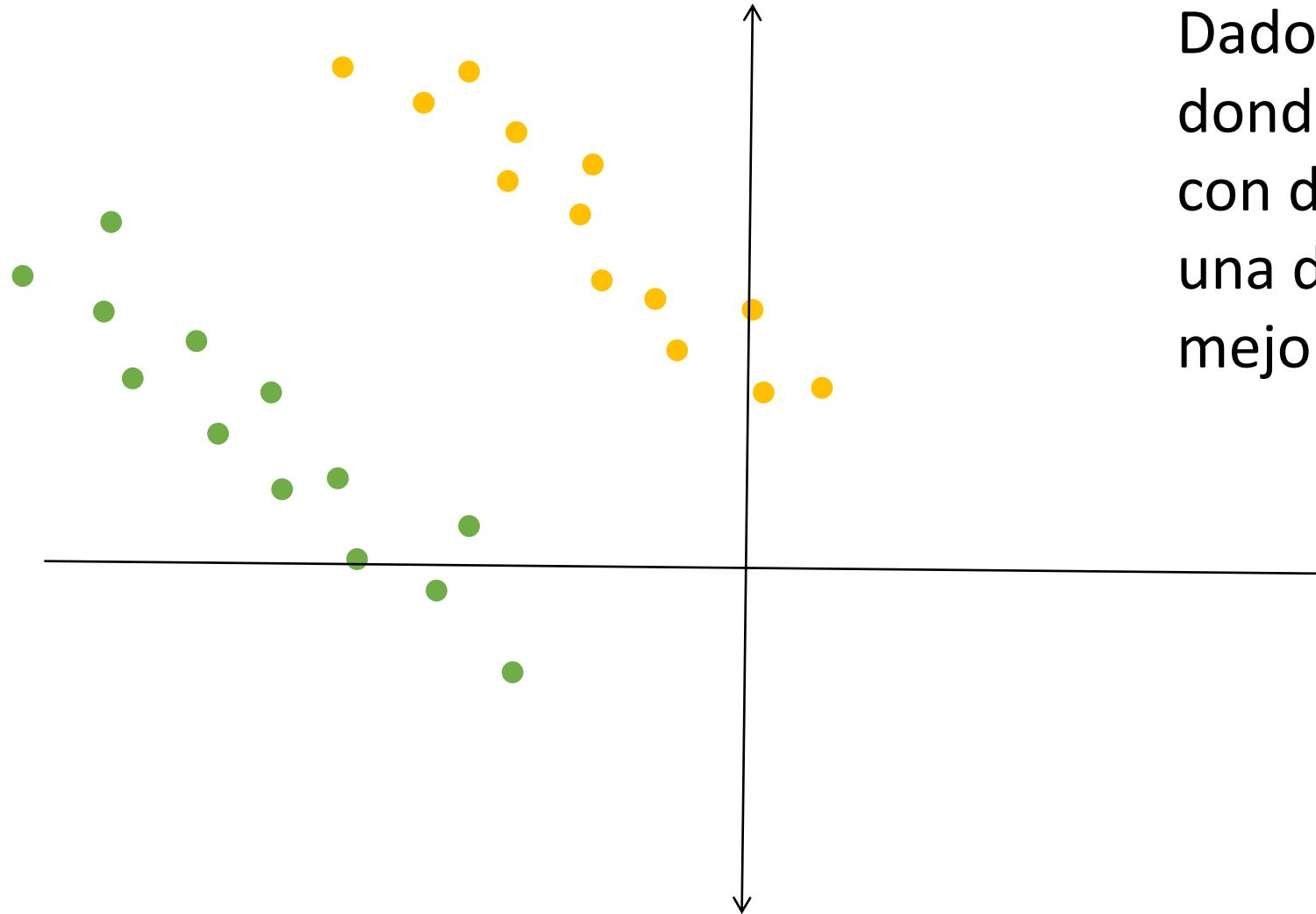
ACP: Conclusiones



Análisis discriminante lineal

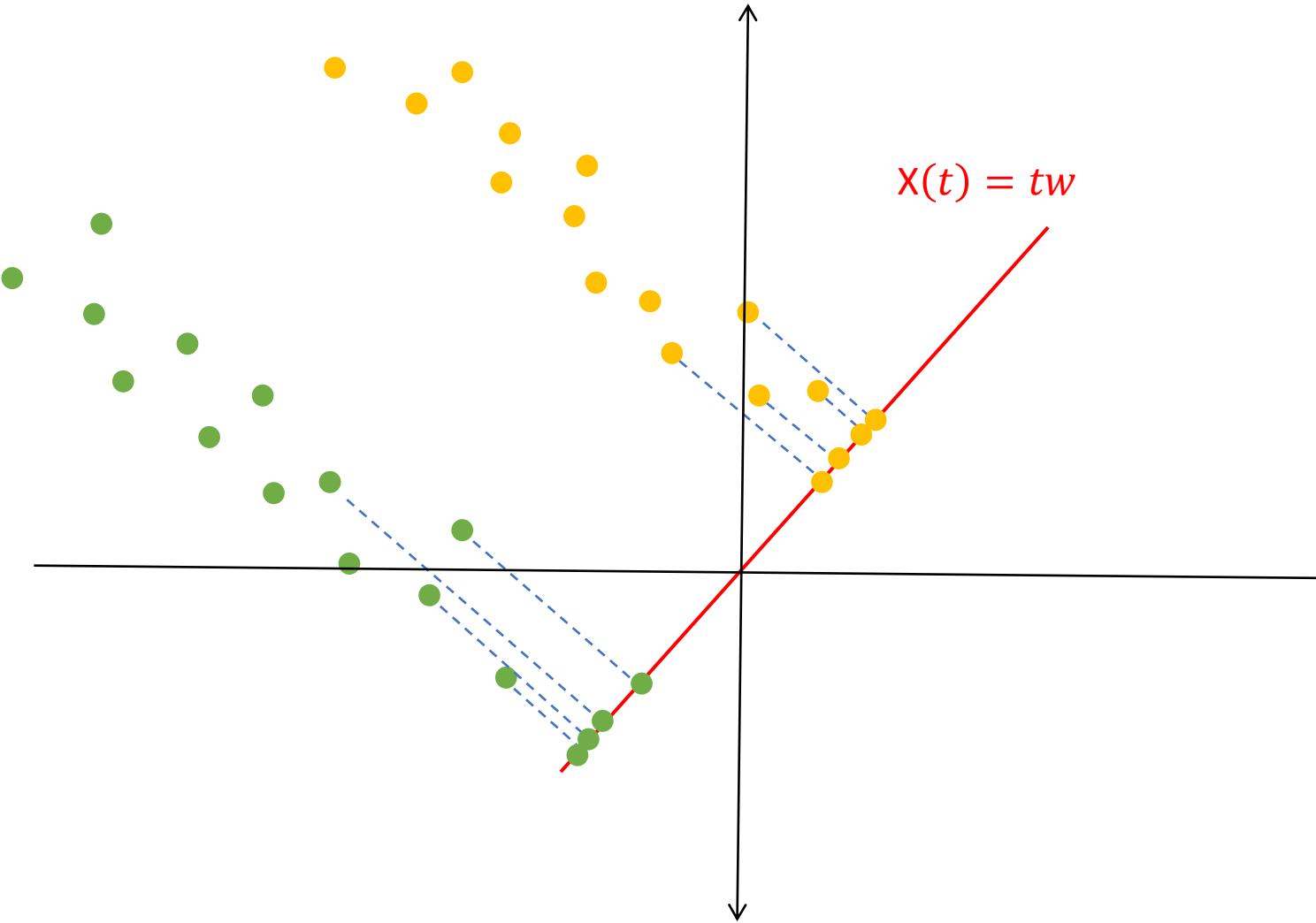


Análisis discriminante lineal: El problema de dos clases

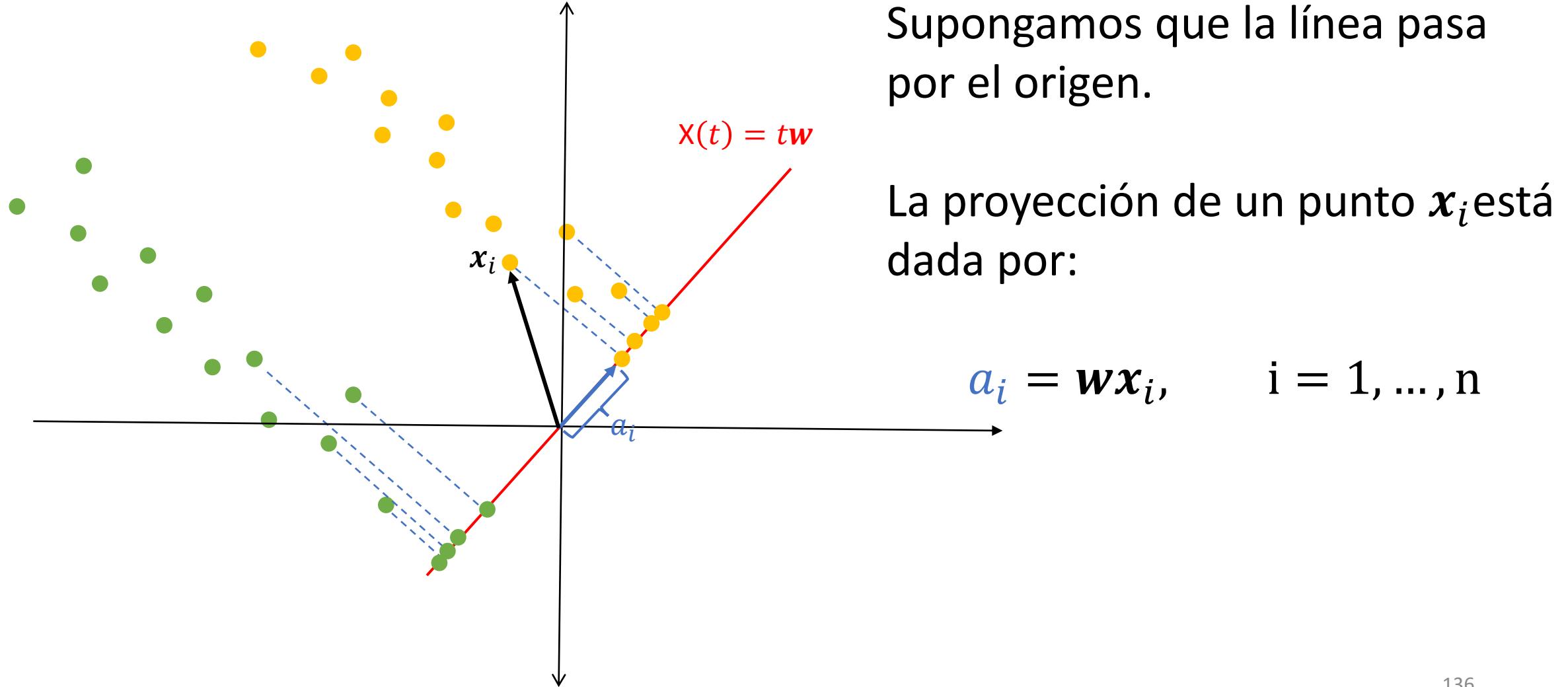


Dado un conjunto de datos $\{x_n\}$ donde $n = 1, \dots, N$ con $x_n \in \mathbb{R}^D$ con dos clases C_1 y C_2 encontrar una dirección que discrimina “lo mejor posible” a las dos clases.

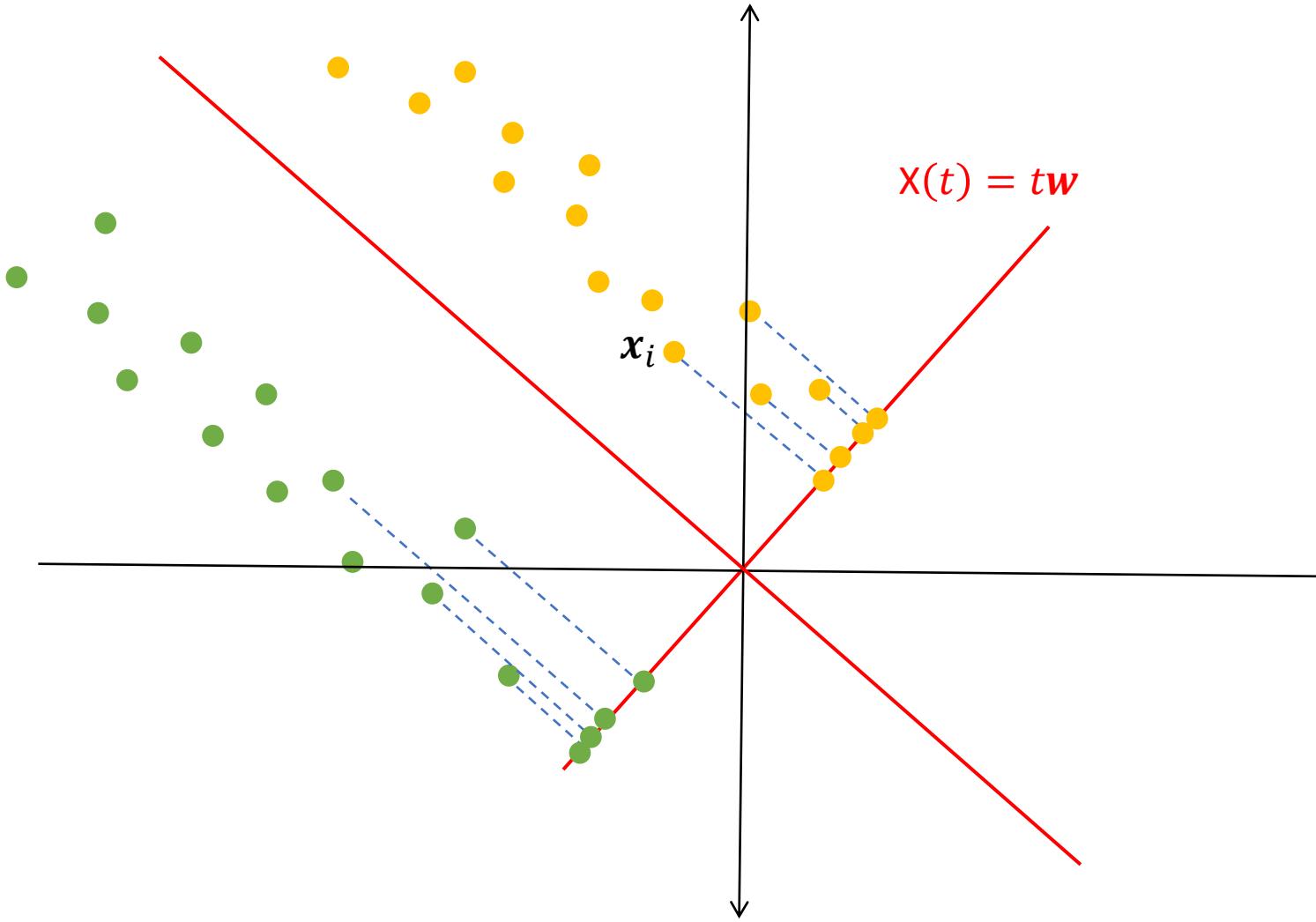
Análisis discriminante lineal: El problema de dos clases



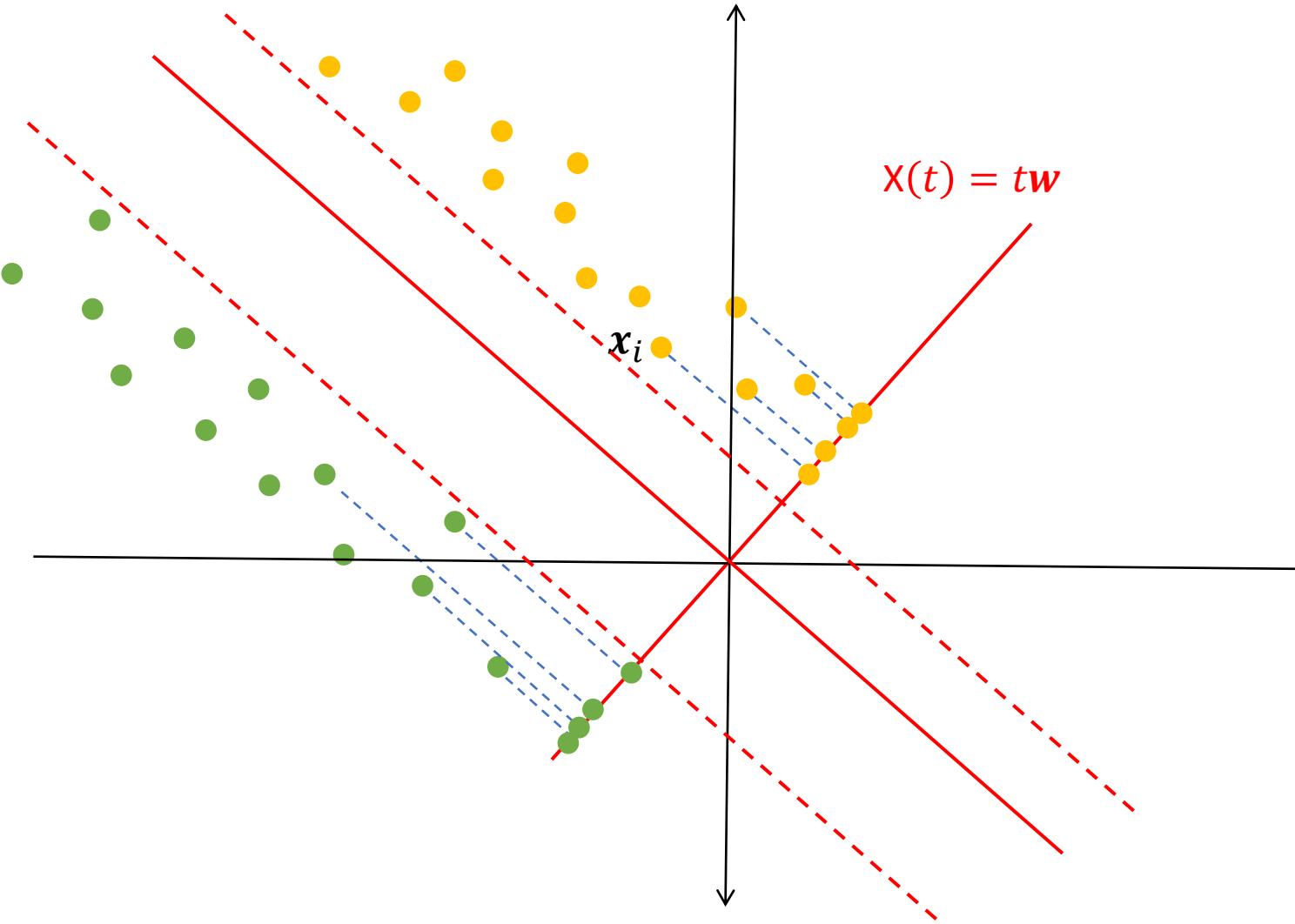
Análisis discriminante lineal: El problema de dos clases



