

UNIVERSIDAD
NACIONAL
DE COLOMBIA

MÉTODOS DE CLASIFICACIÓN

Carlos A. Mera B.

Departamento de Ciencias de la Computación y de la Decisión
Investigador del Grupo de I+D en Inteligencia Artificial – GIDIA

camerab@unal.edu.co

Contenido

1. Motivación
2. Métodos de Aprendizaje de Máquina:
 - a. Métodos Supervisados
 - b. Métodos No Supervisado
3. Regresión Lineal
4. Regresión Logística
5. Naïve Bayes
6. LDA y QDA
7. Máquinas de Vectores de Soporte

Motivación

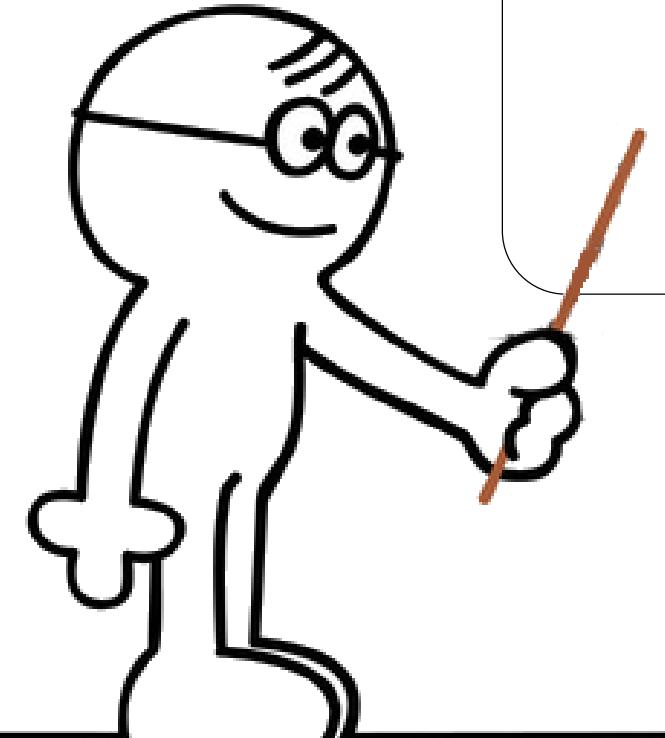


OBSERVE EL VIDEO Y RESPONDA A LAS SIGUIENTES PREGUNTAS:

- ✓ ¿Cuántos datos se requieren para entrenar un sistema de visión artificial?
- ✓ ¿Es posible decir que los computadores ya sobrepasaron la capacidad humana?
- ✓ ¿Qué problemas evidencian los sistemas de visión artificial, y en general de los sistemas de Reconocimiento de Patrones?

The image shows a TED Talk video player. At the top left is the TED logo with the tagline "Ideas worth spreading". Below the logo is a photo of Fei-Fei Li, a woman with dark hair, wearing a colorful abstract patterned dress, standing on a stage and speaking. To her right is a large play button icon. In the bottom right corner of the video frame, there is a grid of numbers (143, 49, 90, 179, 10, 139, etc.) and several social media interaction icons: Share, Add to list, Like, and Recommend. The video title "Cómo estamos enseñando a las computadoras a entender imágenes" is displayed prominently in white text at the bottom of the video frame. The bottom right corner of the video frame also shows a timestamp "17:59" and some volume control icons.

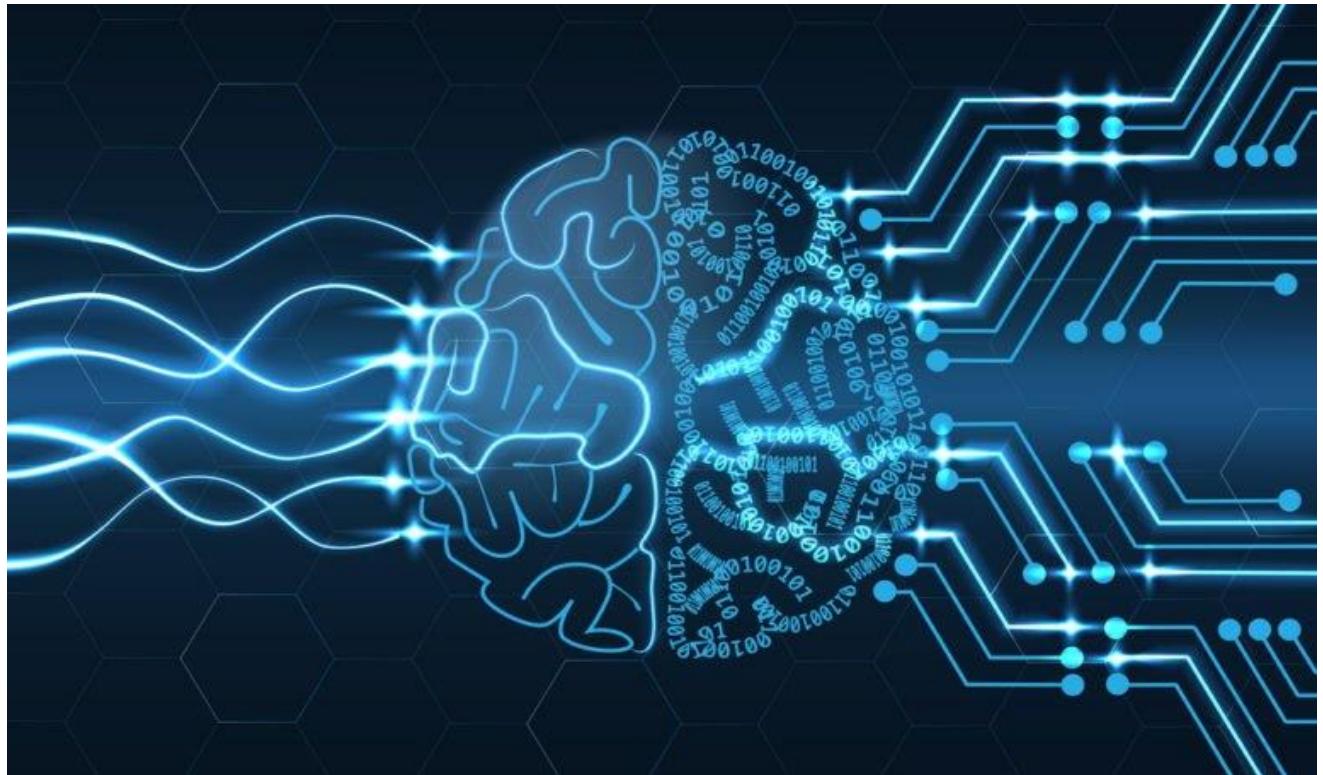
https://www.ted.com/talks/fei_fei_li_how_we_re_teaching_computers_to_understand_pictures?language=es