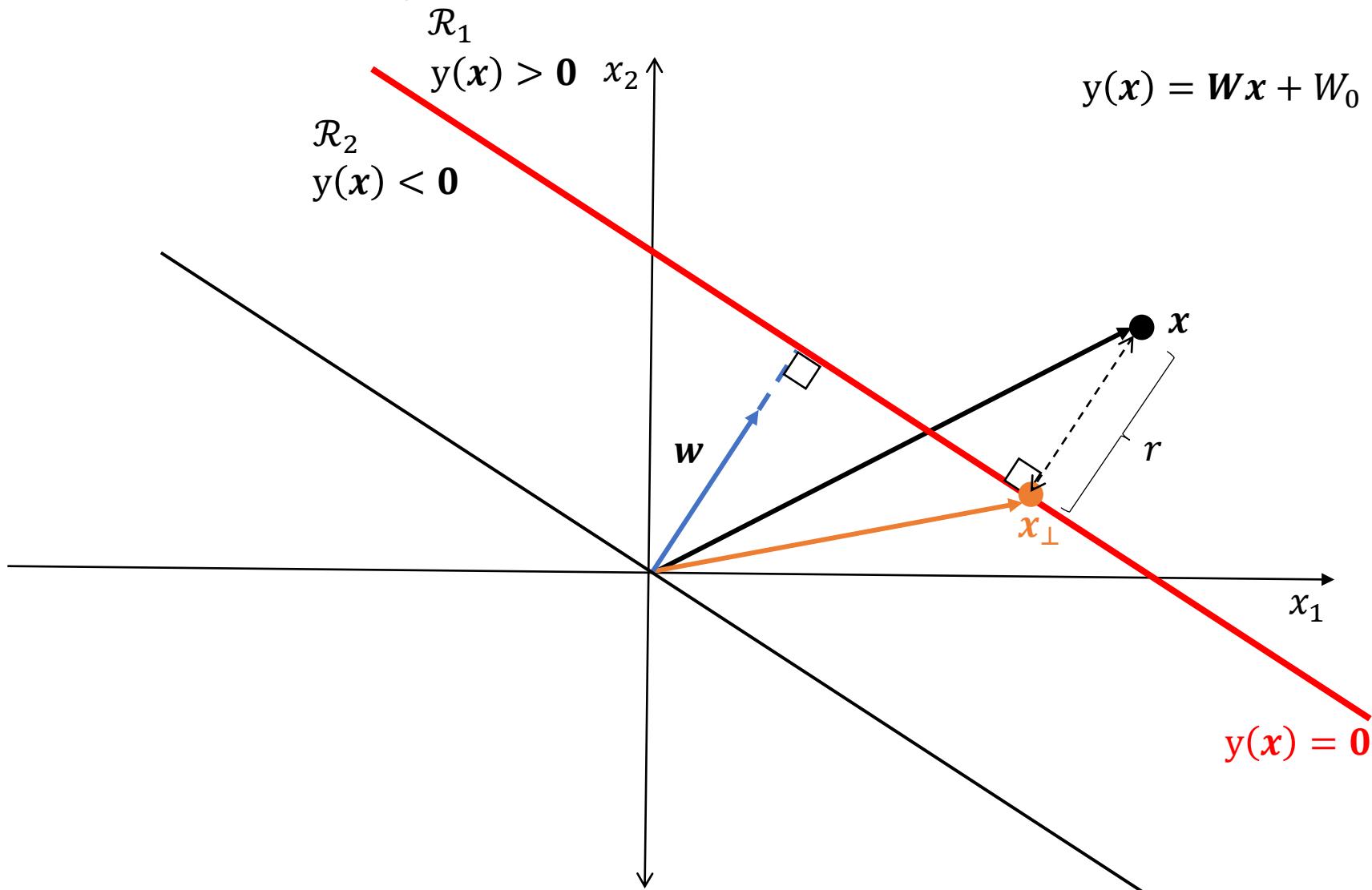
Análisis discriminante lineal: El problema de dos clases



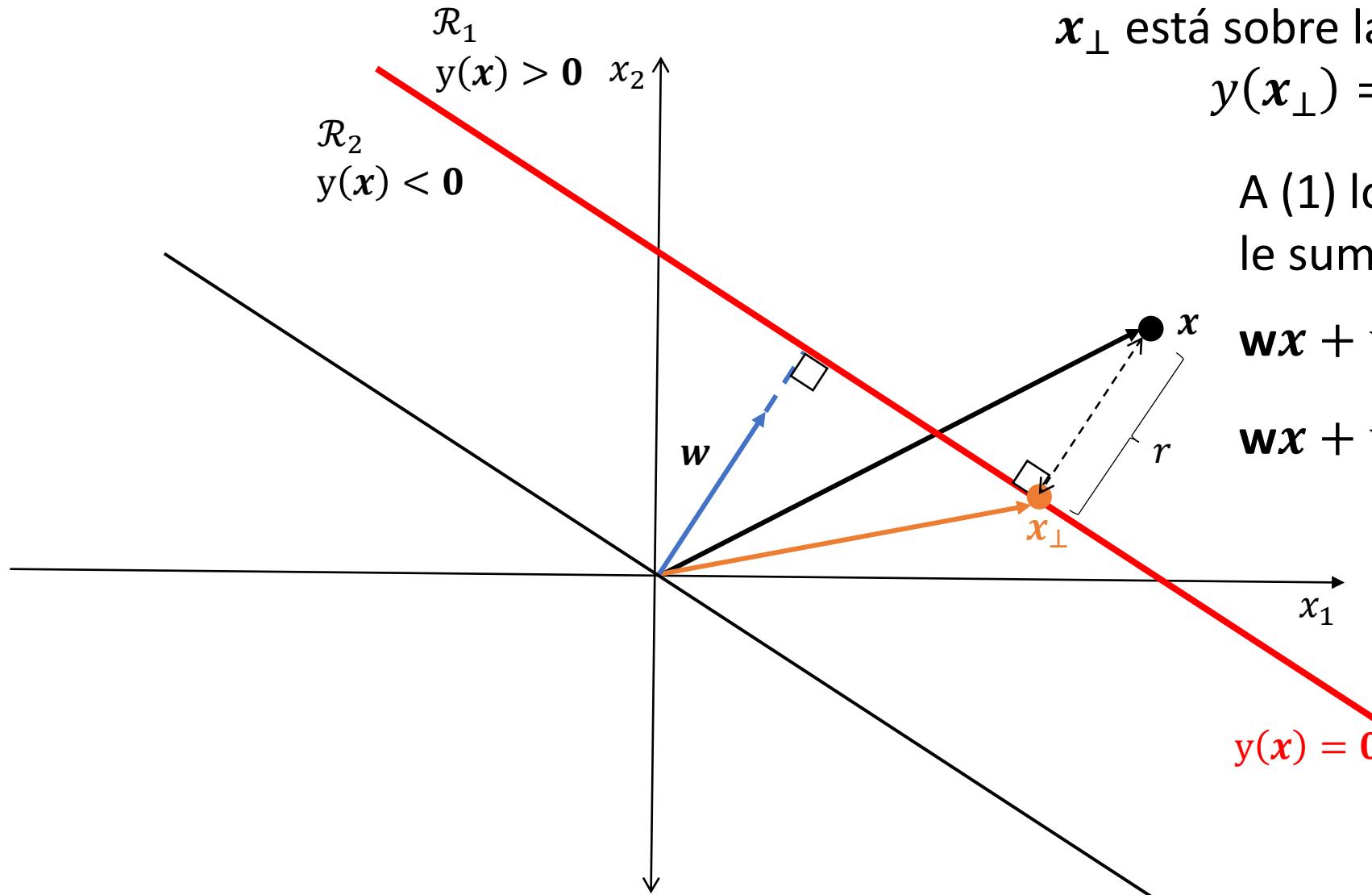
Análisis discriminante lineal: El problema de dos clases



¿Cómo lo plateamos?



¿Cómo lo plateamos?



Podemos plantear:

$$\mathbf{x} = \mathbf{x}_\perp + r \frac{\mathbf{w}}{|\mathbf{w}|} \quad (1)$$

\mathbf{x}_\perp está sobre la línea roja, por lo tanto:

$$y(\mathbf{x}_\perp) = \mathbf{w}\mathbf{x}_\perp + w_0 = 0$$

A (1) lo multiplicamos por \mathbf{w} y le sumamos w_0 :

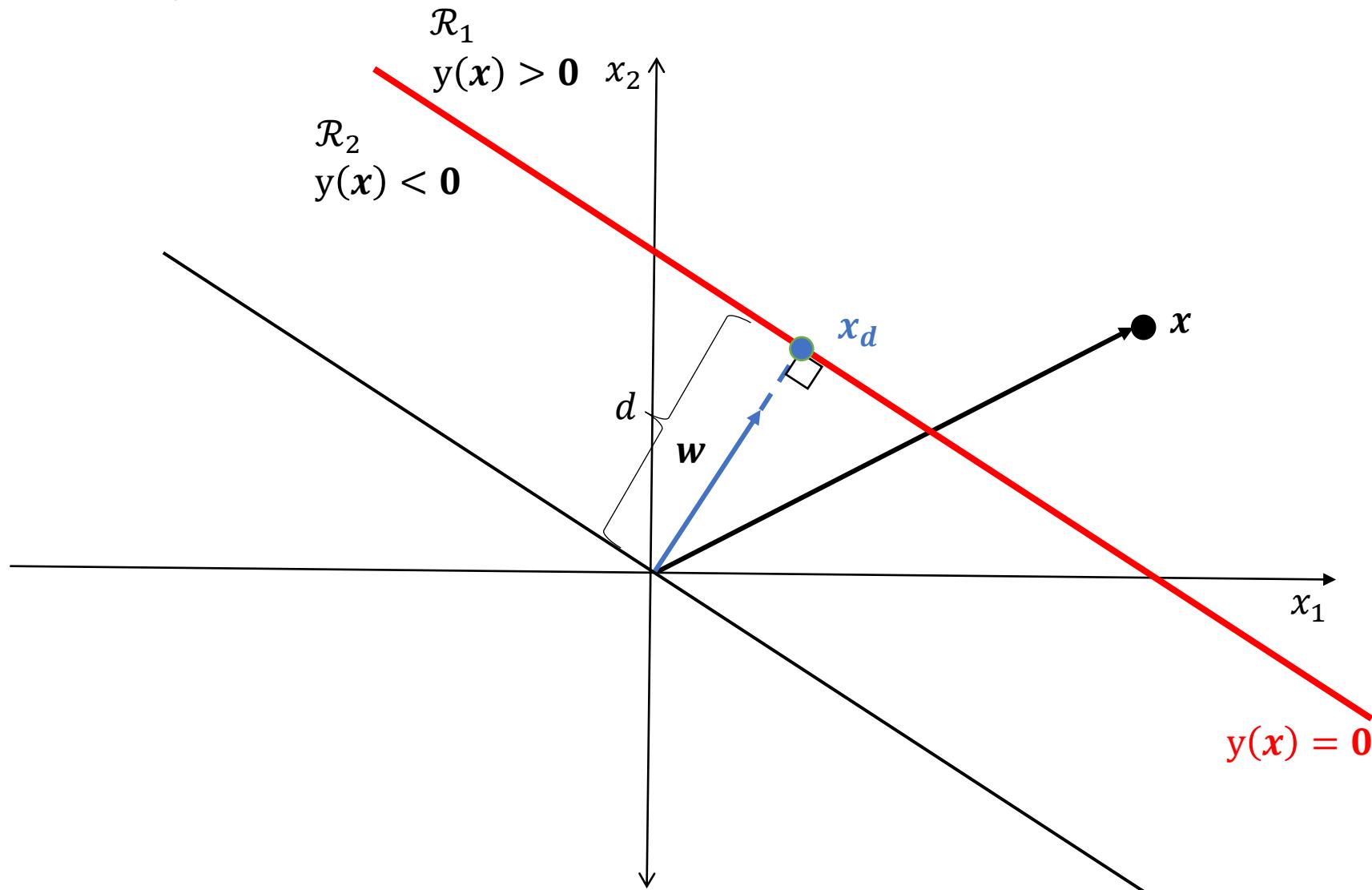
$$\mathbf{w}\mathbf{x} + w_0 = \mathbf{w}\mathbf{x}_\perp + r \frac{\mathbf{w}^T \mathbf{w}}{|\mathbf{w}|} + w_0$$

$$\mathbf{w}\mathbf{x} + w_0 = \mathbf{w}\mathbf{x}_\perp + r \frac{|\mathbf{w}| |\mathbf{w}|}{|\mathbf{w}|} + w_0$$

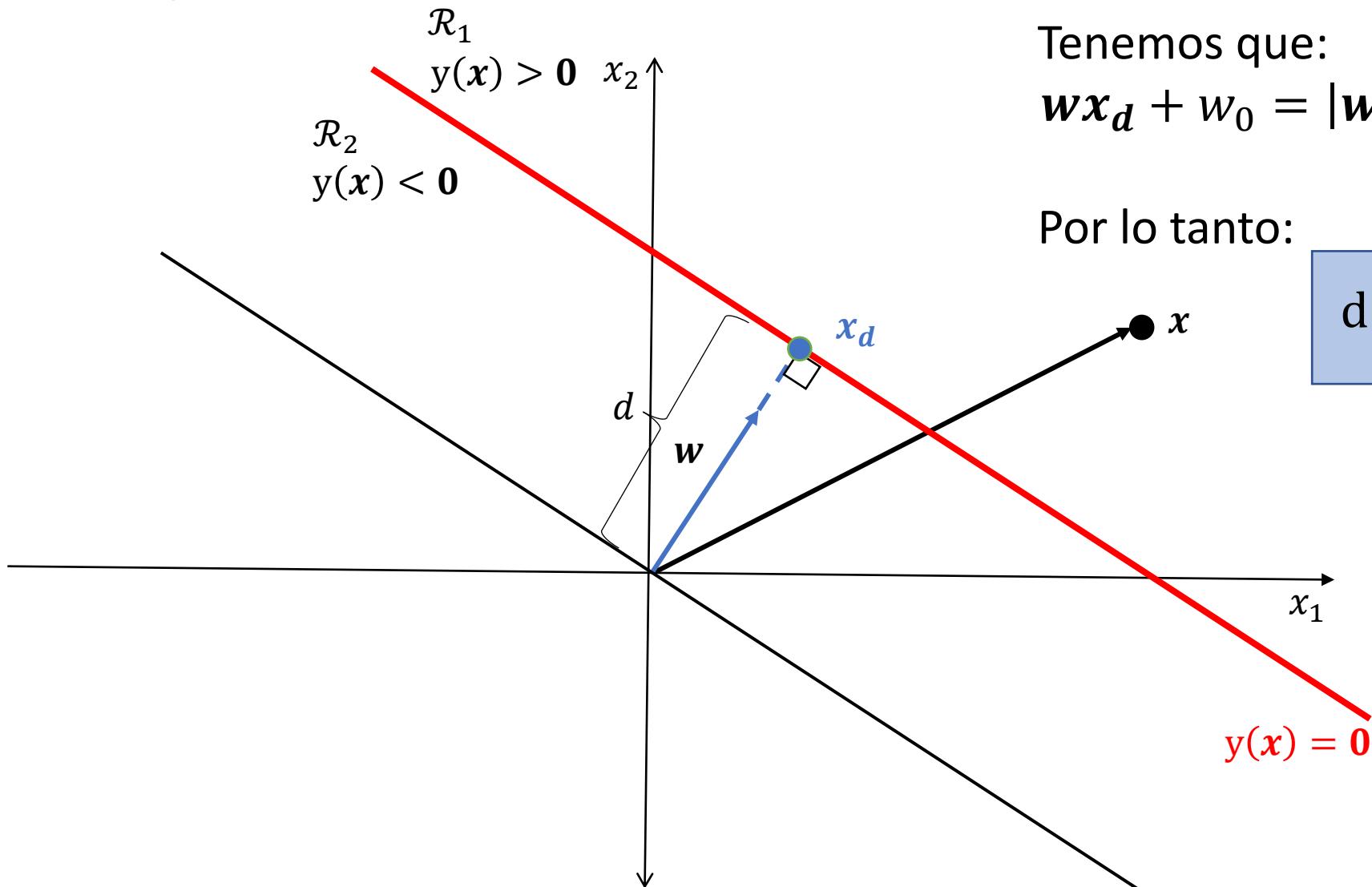
$$y(\mathbf{x}) = r |\mathbf{w}|$$

$$r = \frac{y(\mathbf{x})}{|\mathbf{w}|}$$

¿A qué distancia está la línea del origen?



¿A qué distancia está la línea del origen?