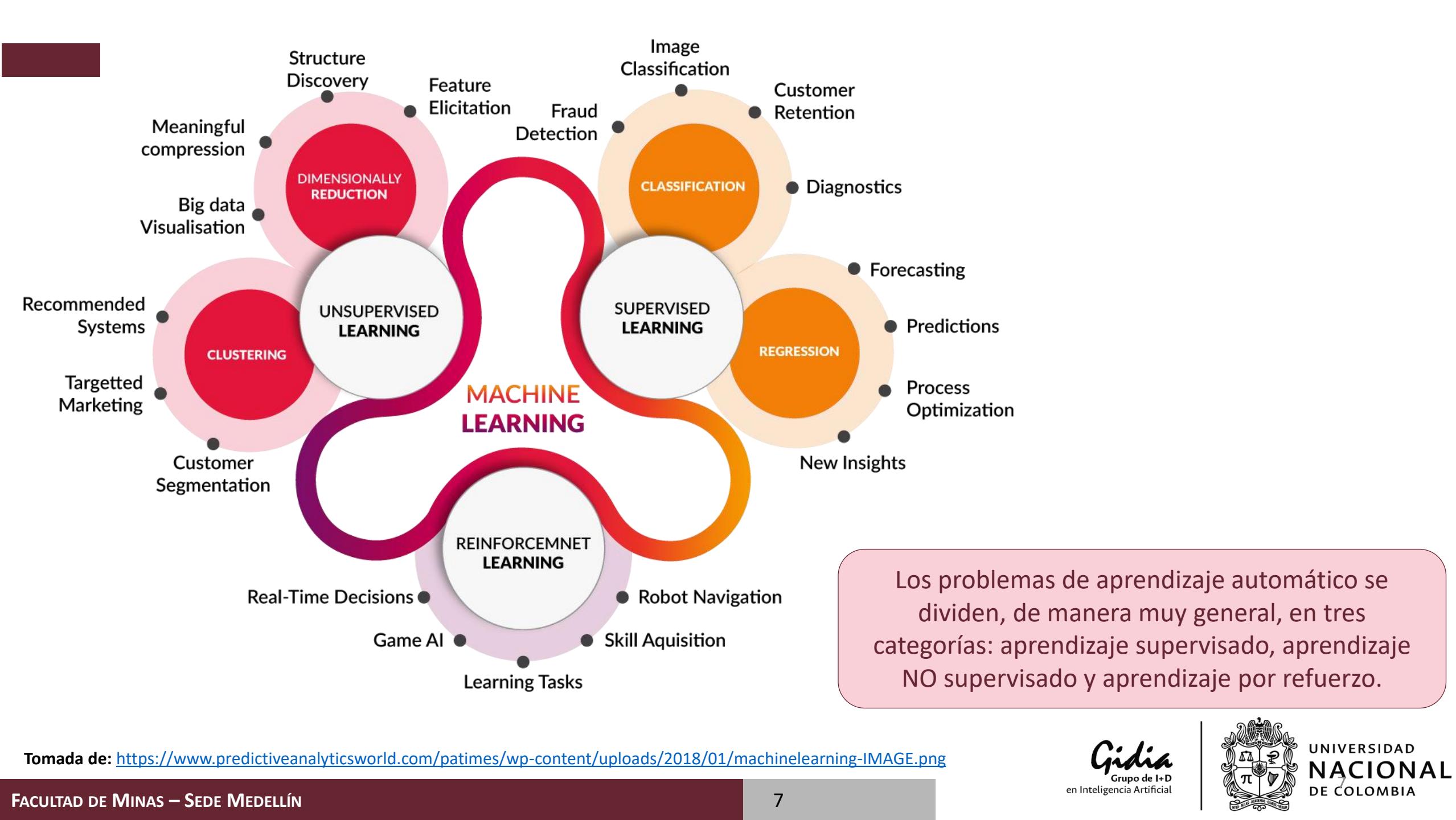
¿QUÉ SON LOS MÉTODOS DE APRENDIZAJE DE MÁQUINA?

Métodos de Aprendizaje de Máquina

- Los métodos de aprendizaje automático tienen como objetivo **crear un modelo** de los datos optimizando un criterio de rendimiento, esto a partir de un conjunto de datos de entrenamiento.
- ¿Por qué hacer que las máquinas aprendan?**
 - La experiencia humana no existe (navegar en Marte)
 - Los humanos no pueden explicar su experiencia (reconocimiento de voz)
 - La solución cambia en el tiempo
 - La solución debe adaptarse a casos particulares (biometría del usuario)

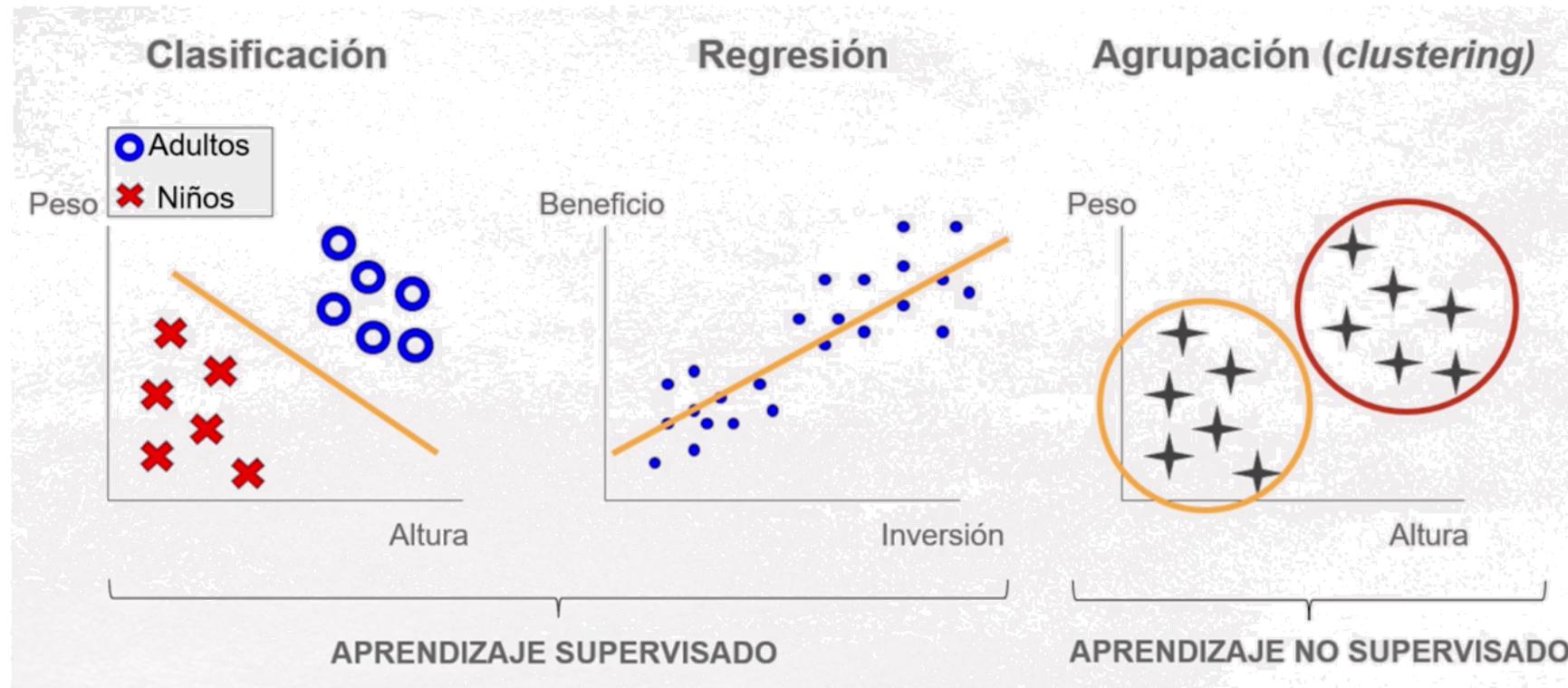


Machine Learning



Tomada de: <https://www.predictiveanalyticsworld.com/patimes/wp-content/uploads/2018/01/machinelearning-IMAGE.png>