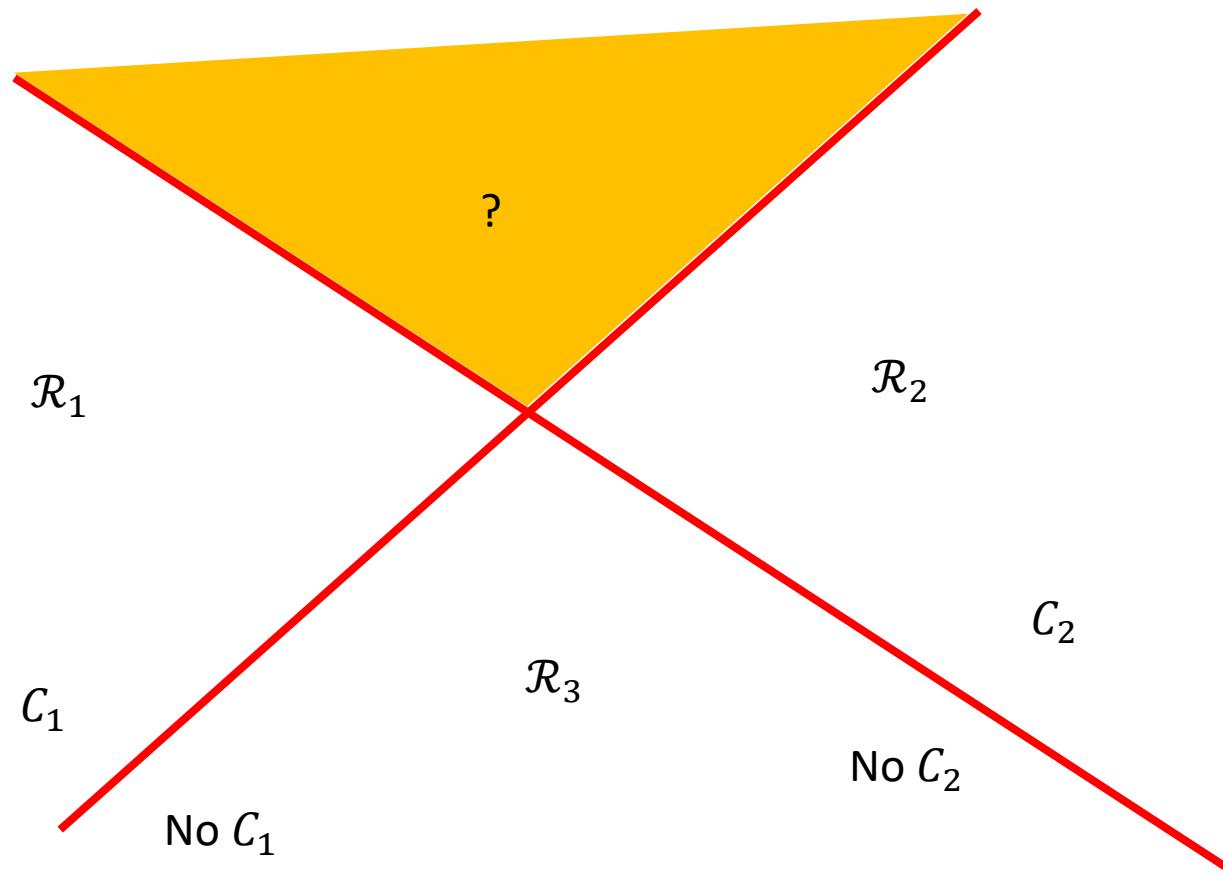
Tenemos que:

$$\mathbf{w}\mathbf{x}_d + w_0 = |\mathbf{w}||\mathbf{x}_d| \cos(\theta) + w_0 = 0$$

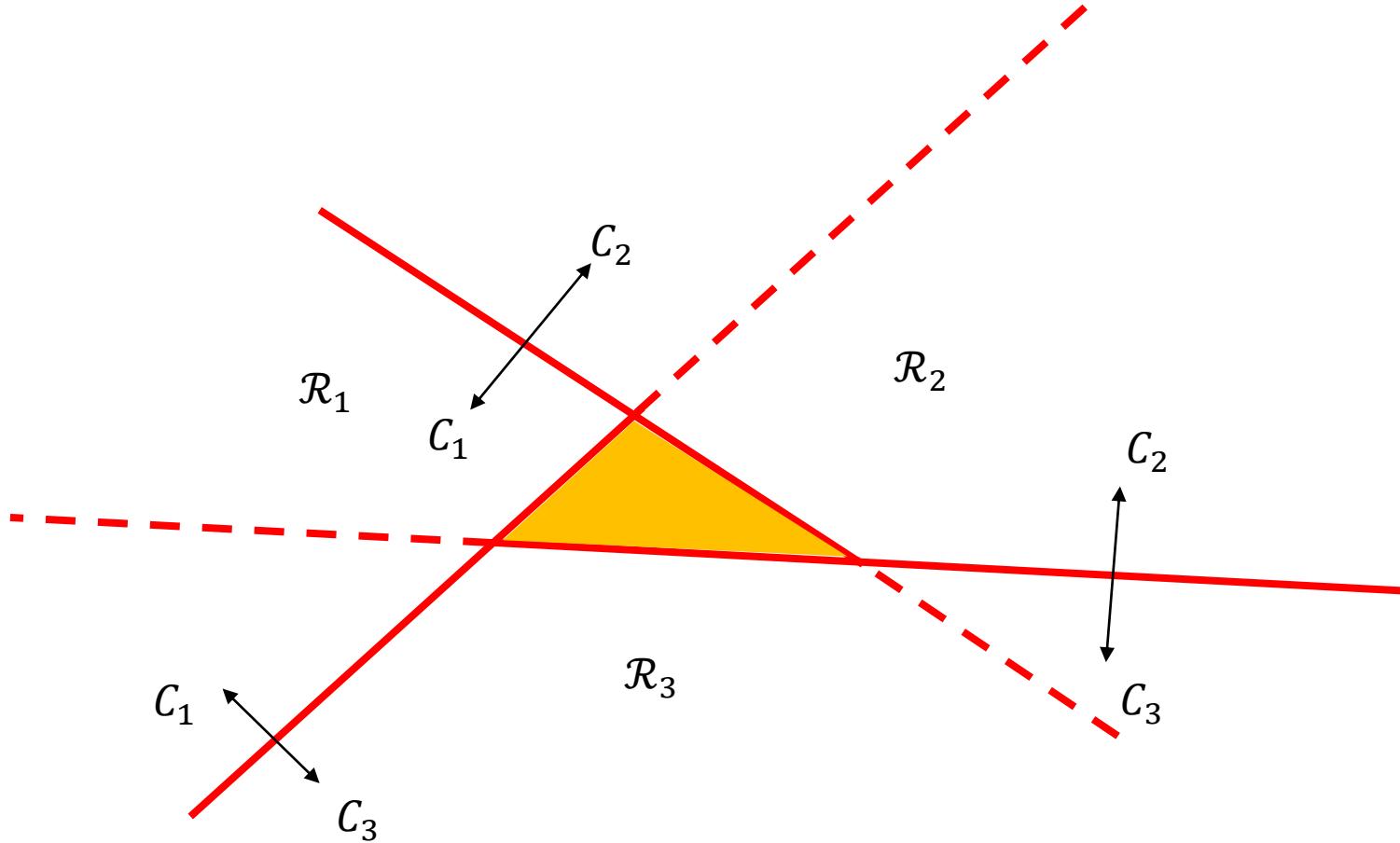
Por lo tanto:

$$d = -\frac{w_0}{|\mathbf{w}|}$$

¿Cómo lo aplicamos a varias regiones?



¿Cómo lo aplicamos a varias regiones?



La solución para aplicarlo a varias regiones

Considerar K clases discriminantes:

$$y_k(\mathbf{x}) = \mathbf{w}_k \mathbf{x} + w_{k0}$$

Y asignar el punto \mathbf{x} a la clase C_k si $y_k(\mathbf{x}) > y_j(\mathbf{x})$ para $j \neq k$.

Las fronteras entre dos clases C_k y C_j está dada donde $y_k(\mathbf{x}) = y_j(\mathbf{x})$
Y es un hiperplano de dimensión $D - 1$ definido por:

$$(\mathbf{w}_k - \mathbf{w}_j)^T \mathbf{x} + (w_{k0} - w_{j0}) = 0$$

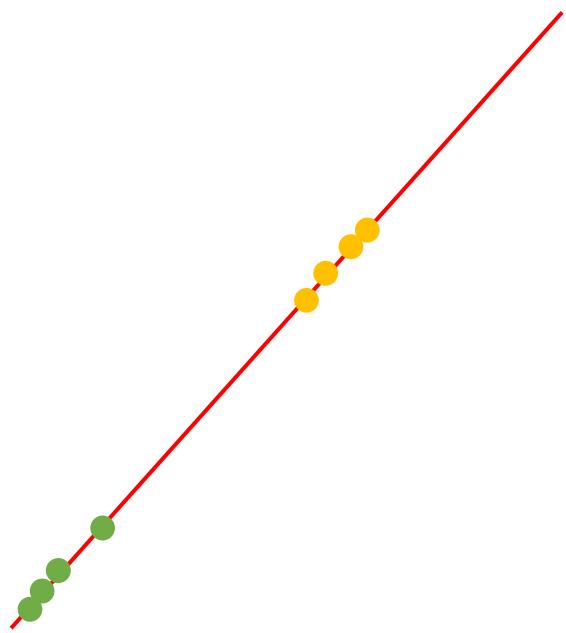


14
6

Actividad 2

Clasificación por medio de análisis de discriminante
lineal

¿Cómo resolverlo?: El problema de dos clases

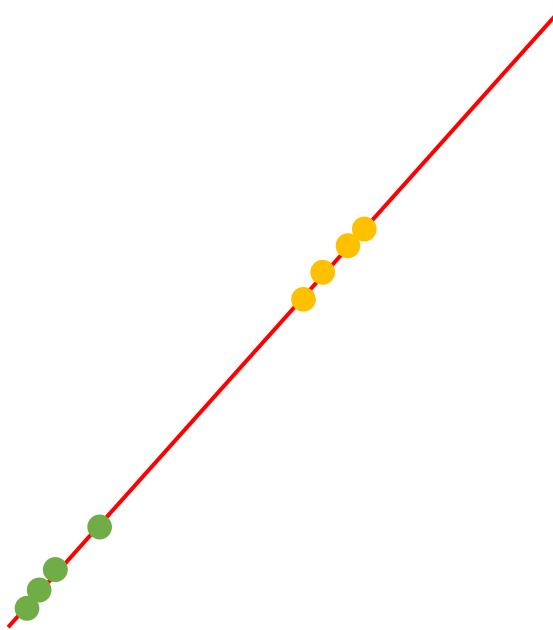


¿Cómo podemos decir que un vector w_a es mejor que otro w_b ?

Una idea sería medir la distancia entre las medias de las dos clases en la proyección 1D:

$$|\mu_1 - \mu_2|$$

¿Cómo resolverlo?: El problema de dos clases



Tenemos:

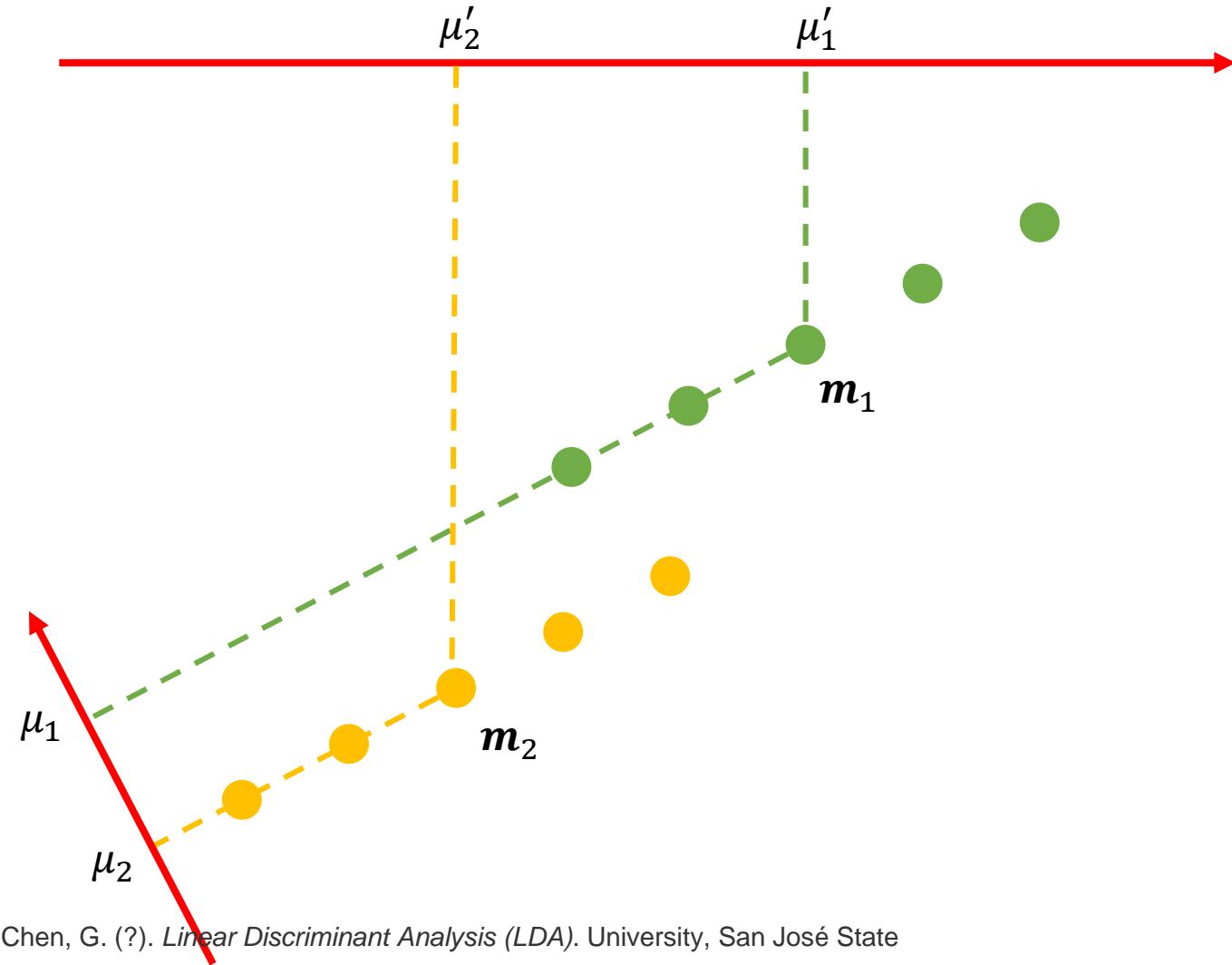
$$\begin{aligned}\mu_1 &= \frac{1}{n_1} \sum_{i \in C_1} a_i = \frac{1}{n_1} \sum_{i \in C_1} \mathbf{w} \mathbf{x}_i \\ &= V^T \frac{1}{n_1} \sum_{i \in C_1} \mathbf{x}_i = \mathbf{w} \mathbf{m}_1\end{aligned}$$

Similarmente:

$$\mu_2 = \mathbf{w} \mathbf{m}_2, \text{ con}$$

$$\mathbf{m}_2 = \frac{1}{n_2} \sum_{i \in C_2} \mathbf{x}_i$$

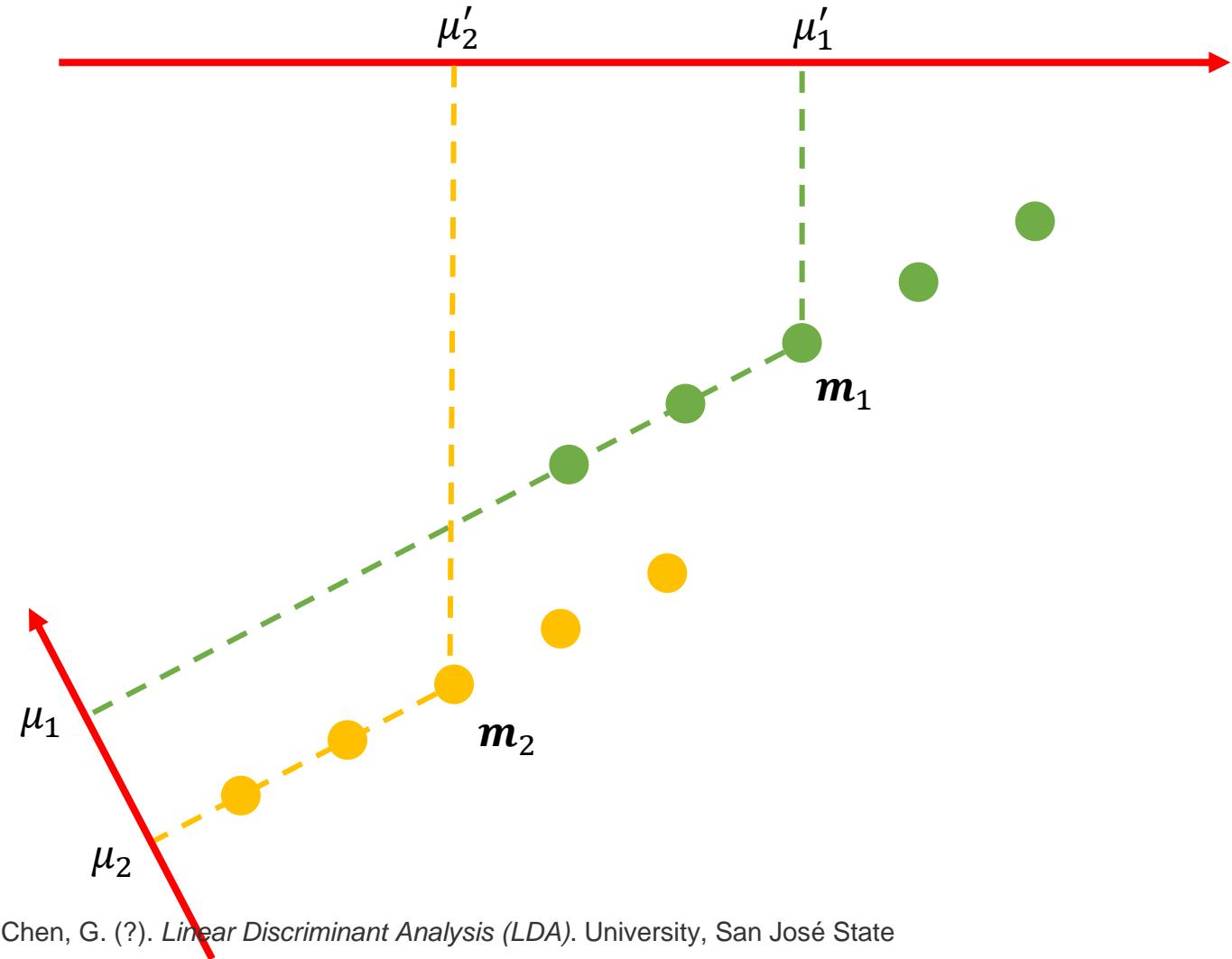
¿Cómo resolverlo?: El problema de dos clases



Ahora tenemos que resolver el siguiente problema:

$$\max_{v: \|v\|=1} |\mu_1 - \mu_2|$$

¿Cómo resolverlo?: El problema de dos clases



Debemos tener cuidado con las varianzas:

$$S_1^2 = \frac{1}{n_1} \sum_{i \in C_1} (a_i - \mu_1)^2$$

$$S_2^2 = \frac{1}{n_2} \sum_{i \in C_2} (a_i - \mu_2)^2$$

Idealmente queremos medias lejanas y varianzas pequeñas.