Aprendizaje Supervisado vs No Supervisado



Aprendizaje Supervisado

VS

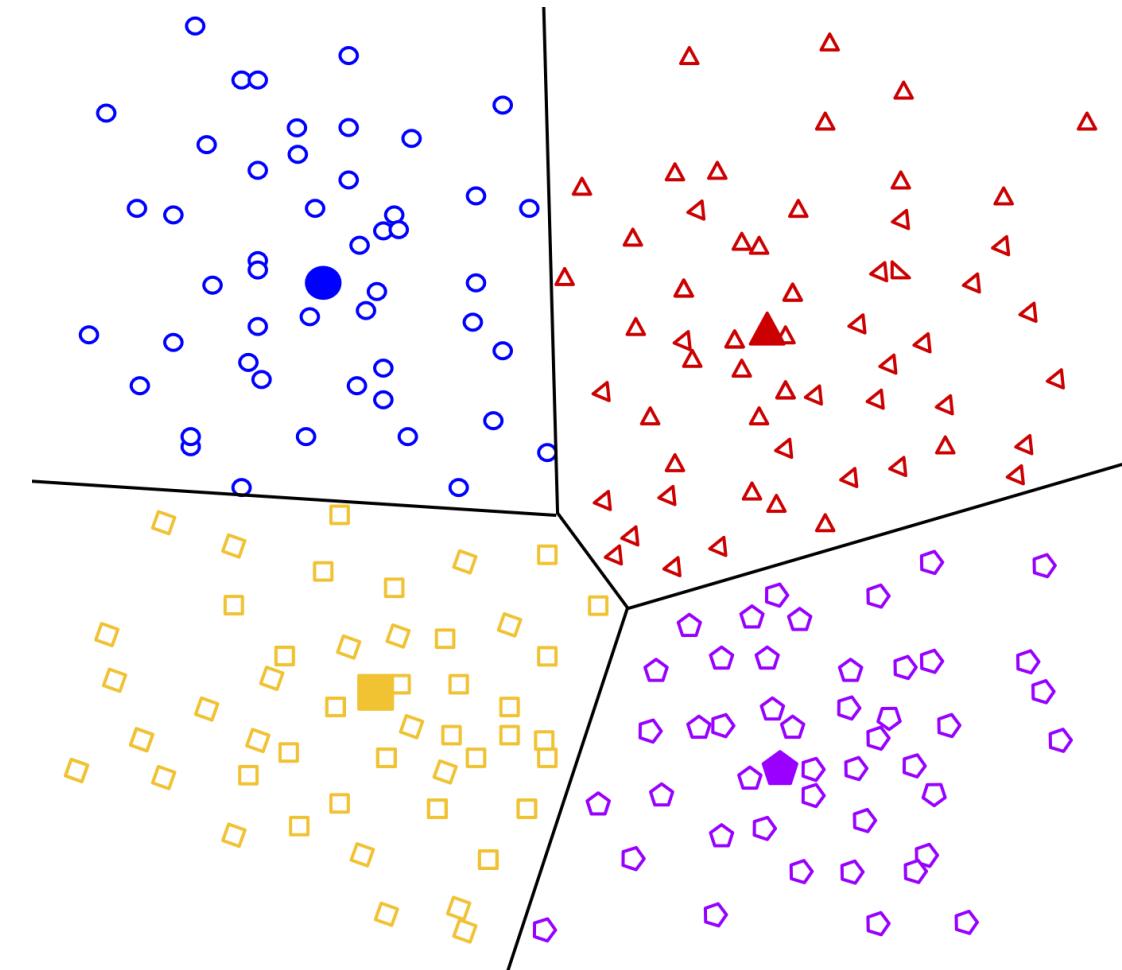
Aprendizaje NO Supervisado

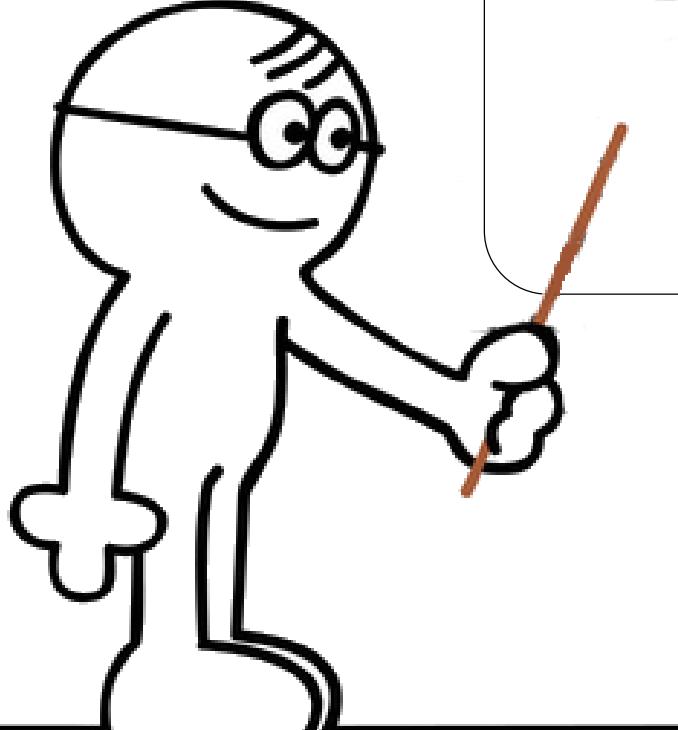
Aprendizaje Supervisado: requiere de un conjunto de datos conocidos a partir del cual se crea un **modelo** para predecir el valor de una variable de salida. El aprendizaje supervisado se puede usar en dos tareas:

- **Clasificación:** en este caso la variable de salida es una etiqueta que determina la clase a la que pertenecen los datos de entrada, es decir, la variable de salida es una **variable discreta**.
- **Regresión:** en este caso los algoritmos de aprendizaje buscan predecir el valor de una **variable continua** a partir de los datos de entrada. Un ejemplo de una tarea de regresión es el de estimar la longitud de un salmón en función de su edad y su peso.