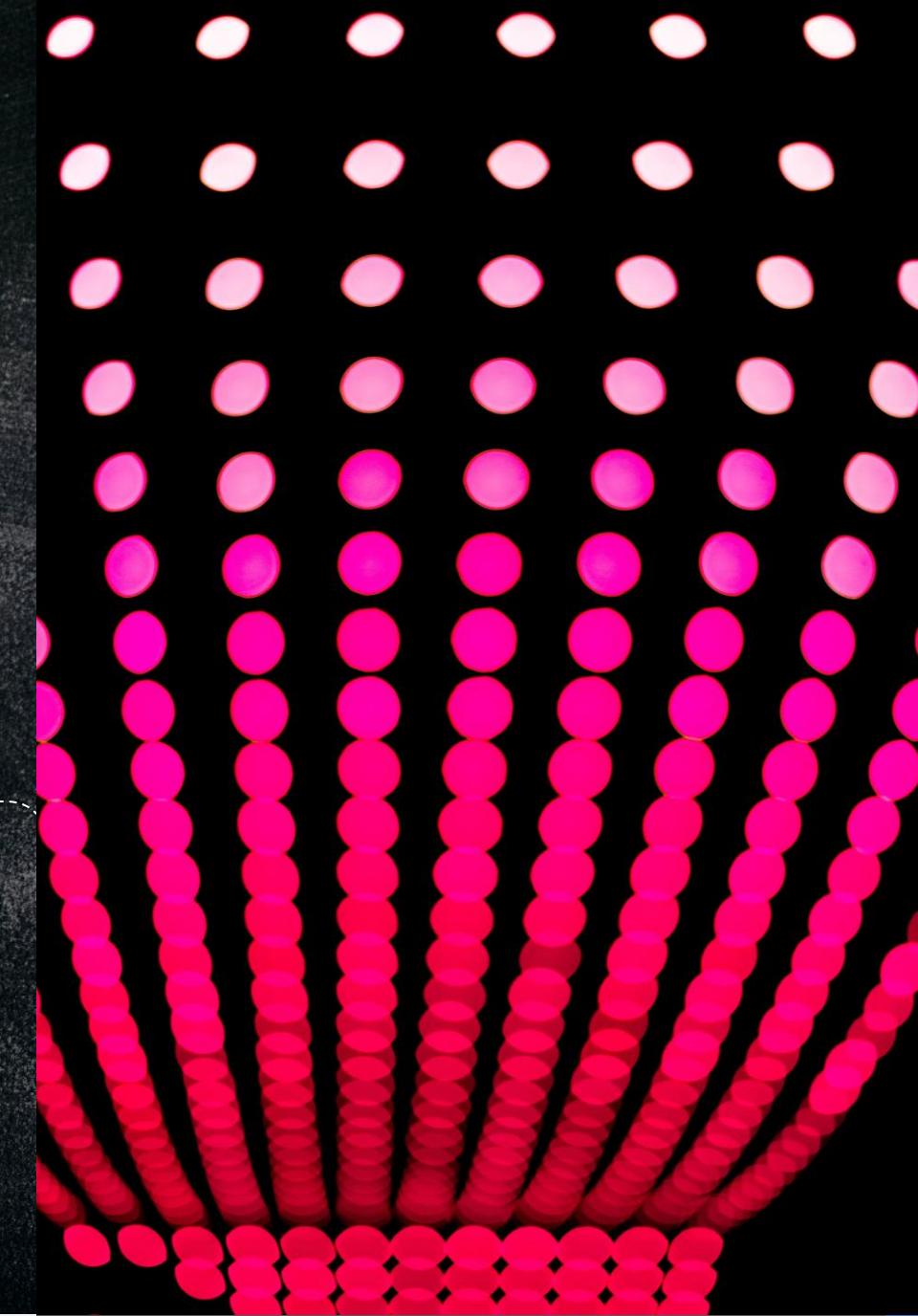
¿Cómo resolverlo?: El problema de dos clases

Pasamos de $\max_{v: \|v\|=1} |\mu_1 - \mu_2|$ a $\max_{v: \|v\|=1} \frac{(\mu_1 - \mu_2)^2}{S_1^2 + S_2^2}$

El valor óptimo es aquel con el v óptimo que:

- $(\mu_1 - \mu_2)^2$: es grande
- S_1^2, S_2^2 : ambos son pequeños

Regresión lineal



Regresión

El objetivo de la regresión es predecir el valor de una o más variables objetivo continuas dado el valor de un vector x de dimensión D correspondiente a variables de entrada.

Dado un conjunto de datos de entrenamiento con N observaciones $\{x_n\}$ donde $n = 1, \dots, N$ junto con los valores objetivo $\{t_n\}$ el objetivo es predecir un nuevo valor de t para un nuevo valor de x .

Regresión

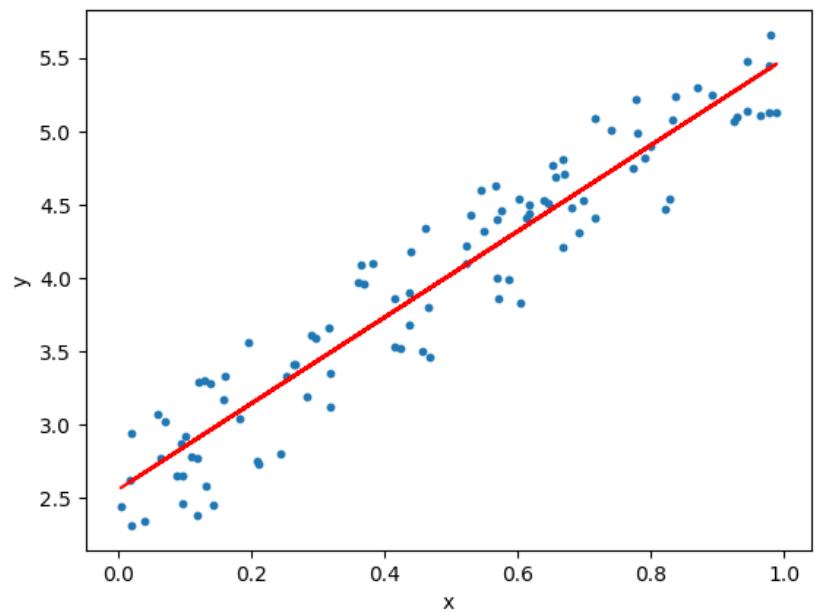
De manera sencilla, esto puede hacerse por la construcción directa de una función apropiada $y(x_i)$, cuyos valores generados para las nuevas entradas x constituyen la predicción de los valores correspondientes de t .

Regresión

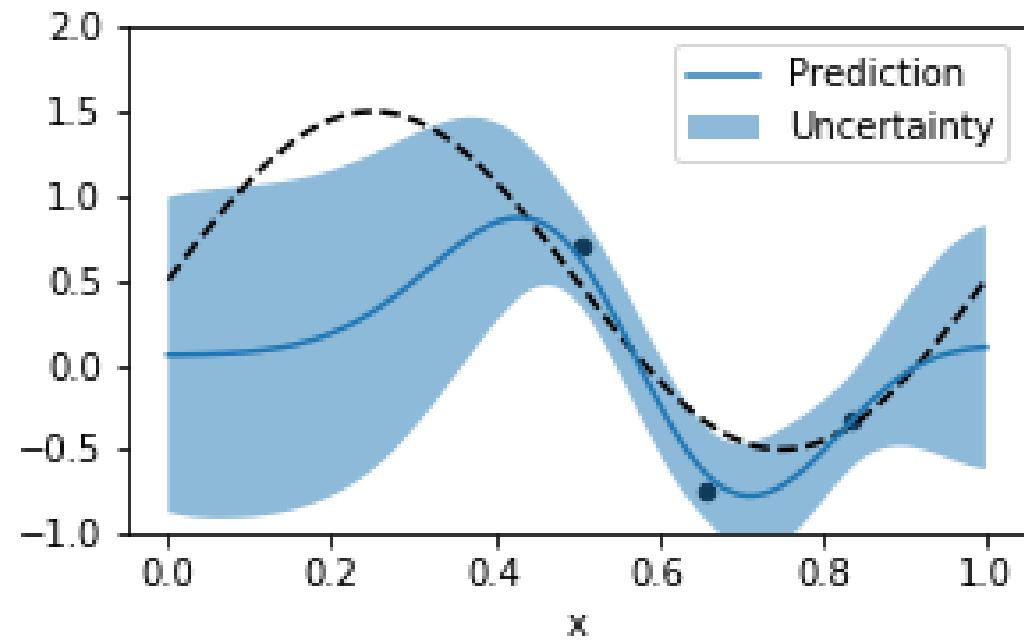
De manera más general, de un punto de vista probabilístico, el objetivo es modelar una distribución predictiva $P(t|x)$. Con esto

- expresamos la incertidumbre sobre el valor de t para cada valor de x ;
- se pueden hacer predicciones para cada nuevo valor de x que se minimice el valor esperado de una **función de perdida** (*loss function*) por ejemplo, la función de pérdida cuadrática (*Quadratic loss function*).

Regresión



$$f(x)$$



$$P(t|x)$$

Regresión lineal simple y múltiple

La regresión lineal modela la relación entre una variable continua y una o más variables independientes mediante el ajuste de una ecuación lineal.

- **Regresión lineal simple:** hay una variable independiente.
- **Regresión lineal múltiple:** hay más de una variable independiente.

A la variable modelada se le llama **variable dependiente** o **variable respuesta** (y), y a las variables independientes como **regresores**, **predictores** o **features** (x).

Regresión lineal, definición matemática

La regresión más sencilla es la regresión lineal de las variables de entrada

$$y(\mathbf{x}, \mathbf{W}) = w_0 + w_1 x_1 + \cdots + w_D x_D$$

$$\mathbf{x} = (x_1 \cdots x_D), \mathbf{W} = (w_0 \cdots w_D).$$

- x_i : predictor i
- w_0 : ordenada origen, cuando todos los predictores son 0.
- w_i : efecto promedio que tiene el incremento de una unidad del predictor x_i en la variable respuesta (coeficiente parcial de regresión)
- ε : error de la predicción, residuo entre el valor estimado y el observado.

Regresión lineal, definición matemática

	Variable 1	Variable 2	Variable 3	Variable 4		Variable m
Observación 1						
Observación 2						
Observación 3						
Observación 4						
Observación 5						
Observación n						

Ejemplo

	Variable 1	Variable 2	Variable 3	Variable 4	Variable m
Observación 1					
Observación 2					
Observación 3					
Observación 4					
Observación 5					
Observación n					

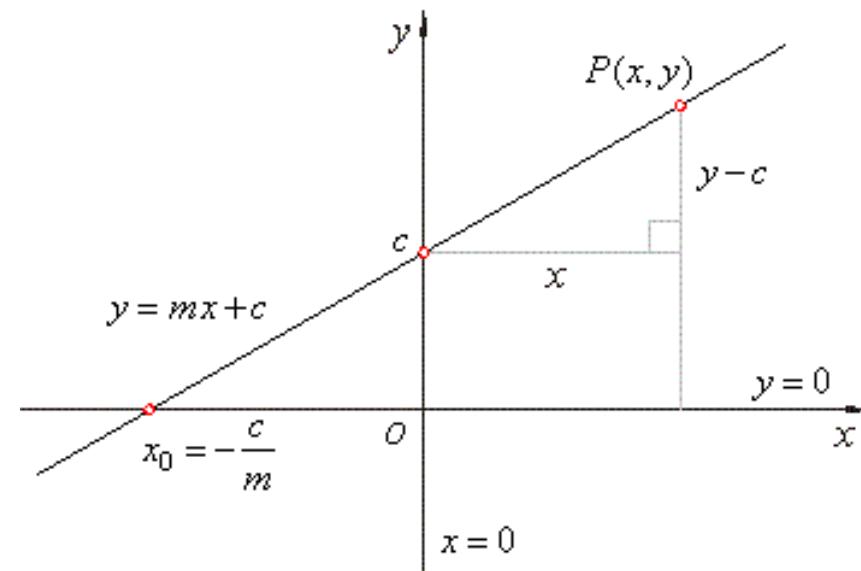
Algoritmo de
aprendizaje
automático

$$W = (w_0 \cdots w_D)$$

Variable m

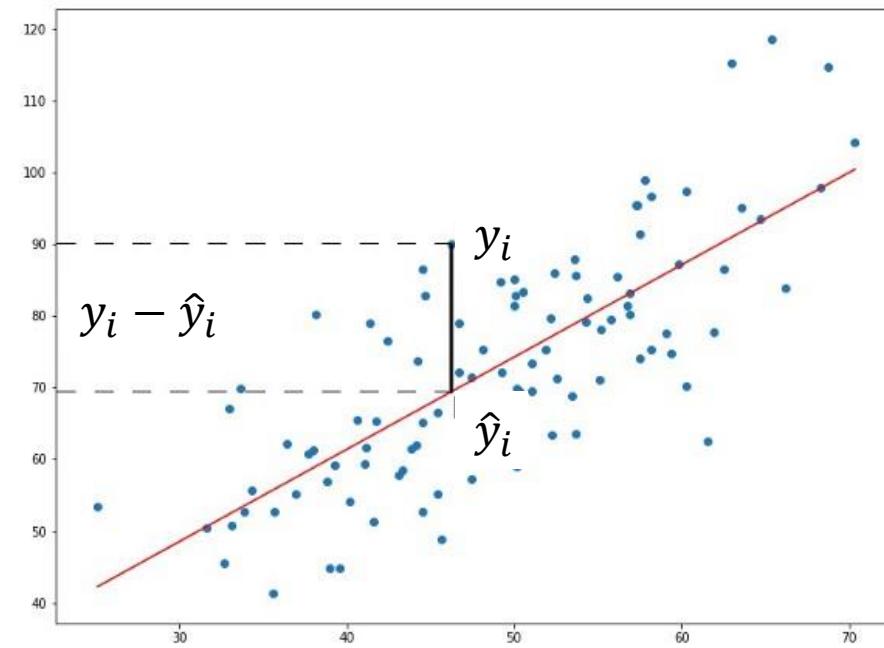
Regresión lineal, definición matemática

$$y = mx + c$$

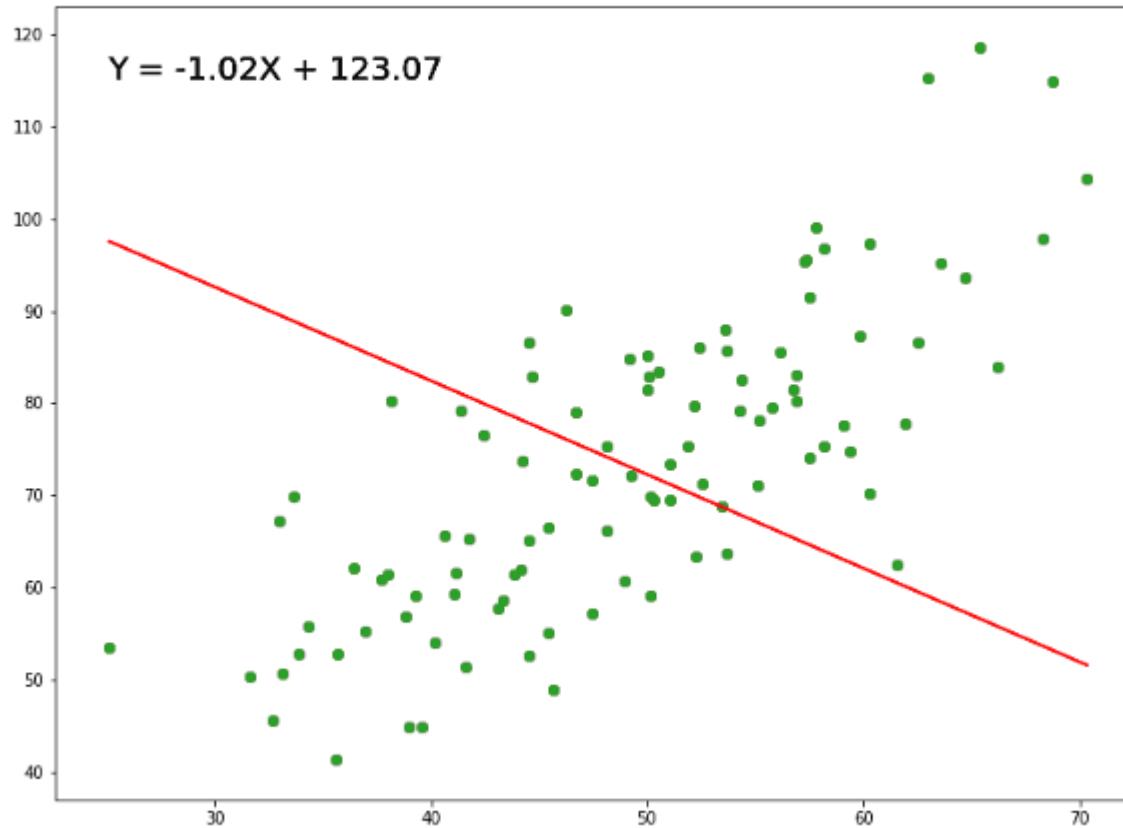


Regresión lineal, definición matemática

$$L(x) = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



Regresión lineal, definición matemática



Regresión lineal, definición matemática

A partir de

$$\hat{y}(x, \mathbf{W}) = w_0 + w_1 x_1 + \cdots + w_D x_D$$

Geralizamos para todos los datos

$$\mathbf{y} = \mathbf{X}^T \mathbf{W} + \boldsymbol{\varepsilon}, \quad \hat{\mathbf{y}} = \mathbf{X}^T \mathbf{W}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,D} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & \cdots & x_{N,D} \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} w_0 \\ \vdots \\ w_N \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

Regresión lineal, definición matemática

El modelo se puede generalizar considerando una combinación lineal de funciones no lineales de las variables de entrada, de la forma

$$\hat{y}(x, W) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(x)$$

Donde $\phi_j(x_j)$ son conocidas como **funciones base**. Ejemplo

$$\hat{y}(x, W) = w_0 + w_1 x_1 + w_2 \log(x_2)$$

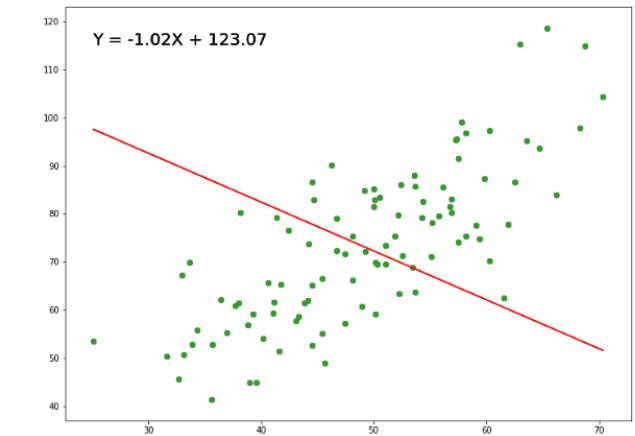
Regresión lineal, definición matemática

El problema es encontrar

$$\widehat{\mathbf{W}} = \arg \min_{\mathbf{W}} (\mathbf{y} - \mathbf{x}^T \mathbf{W})^2$$

o lo que es lo mismo minimizar $(\mathbf{y} - \widehat{\mathbf{y}})^2$

Llamado **minimización del error cuadrático**.



A esto se le conoce como maximizar la **verosimilitud** (*likelihood*), es decir, encontrar los coeficientes de regresión que dan lugar al modelo que con mayor probabilidad puede haber generado los datos observados.

Regresión lineal, definición matemática

Para iniciar la búsqueda de una solución lo primero que se debe hacer es **estandarizar los datos**: sustraer la media y dividir entre la desviación estándar de cada valor de las variables.

Enseguida resolver

$$\widehat{\mathbf{W}} = \arg \min_{\mathbf{W}} (\mathbf{y} - \mathbf{x}^T \mathbf{W})^2$$

Regresión lineal, definición matemática

Bondad del ajuste del modelo

Las métricas más utilizadas para medir la calidad del ajuste son:

- **Suma de cuadrados residuales** (*Residual Sum of Squares*, RSS)

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- **Error cuadrático medio** (*Root Mean Square Error*, RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$$

Regresión lineal, definición matemática

Bondad del ajuste del modelo

- el **error estándar de los residuos** (*Residual Standard Error*, RSE)

$$RSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p}}$$

donde p es el número de parámetros (p. ej. 2 si se estima w_0 y w_1 para modelos lineales simples).

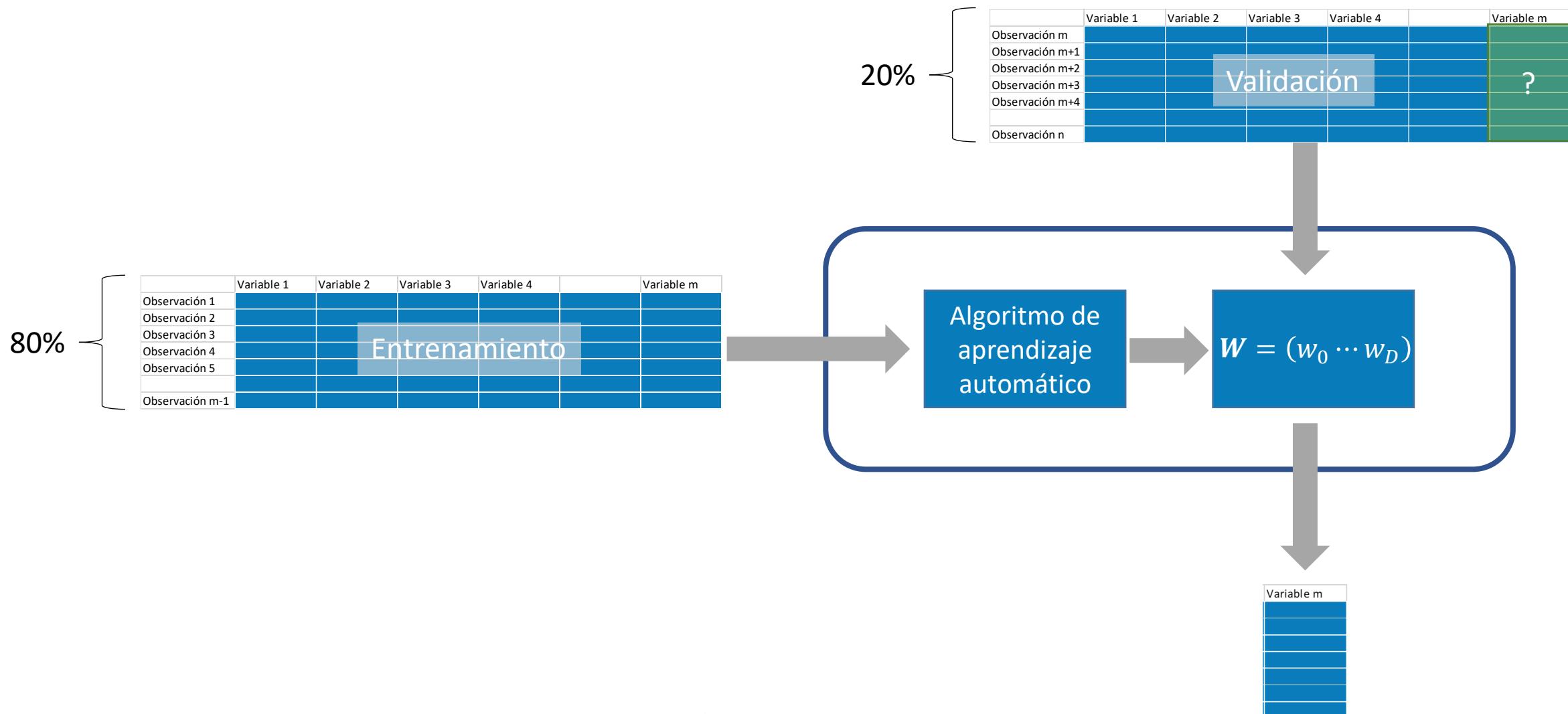
Regresión lineal, definición matemática

Bondad del ajuste del modelo

- el **coeficiente de determinación** R^2 describe la proporción de varianza de la variable respuesta explicada por el modelo y relativa a la varianza total

$$1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2}$$

Validación del modelo



Condiciones para la regresión lineal

Se deben cumplir las siguientes condiciones, aún que no siempre es posible demostrar todas ellas

- Relación lineal entre los predictores numéricos y la variable respuesta.
- No colinealidad o multicolinealidad entre predictores.
- Distribución normal de la variable respuesta.
- Varianza constante de la variable respuesta (homocedasticidad).
- No autocorrelación (Independencia).
- Valores atípicos, con alto *leverage* o influyentes.
- Tamaño de la muestra (que el número de observaciones sea como mínimo entre 10 y 20 veces el número de predictores del modelo.)
- Parsimonia (simplicidad, menor número de predictores)

Ejemplo 03: Regresión lineal simple 1

Ejercicio 01: Regresión lineal simple

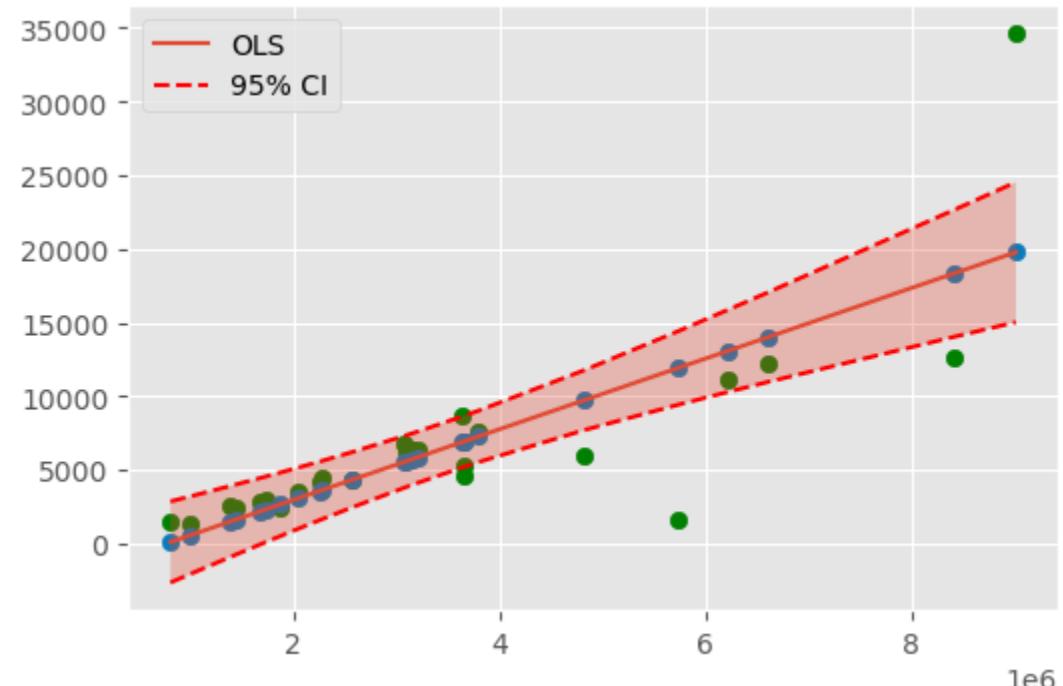


Ejemplo 04: Regresión lineal simple 2

Valores atípicos (*outliers*)

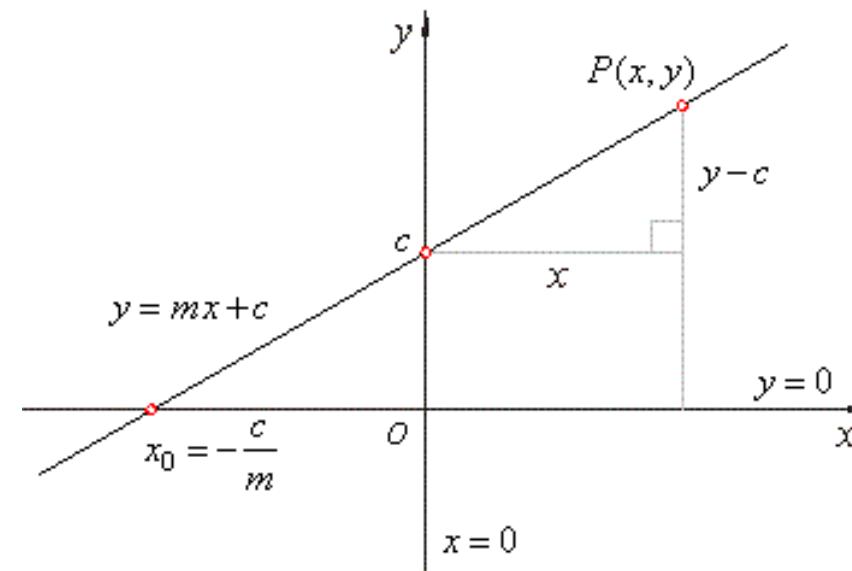
¿Hay algún valor influyente que afecte la regresión?

- Analizar muy bien si se deben eliminar de acuerdo a lo que se pretende hacer con el modelo.
- Para la predicción muchas veces es mejor no tener valores atípicos.
- Lo mejor es realizar un modelo con los valores atípicos y otro sin ellos para poder realizar comparaciones.



Realizamos regresión lineal simple

$$y = mx + c$$



Ahora realicemos regresión lineal múltiple

$$y(\mathbf{X}, \mathbf{W}) = w_0 + w_1 x_1 + \cdots + w_D x_D$$

$$\mathbf{X} = (x_1 \cdots x_D), \mathbf{W} = (w_0 \cdots w_D).$$

- x_i : predictor i
- w_0 : ordenada origen, cuando todos los predictores son 0.
- w_i : efecto promedio que tiene el incremento de una unidad del predictor x_i en la variable respuesta (coeficiente parcial de regresión)
- ε : error de la predicción, residuo entre el valor estimado y el observado.

Ejemplo 05: Regresión lineal múltiple

Ejercicio 02: Regresión lineal múltiple



¿Podemos descartar algún predictor?

Significancia de los predictores

Para cada uno de los coeficientes de regresión w_i se puede calcular su significancia llamado ***p-value*** y su intervalo de confianza. El *p-value* es la evidencia **contra** una hipótesis nula.

- **Hipótesis nula, H_0** : el predictor x_i no contribuye al modelo ($w_i = 0$), en presencia del resto de predictores, esto es no existe relación lineal entre la variable respuesta y y x_i por lo que la pendiente del modelo es cero $w_i = 0$.
- **Hipótesis alternativa, H_a** : el predictor x_i sí contribuye al modelo ($w_i \neq 0$), en presencia del resto de predictores, es decir sí existe relación lineal entre la variable respuesta y y x_i por lo que la pendiente del modelo es diferente de cero $w_i \neq 0$

Significancia de los predictores

Para cada uno de los coeficientes de regresión w_i se puede calcular su significancia llamado ***p-value*** y su intervalo de confianza. El *p-value* es la evidencia **contra** una hipótesis nula.

- **Hipótesis nula, H_0** : el predictor x_i no contribuye al modelo ($w_i = 0$), en presencia del resto de predictores, esto es no existe relación lineal entre la variable respuesta y y x_i por lo que la pendiente del modelo es cero $w_i = 0$.
- **Hipótesis alternativa, H_a** : el predictor x_i sí contribuye al modelo ($w_i \neq 0$), en presencia del resto de predictores, es decir sí existe relación lineal entre la variable respuesta y y x_i por lo que la pendiente del modelo es diferente de cero $w_i \neq 0$.

Significancia de los predictores

Entre más grande el valor del *p-value* más fuerte es la evidencia de que debe rechazarse la hipótesis nula. Se expresa en decimales y lo mejor es pensarla en porcentaje. Por ejemplo si tenemos que el p-value = 0.037 significa que hay una probabilidad del 3.7% de que el resultado sea aleatorio.

Así que, entre más bajo el *p-value* más significado (importancia) tienen los resultados.

Significancia de los predictores

p-value vs nivel alfa

$$\text{nivel alfa} = 100\% - \text{nivel de confianza}$$

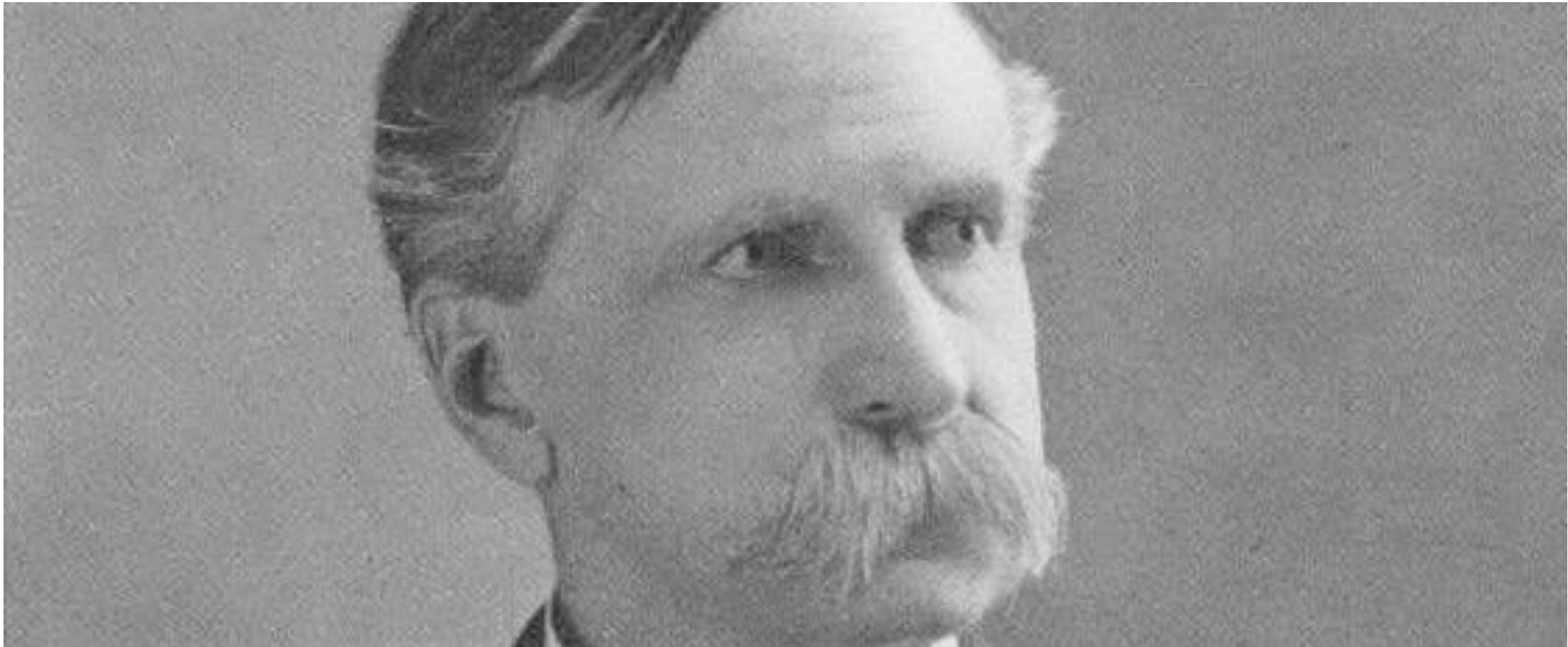
De esta forma:

Un **p-value** \leq **nivel alfa** : rechazar la hipótesis nulla.

Un **p-value** $>$ **nivel alfa** : aceptamos la hipótesis nula.