En el **Aprendizaje no supervisado** se cuenta con un conjunto de datos de entrenamiento, pero no hay una variable específica de salida (se desconocen las clases). En este sentido, el objetivo de los problemas del aprendizaje no supervisado es, por ejemplo, el de agrupar los datos de entrada con base en algún criterio de similitud o disimilitud o determinar la distribución estadística de los datos, conocida como estimación de la densidad.

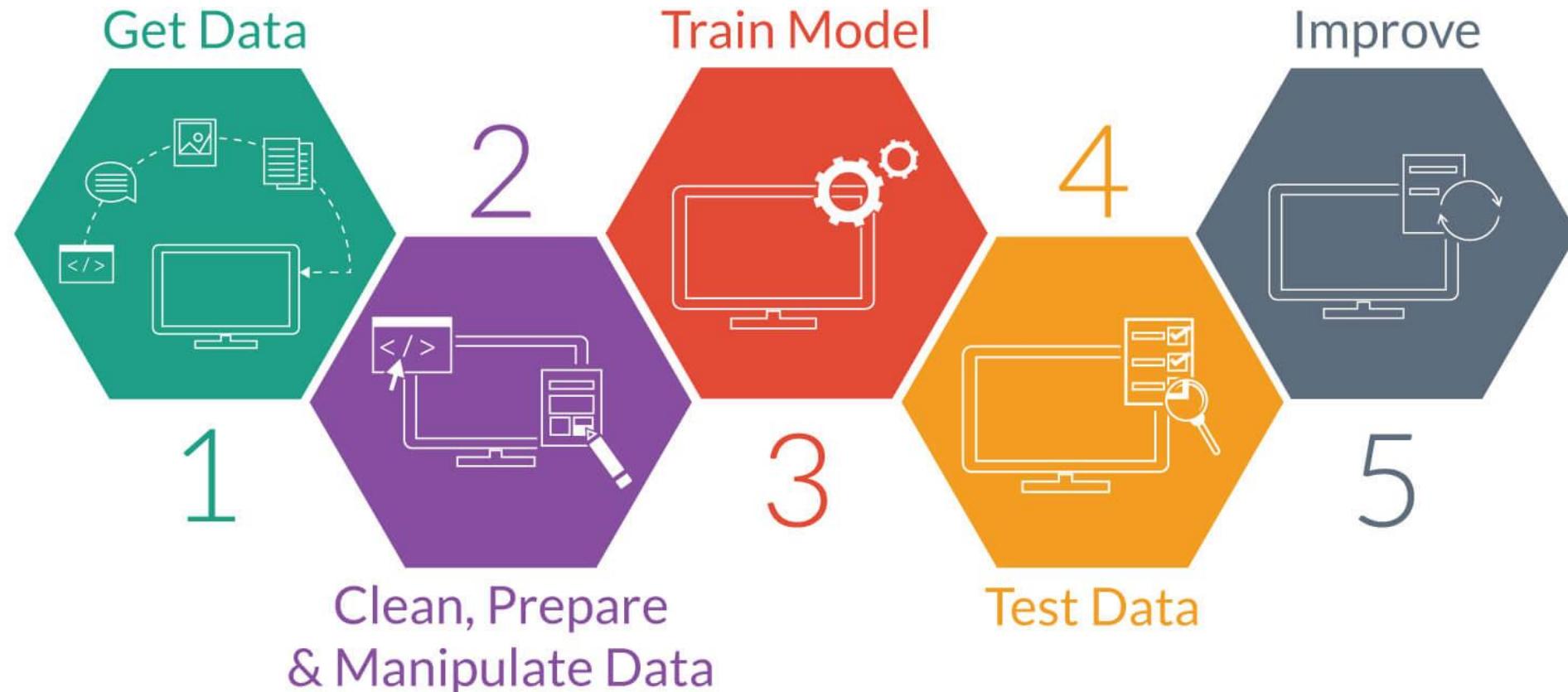
Aprendizaje Supervisado vs Aprendizaje No Supervisado





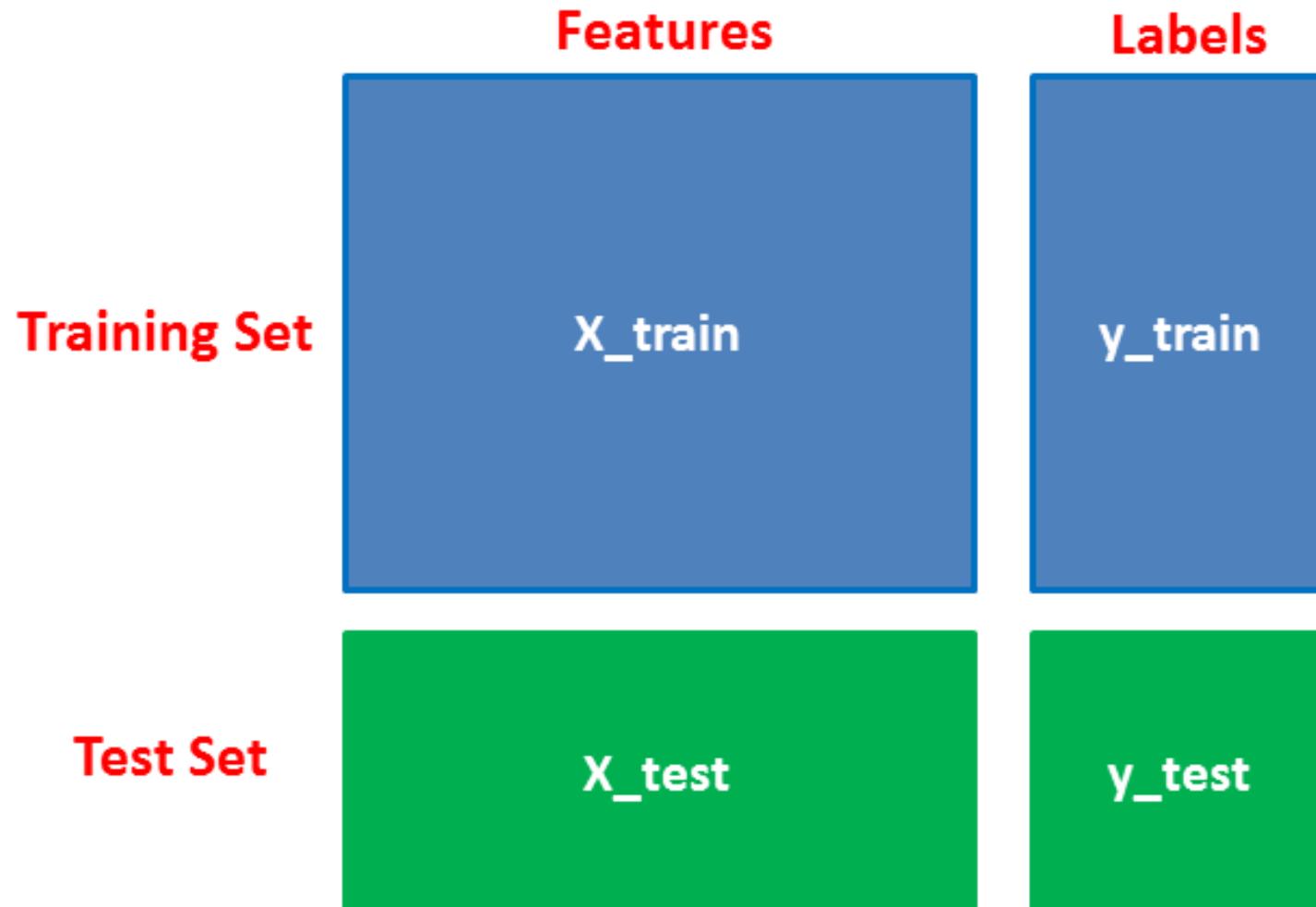
FLUJO DE TRABAJO DE LOS MÉTODOS DE APRENDIZAJE AUTOMÁTICO

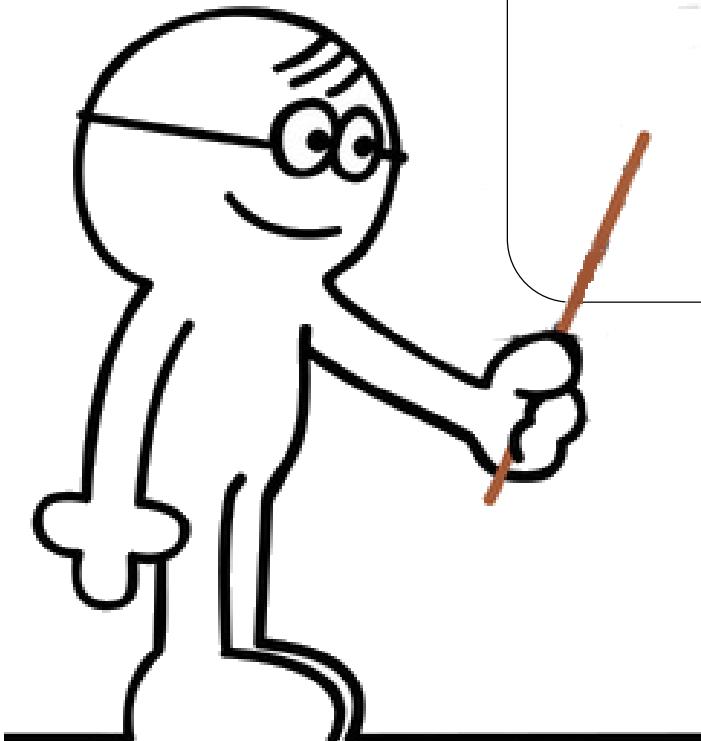
Flujo de Trabajo



Tomada de: <https://newtiummedia.blob.core.windows.net/images/Steps-to-Predictive-Modelling.jpg>

Partición del Conjunto de Datos





¿QUÉ ES LA REGRESIÓN LINEAL?