Ejemplo 06: Análisis de regresión lineal múltiple 1

Figures don't lie, but liars figure.
Atribuido a Carroll Davidson Wright



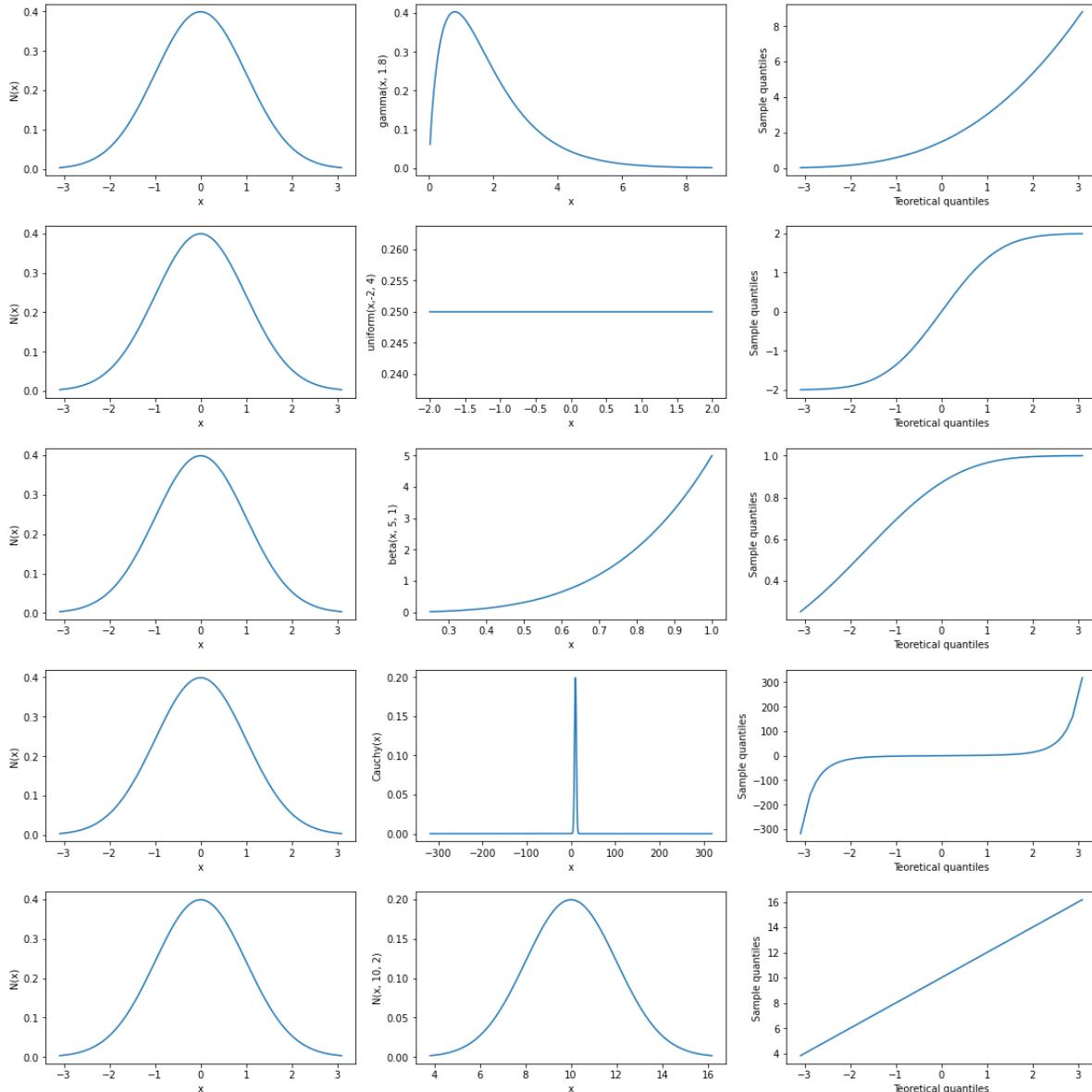
Ejemplo 07: Análisis de regresión lineal múltiple 2

Gráfica Q-Q (*Q-Q plot*)

¿Mi conjunto de datos representa una distribución normal?

Si es así, el resultado de la gráfica Q-Q debe ser una línea recta.

La evaluación de la normalidad es importante ya que muchos de los procesos de inferencia suponen que las muestras vienen de una población con distribución normal.



Gráfica Q-Q (Q-Q *plot*)

Mis datos podrían no seguir una distribución normal por

- Exceso de inclinación, asimetría.
- Curtosis, colas muy largas
- Distribución bimodal

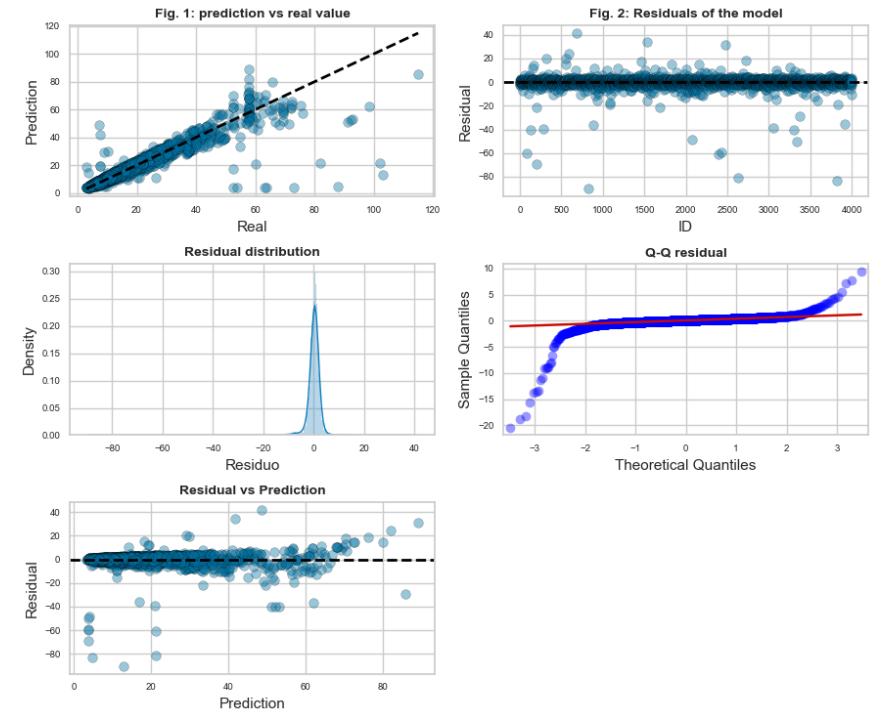
Las herramientas pueden ser

- Histogramas
- Graficas “Steam and leaf”
- Diagramas de caja (Box plots)
- Grafica P-P (P-P plot)
- Gráfica Q-Q (Q-Q plot)

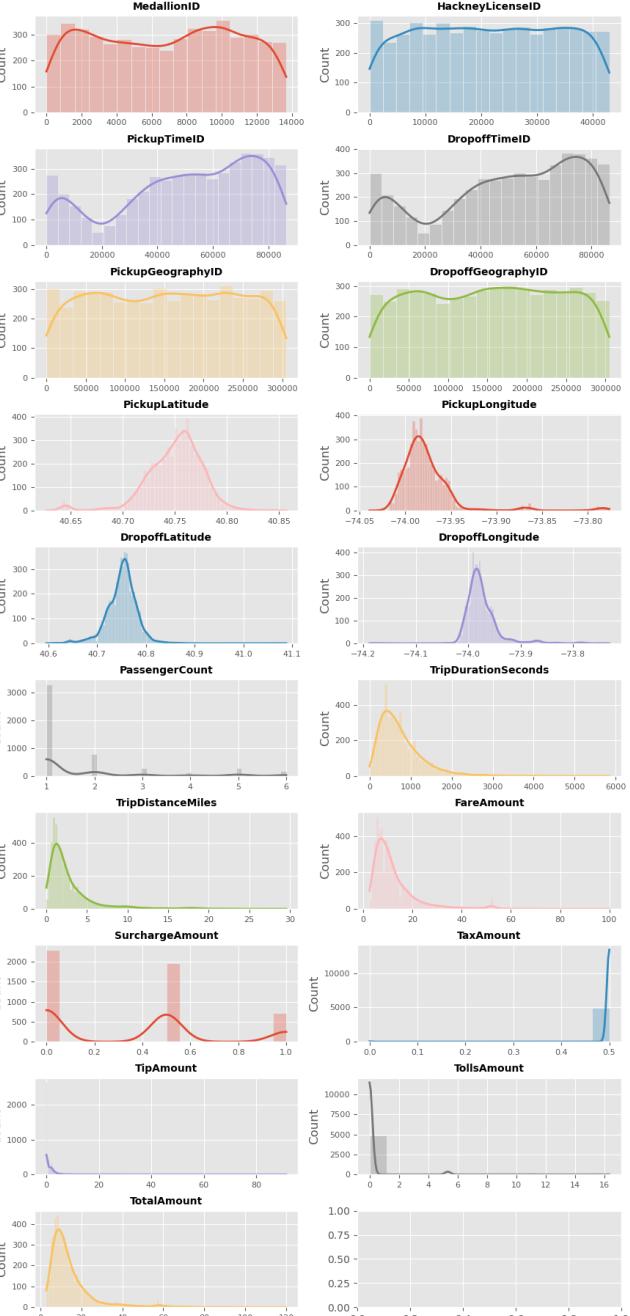
Tener en mente

- Si los datos violan una suposición se podrían obtener resultados que no son muy útiles.
- Es recomendable ver los datos gráficamente antes de iniciar los análisis.
- Siempre buscar entender lo que sucede.

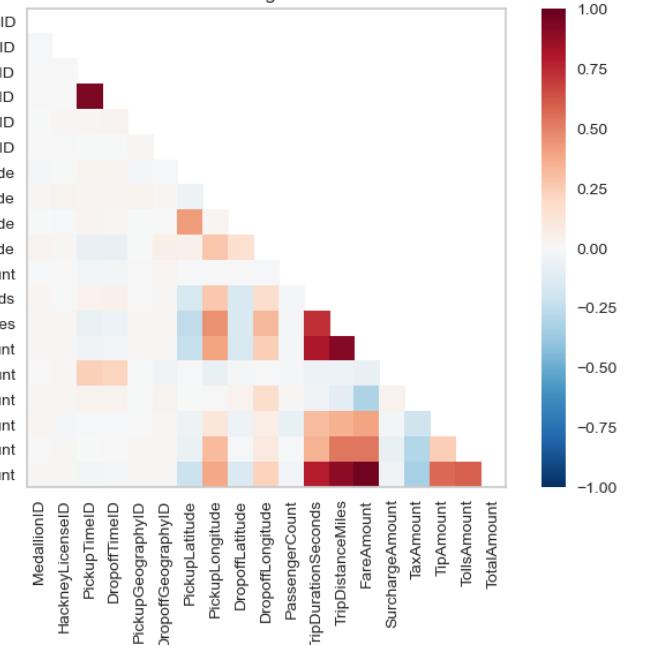
Residual diagnosis



Distribution of the variables



Pearson Ranking of 19 Features



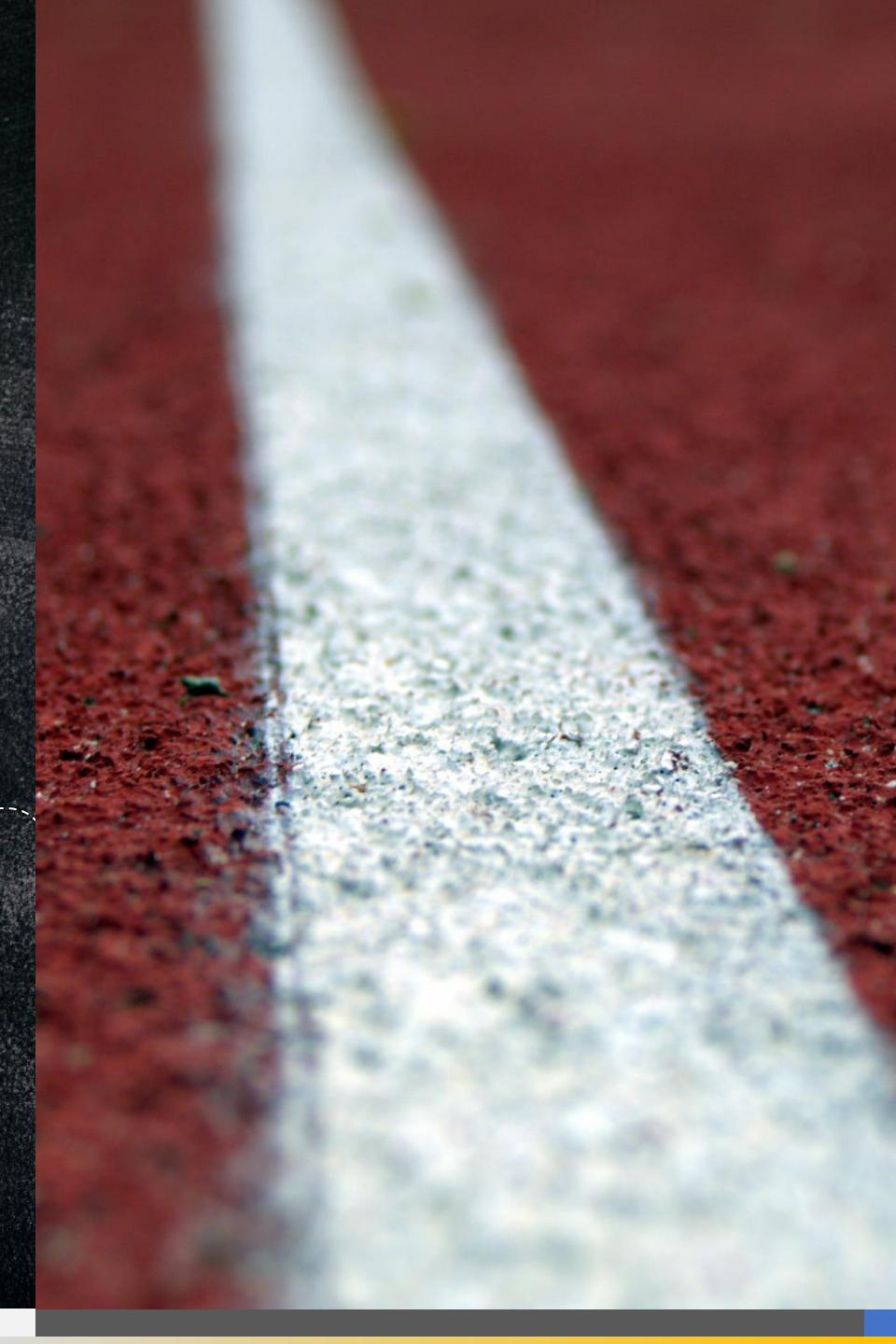
OLS Regression Results

Dep. Variable:	y	R-squared:	0.843			
Model:	OLS	Adj. R-squared:	0.842			
Method:	Least Squares	F-statistic:	5343.			
Date:	Thu, 22 Jul 2021	Prob (F-statistic):	0.00			
Time:	20:23:06	Log-Likelihood:	-11620.			
No. Observations:	4000	AIC:	2.325e+04			
Df Residuals:	3995	BIC:	2.328e+04			
Df Model:	4					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	3.7827	0.205	18.409	0.000	3.380	4.186
x1	0.0054	0.000	29.359	0.000	0.005	0.006
x2	2.4196	0.032	75.036	0.000	2.356	2.483
x3	-0.1626	0.051	-3.217	0.001	-0.262	-0.064
x4	4.212e-07	2.87e-06	0.147	0.883	-5.2e-06	6.04e-06
Omnibus:	6200.792	Durbin-Watson:	1.977			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4484808.811			
Skew:	9.582	Prob(JB):	0.00			
Kurtosis:	165.916	Cond. No.	1.65e+05			

Notes:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
- [2] The condition number is large, 1.65e+05. This might indicate that there are strong multicollinearity or other numerical problems.

1.4 Regresión Logística.



Regresión

El objetivo de la regresión es predecir el valor de una o más variables objetivo continuas dado el valor de un vector X de dimensión D correspondiente a variables de entrada.

Dado un conjunto de datos de entrenamiento con N observaciones $\{X_n\}$ donde $n = 1, \dots, N$ junto con los valores objetivo $\{t_n\}$ el objetivo es predecir un nuevo valor de t para un nuevo valor de X .

Regresión

De manera sencilla, esto puede hacerse por la construcción directa de una función apropiada $y(X)$, cuyos valores generados para las nuevas entradas X constituyen la predicción de los valores correspondientes de t .

Regresión

De manera más general, de un punto de vista probabilístico, el objetivo es modelar una distribución predictiva $P(t|X)$. Con esto

- expresamos la incertidumbre sobre el valor de t para cada valor de X ;
- se pueden hacer predicciones para cada nuevo valor de X que se minimice el valor esperado de una **función de perdida** (*loss function*) por ejemplo, la función de pérdida cuadrática (*Quadratic loss function*).

Regresión logística

La regresión logística es una regresión que trata de modelar la probabilidad de una **variable cualitativa binaria** en función de los regresores.

La principal aplicación de la regresión logística es la **clasificación binaria**.

Regresión logística

Variables cuantitativas vs cualitativas

- **Cuantitativas:** otorgan como resultado un valor numérico.
- **Cualitativas:** describe las cualidades, circunstancias o características de un objeto o persona, el uso de números no es necesario.



Regresión lineal, definición matemática

La regresión más sencilla es la regresión lineal de las variables de entrada

$$y(x, \mathbf{W}) = w_0 + w_1 x_1 + \cdots + w_D x_D$$

$$\mathbf{x} = (x_1 \cdots x_D), \mathbf{W} = (w_0 \cdots w_D).$$

- x_i : predictor i
- w_0 : ordenada origen, cuando todos los predictores son 0.
- w_i : efecto promedio que tiene el incremento de una unidad del predictor x_i en la variable respuesta (coeficiente parcial de regresión)
- ε : error de la predicción, residuo entre el valor estimado y el observado.

Ejemplo 08: Regresión lineal simple con variable cualitativa

Regresión logística

El término "lineal" en "la regresión lineal" se refiere a que **los parámetros se incorporan en la ecuación de forma lineal**, no a que necesariamente la relación entre cada predictor y la variable respuesta tenga que seguir un patrón de recta.

Dos problemas si usamos la regresión lineal para nuestro nuevo problema:

- Se pueden predecir valores distintos de 0 y 1 y esto **no** es una respuesta binaria.
- No se pueden calcular probabilidades de pertenencia a cada clase, podrían obtenerse valores fuera de intervalo [0, 1].