Regresión Lineal



CORRELACIÓN ENTRE DOS VARIABLES:

Se considera que dos variables cuantitativas (x e y) están **correlacionadas** cuando una de ellas (y) varía sistemáticamente con respecto a los valores de la otra (x).

Por ejemplo:

- ✓ ¿Hay una correlación entre la Temperatura y el número de Helados Vendidos en una Heladería?
- ✓ ¿Puede identificar otras correlaciones?

Claro está, si sabemos que la variable x está correlacionada con y , quiere decir que podemos **predecir** la variable y a partir de x .

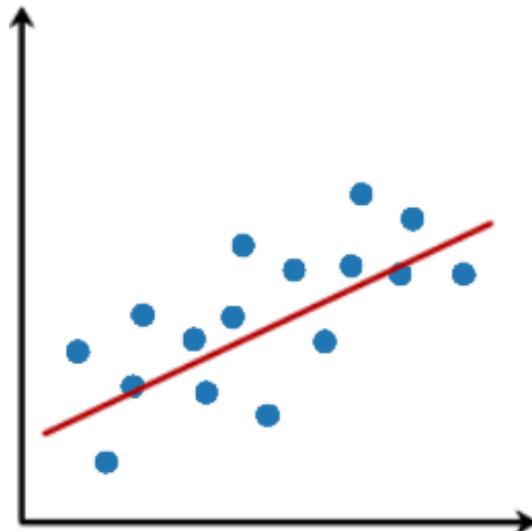
Estamos en el terreno de la PREDICCIÓN!

Regresión Lineal



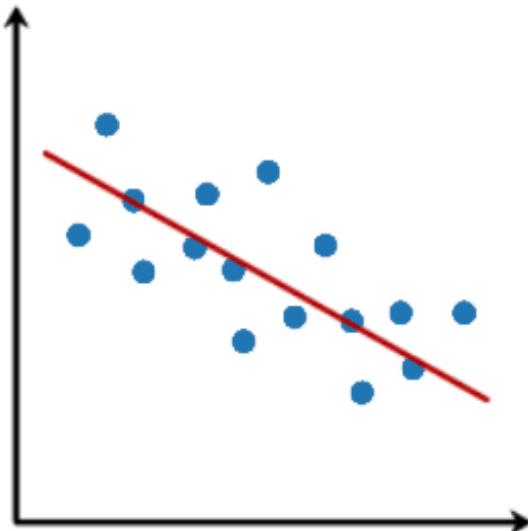
TIPOS DE CORRELACIÓN:

Hay tres tipos básicos de correlación: positiva, negativa y nula (sin correlación).