Regresión logística

True value vs predicted: 1 0.60

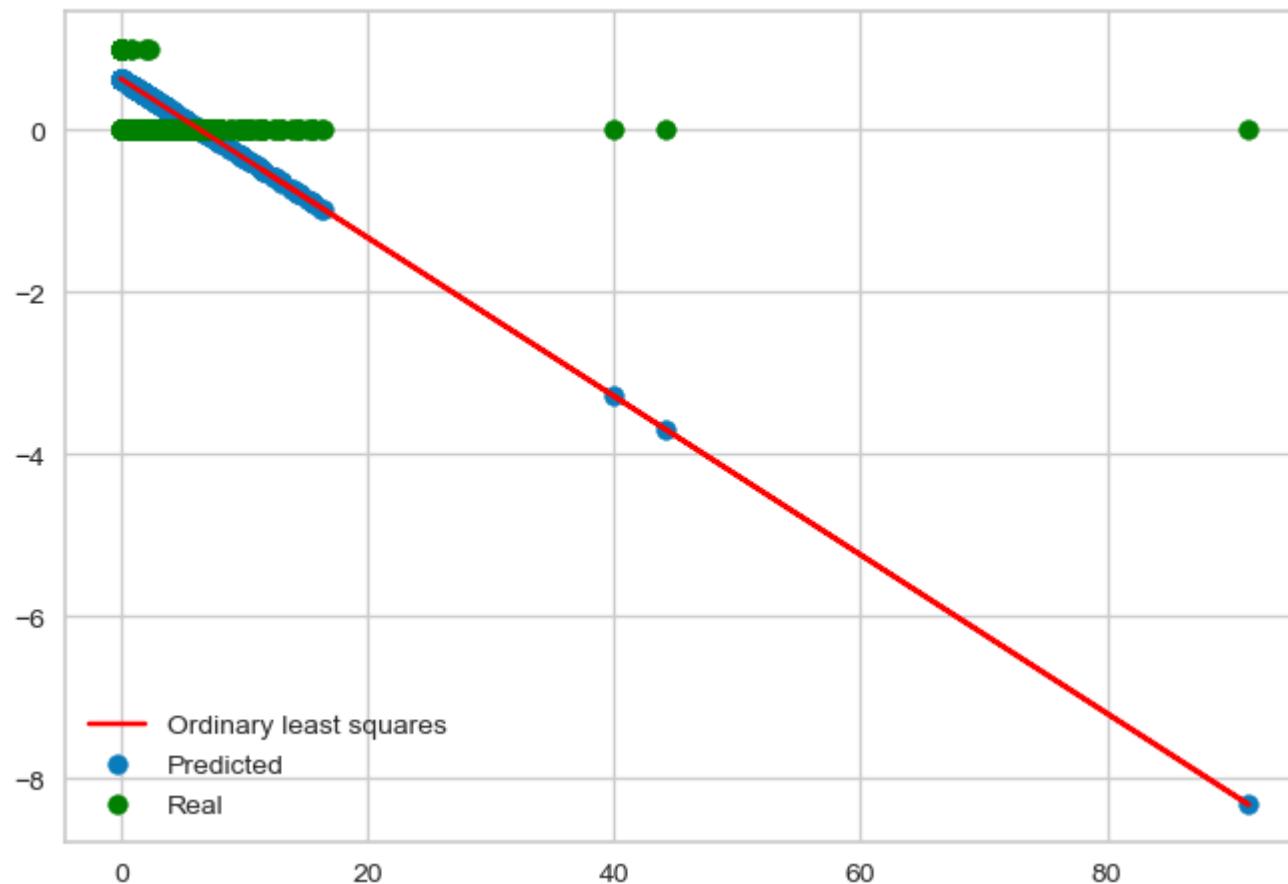
True value vs predicted: 0 0.41

True value vs predicted: 0 -0.10

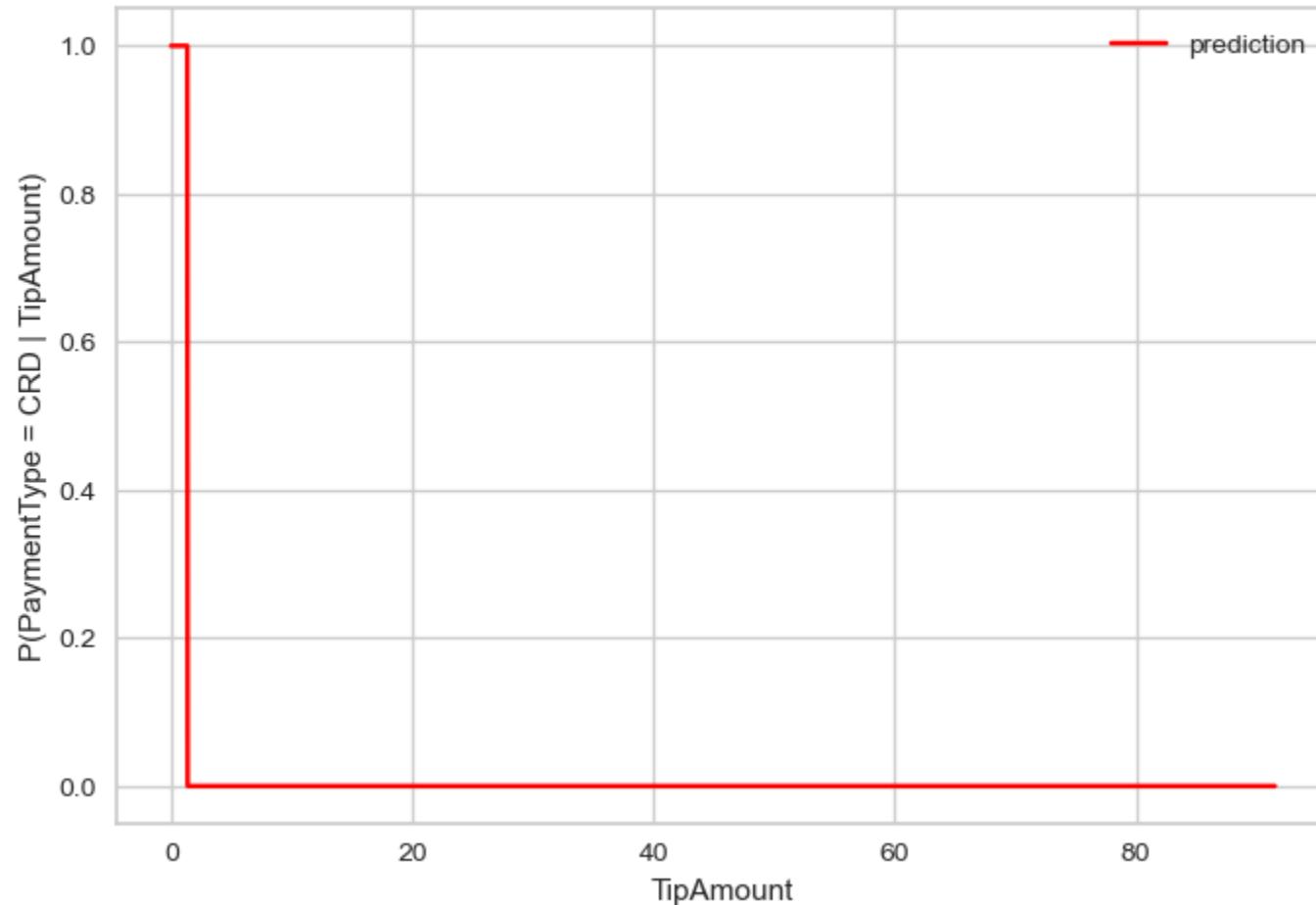
True value vs predicted: 1 0.60

True value vs predicted: 1 0.60

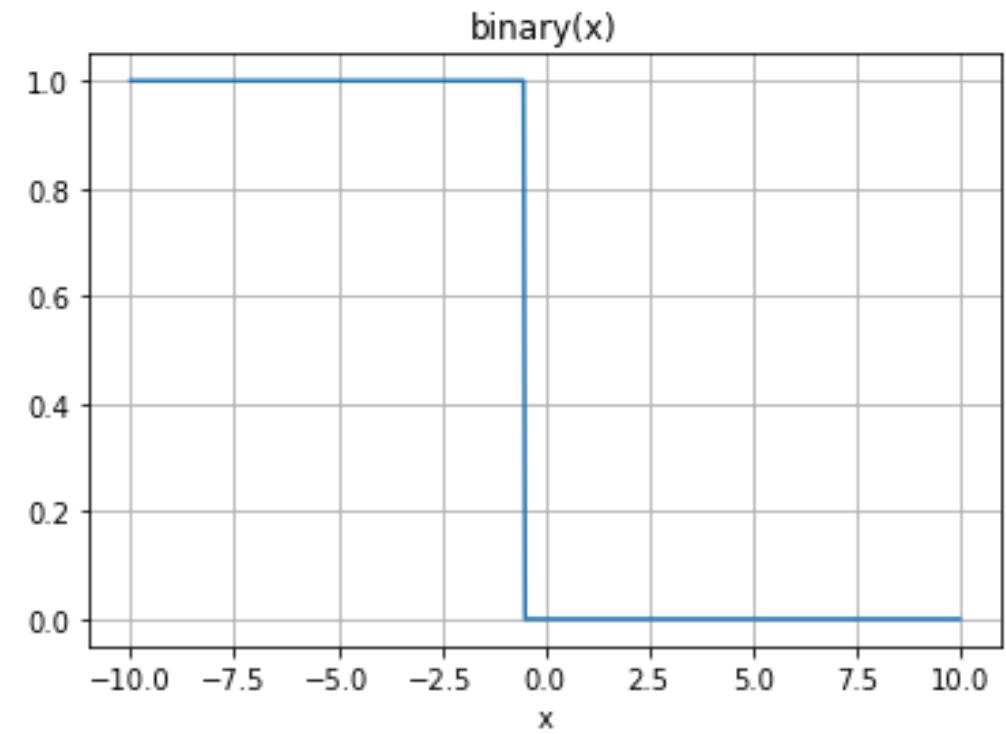
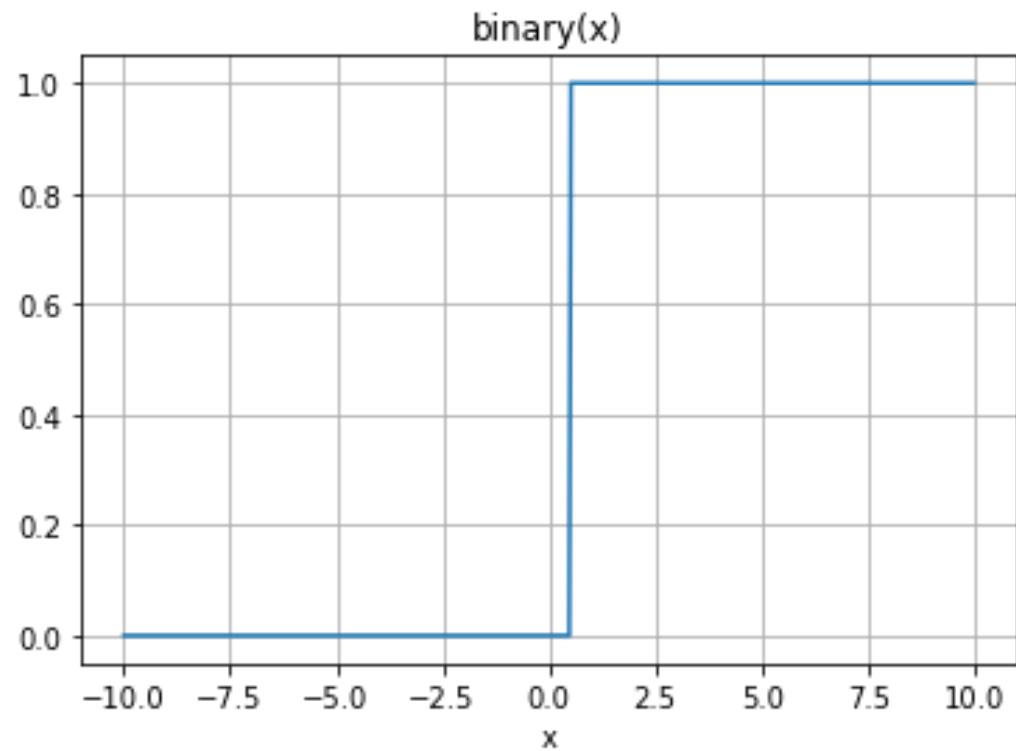
Regresión logística



Regresión logística



Regresión logística



Regresión logística

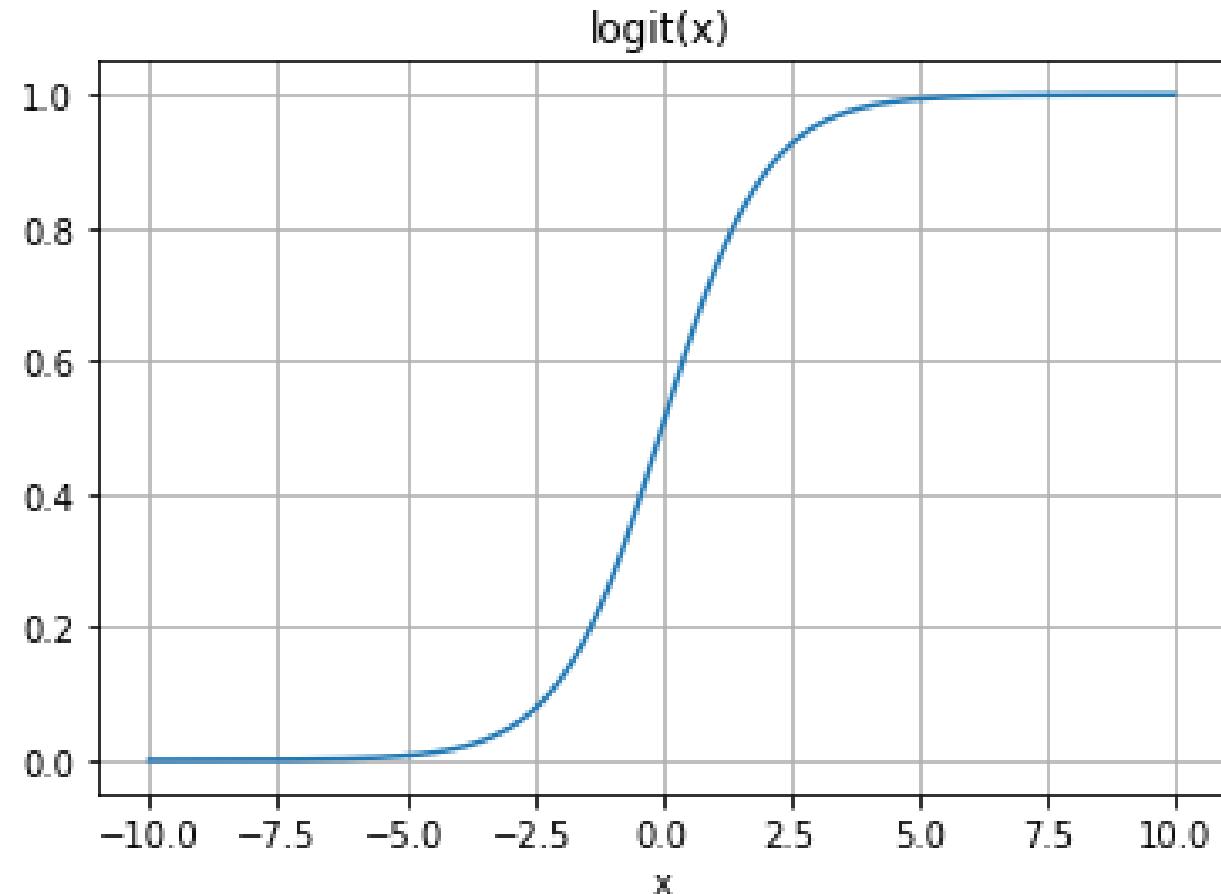
¿Cómo resolvemos el problema?

Integremos una función que transforme los valores arrojados por la regresión lineal a un valor entre 0 y 1.

Una de las funciones más conocidas que puede hacer esto es la función logística o función sigmoide:

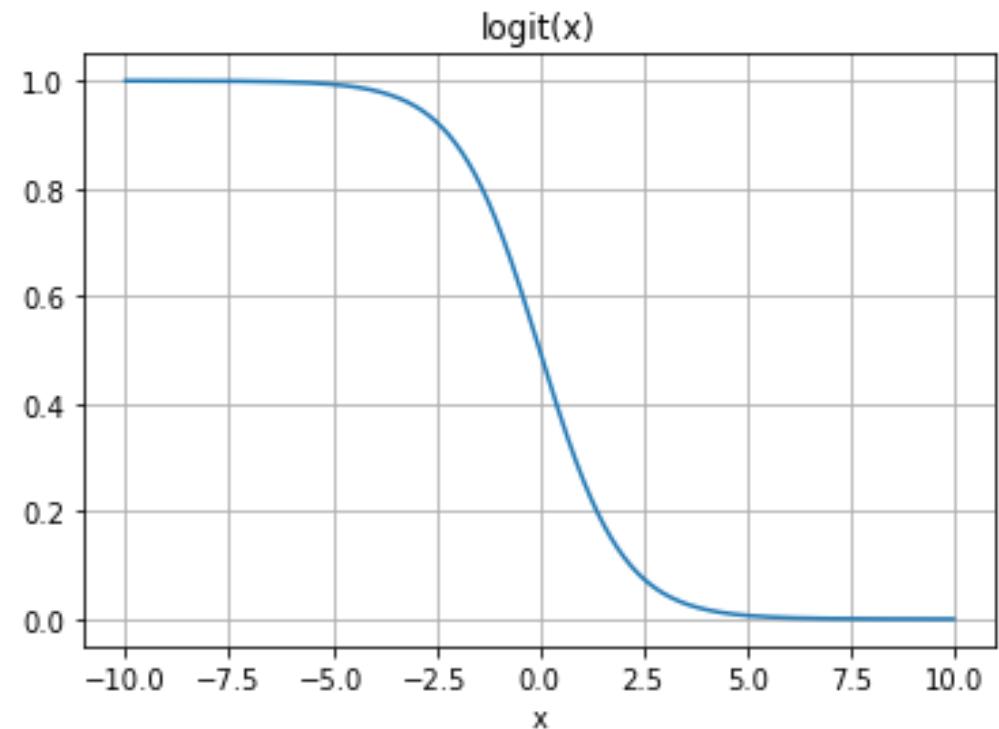
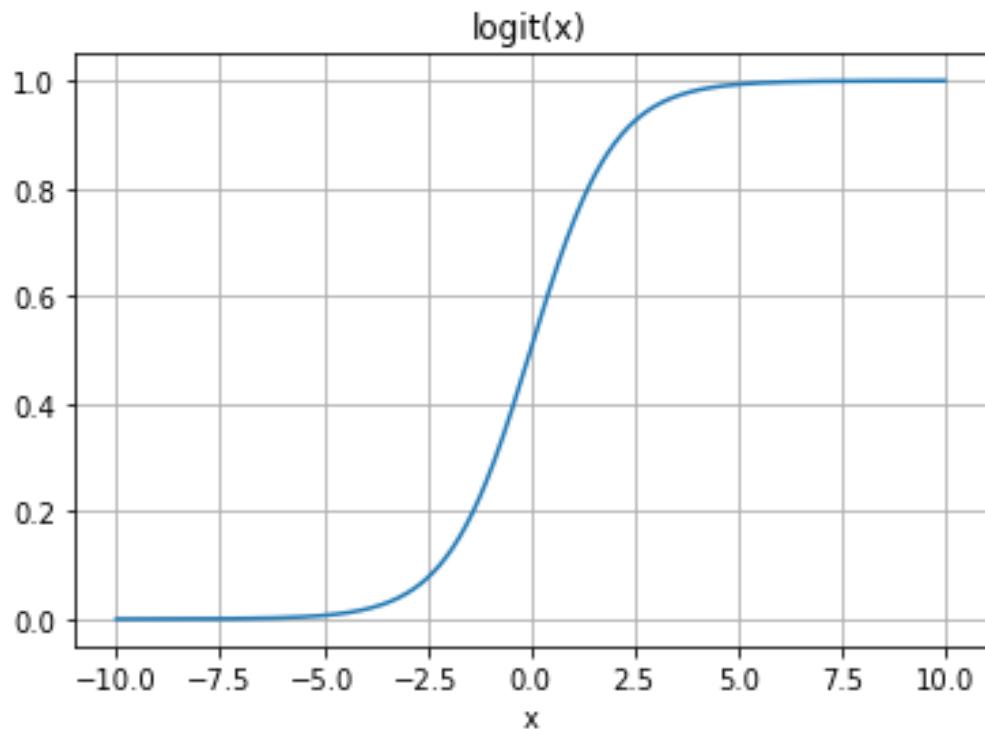
$$\sigma(y) = \frac{1}{1 + e^{-y}}$$

Regresión logística



$$\sigma(y) = \frac{1}{1 + e^{-y}}$$

Regresión logística



Regresión logística

Tenemos que

$$y(\mathbf{x}, \mathbf{W}) = w_0 + w_1 x_1 + \cdots + w_D x_D$$

$y(\mathbf{x}, \mathbf{W})$ es el valor devuelto de la regresión así que lo substituimos en

$$\sigma(y) = \frac{1}{1 + e^{-y}}$$

obtenemos,

$$\sigma(y(\mathbf{x}, \mathbf{W})) = \frac{1}{1 + e^{-(w_0 + w_1 x_1 + \cdots + w_D x_D)}}$$

Regresión logística

$$\sigma(y(\mathbf{x}, \mathbf{w})) = \frac{1}{1 + e^{-(w_0 + w_1x_1 + \dots + w_Dx_D)}}$$

se puede reescribir como (o representa)

$$P(y = 1 | X = \mathbf{x}) = \frac{1}{1 + e^{-(w_0 + w_1x_1 + \dots + w_Dx_D)}}$$

Regresión logística

El problema es entonces encontrar

$$\mathbf{w} = \begin{bmatrix} w_0 \\ \vdots \\ w_N \end{bmatrix}$$

Tal que se maximiza la precisión (*accuracy*) de

$$P(y = 1 | X = \mathbf{x}) = \frac{1}{1 + e^{-(w_0 + w_1 x_1 + \dots + w_D x_D)}}$$

Con los datos de entrenamiento

Regresión logística

El con la regresión lineal no lo podemos hacer por que

$$P(y = 1 | X = \mathbf{x}) = \frac{1}{1 + e^{-(w_0 + w_1x_1 + \dots + w_Dx_D)}}$$

¡No es una ecuación lineal!

Regresión logística

La cuota o *odds* para un evento se define de la siguiente manera:

$$\text{cuota} = \frac{p}{(1 - p)}$$

Por lo que entre más grande es la cuota más grande es la probabilidad de que se cumpla el evento.

Regresión logística

La cuota o *odds* para un evento se define de la siguiente manera:

$$\text{cuota} = \frac{p}{(1 - p)}$$

Por lo que entre más grande es la cuota más grande es la probabilidad de que se cumpla el evento.

!Podemos usar una regresión lineal sobre el logaritmo de la cuota y no sobre la probabilidad!

Regresión logística

Primero veamos que

$$\begin{aligned} P(y = 1 | X = \mathbf{x}) &= \frac{1}{1 + e^{-(w_0 + w_1 x_1 + \dots + w_D x_D)}} \\ &= \frac{1}{\frac{e^{w_0 + w_1 x_1 + \dots + w_D x_D}}{e^{w_0 + w_1 x_1 + \dots + w_D x_D} + 1} + \frac{1}{e^{w_0 + w_1 x_1 + \dots + w_D x_D}}} \\ &= \frac{1}{\frac{e^{w_0 + w_1 x_1 + \dots + w_D x_D} + 1}{e^{w_0 + w_1 x_1 + \dots + w_D x_D}}} \\ &= \frac{e^{w_0 + w_1 x_1 + \dots + w_D x_D}}{e^{w_0 + w_1 x_1 + \dots + w_D x_D} + 1} \end{aligned}$$

Regresión logística

Si

$$P(y = 1 | X = \mathbf{x}) = \frac{e^{w_0 + w_1 x_1 + \dots + w_D x_D}}{e^{w_0 + w_1 x_1 + \dots + w_D x_D} + 1}$$

Entonces

$$\begin{aligned} P(y = 0 | X = \mathbf{x}) &= 1 - \frac{e^{w_0 + w_1 x_1 + \dots + w_D x_D}}{e^{w_0 + w_1 x_1 + \dots + w_D x_D} + 1} \\ &= \frac{e^{w_0 + w_1 x_1 + \dots + w_D x_D} + 1 - e^{w_0 + w_1 x_1 + \dots + w_D x_D}}{e^{w_0 + w_1 x_1 + \dots + w_D x_D} + 1} \\ &= \frac{1}{e^{w_0 + w_1 x_1 + \dots + w_D x_D} + 1} \end{aligned}$$

Regresión logística

Como consecuencia, si

$$P(y = 1|X = \mathbf{x}) = \frac{e^{w_0 + w_1x_1 + \dots + w_Dx_D}}{e^{w_0 + w_1x_1 + \dots + w_Dx_D} + 1}$$

y

$$P(y = 0|X = \mathbf{x}) = \frac{1}{e^{w_0 + w_1x_1 + \dots + w_Dx_D} + 1}$$

Tenemos que

$$\frac{P(y = 1|X = \mathbf{x})}{P(y = 0|X = \mathbf{x})} = \frac{\frac{e^{w_0 + w_1x_1 + \dots + w_Dx_D}}{e^{w_0 + w_1x_1 + \dots + w_Dx_D} + 1}}{\frac{1}{e^{w_0 + w_1x_1 + \dots + w_Dx_D} + 1}}$$

Regresión logística

Esto es

$$\frac{P(y = 1|X = \mathbf{x})}{P(y = 0|X = \mathbf{x})} = e^{w_0 + w_1x_1 + \dots + w_Dx_D}$$

Por lo que

$$\ln\left(\frac{P(y = 1|X = \mathbf{x})}{P(y = 0|X = \mathbf{x})}\right) = w_0 + w_1x_1 + \dots + w_Dx_D$$

¡Esto es exactamente un problema de regresión lineal, sólo que lo es en la cuota (*odds*) de la probabilidad!

Regresión logística

Interpretación del modelo

- x_i : predictor i
- w_0 : ordenada origen, cuando todos los predictores son 0. Es el valor esperado del logaritmo de *odds*. La probabilidad de que pertenezca a la clase 1 es entonces $\frac{e^{w_0}}{1-e^{w_0}}$
- w_i : efecto promedio que tiene el logaritmo de *odds* al incrementar una unidad al predictor x_i , manteniendo constante el resto de las variables. Por cada unidad en el incremento de x_i se multiplican los *odds* de e^{w_i} .

Condiciones para la regresión logística

Se deben cumplir las siguientes condiciones, aún que no siempre es posible demostrar todas ellas

- Relación lineal entre los predictores numéricos y el logaritmo de *odds* de la variable respuesta.
- No colinealidad o multicolinealidad entre predictores.
- No autocorrelación (Independencia).
- Valores atípicos, con alto *leverage* o influyentes.
- Tamaño de la muestra (que el número de observaciones sea como mínimo entre 10 y 20 veces el número de predictores del modelo.)
- Parsimonia (simplicidad, menor número de predictores)

Ejemplo 09: Regresión logística 1

En este ejemplo se usa la librería [Scikit-learn.](#)

Ejemplo 10: Regresión logística 2

En este ejemplo se usa la librería la librería [statsmodels](#).

Ejemplo 11: Regresión logística 3

Regresión logística usando la librería [statsmodels](#) para clasificar correos.

Ejercicio 03: Regresión logística simple



Ejemplo 12: Regresión logística múltiple

Regresión logística múltiple usando la librería [statsmodels](#) para clasificar correos.

One-Vs-Rest for Multi-Class Classification

El método "uno a uno" (OvR), también conocido como "uno a uno" o "OvA", es un método heurístico para utilizar algoritmos de clasificación binaria para la clasificación multiclas.

Consiste en dividir el conjunto de datos multiclas en múltiples problemas de clasificación binaria. A continuación, se entrena un clasificador binario en cada problema de clasificación binaria y se hacen predicciones utilizando el modelo que tenga más confianza.

One-Vs-Rest for Multi-Class Classification

Por ejemplo, dado un problema de clasificación multiclas con ejemplos para cada clase "rojo", "azul" y "verde". Esto podría dividirse en tres conjuntos de datos de clasificación binaria como sigue:

- Problema de clasificación binaria 1: rojo frente a [azul, verde]
- Problema de clasificación binaria 2: azul frente a [rojo, verde].
- Problema de clasificación binaria 3: verde frente a [rojo, azul].

Multinomial Logistic Regression

- Extensión de la regresión logística que usa un modelo de regresión logística para cada par de categorías.
- Si tenemos N categorías tendríamos $N(N-1)$ posibles pares.
- Sólo necesitamos $N-1$ pares

Multinomial Logistic Regression

- Interpretación: todos los modelos dependen en que respuesta se tenga en el denominador y el numerador.
- La categoría correspondiente al denominador es llamada **referencia** o **categoría base** y es usualmente:
 - La primera o última categoría
 - La categoría mas “significativa”
 - La categoría mas frecuente
 - No una categoría rara.

Multinomial Logistic Regression

$$P(y = 1|X = \mathbf{x}) = \frac{1}{1 + e^{-(w_0 + w_1x_1 + \dots + w_Dx_D)}}$$

$$P(y = 1|X = \mathbf{x}) = \frac{1}{1 + e^{-(w_0 + w_1x)}}$$

$$\log\left(\frac{\pi_1}{\pi_3}\right) = w_{0,1} + w_{1,1}x \quad \text{y} \quad \log\left(\frac{\pi_2}{\pi_3}\right) = w_{0,2} + w_{1,2}x$$

Multinomial Logistic Regression

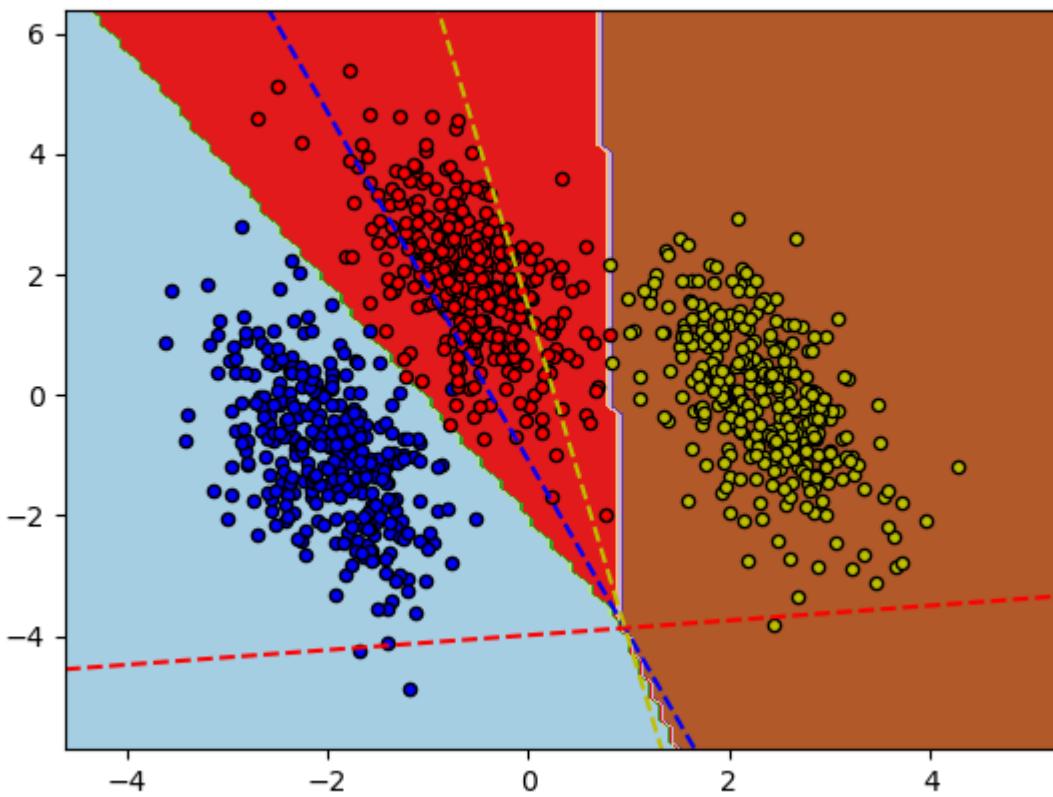
$$\pi_1 = \frac{\exp(w_{0,1} + w_{1,1}x)}{1 + \exp(w_{0,1} + w_{1,1}x) + \exp(w_{0,2} + w_{1,2}x)}$$

$$\pi_2 = \frac{\exp(w_{0,2} + w_{1,2}x)}{1 + \exp(w_{0,1} + w_{1,1}x) + \exp(w_{0,2} + w_{1,2}x)}$$

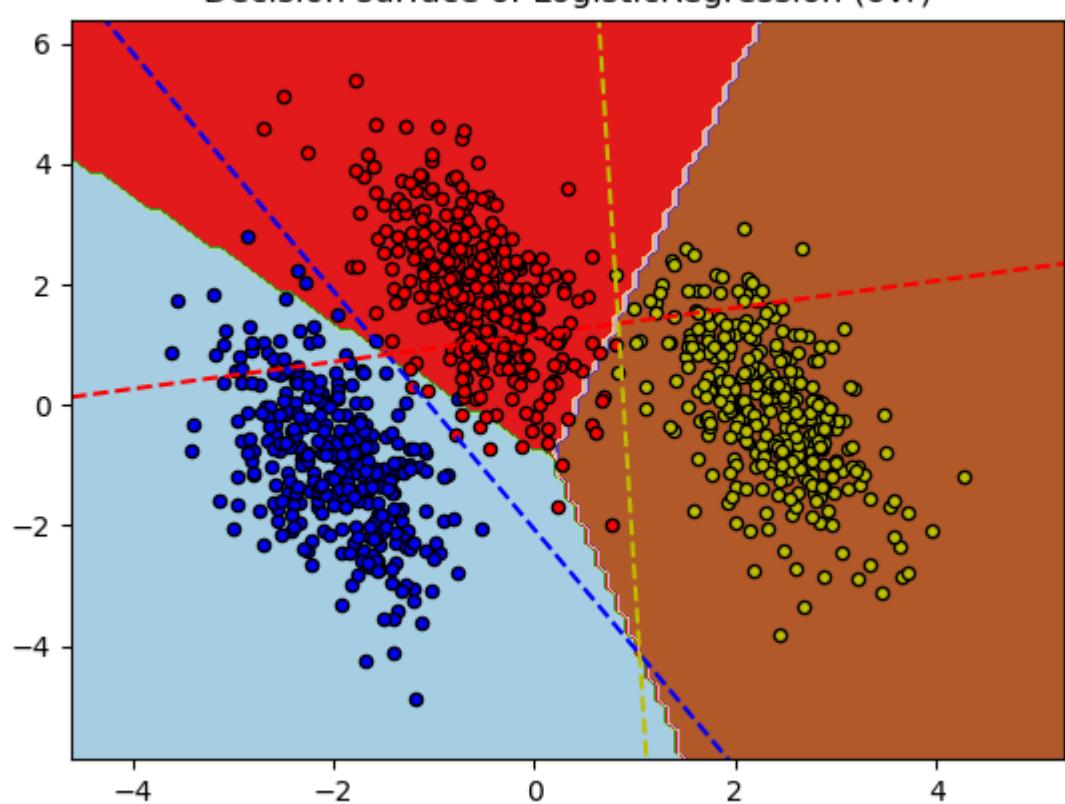
$$\pi_3 = \frac{1}{1 + \exp(w_{0,1} + w_{1,1}x) + \exp(w_{0,2} + w_{1,2}x)}$$

Para más información ver: NCRMUK (2021). (Consultado el 01/09/2022). Multinomial logistic regression, Part 1: Introduction . In Youtube (ed.),
<https://www.youtube.com/watch?v=JcCBIPqcwFo>

Decision surface of LogisticRegression (multinomial)



Decision surface of LogisticRegression (ovr)



Ejemplo 13: Clasificación Multinomial y One-vs-Rest classification