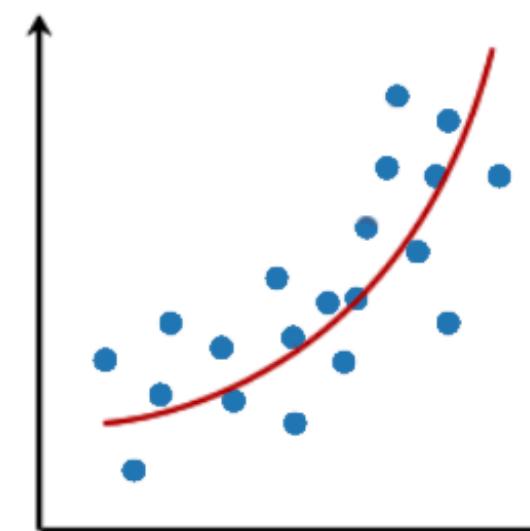
Correlación Lineal Positiva

Ocurre cuando una variable aumenta y la otra también

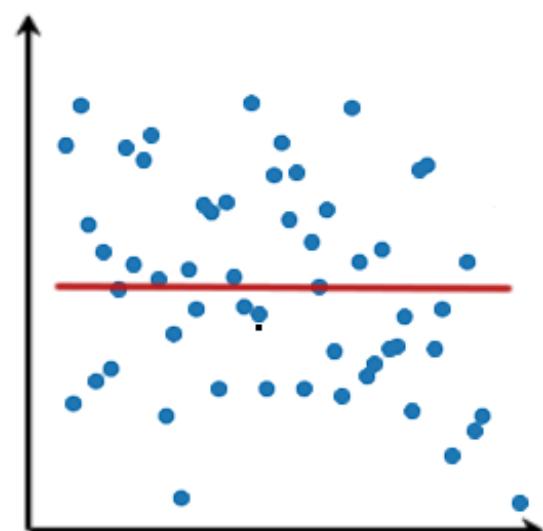


Correlación Lineal Negativa

Ocurre cuando una variable aumenta y la otra disminuye

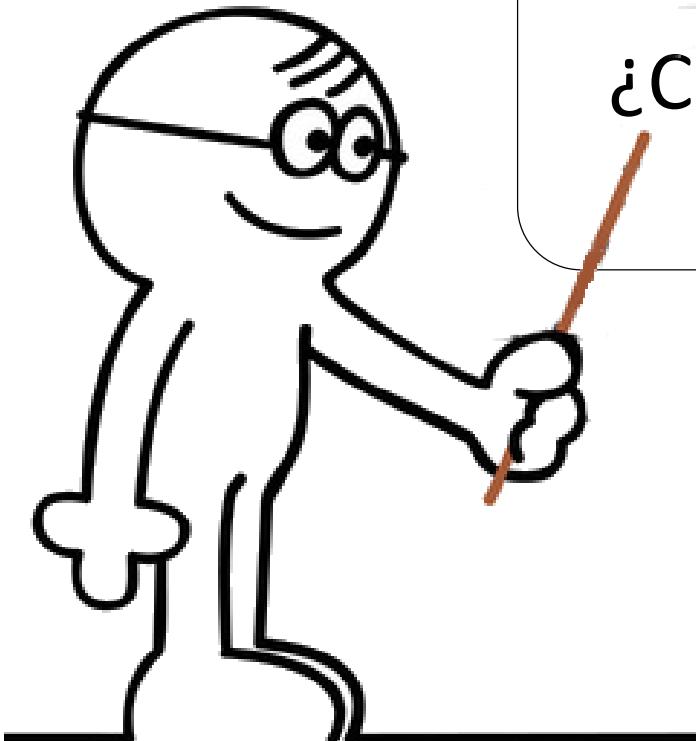


Correlación No Lineal



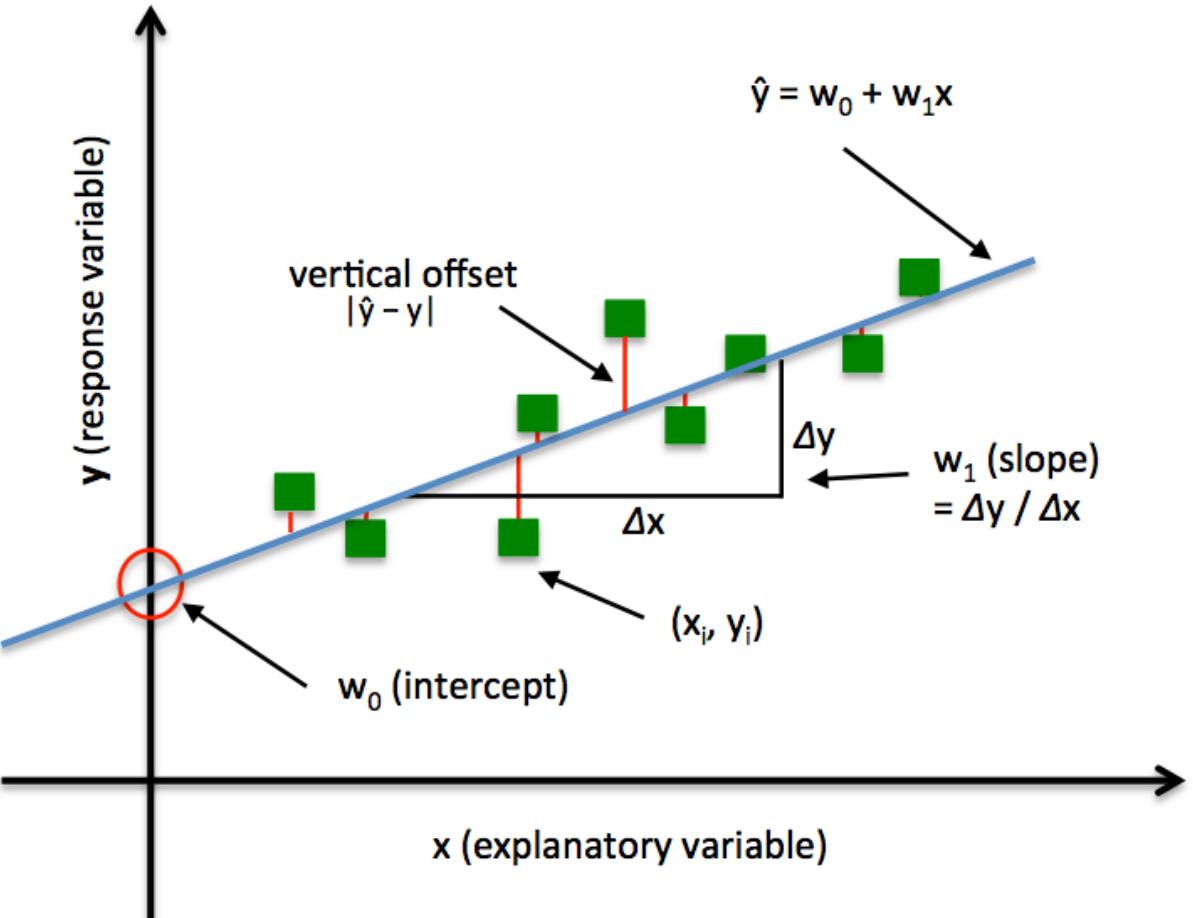
Sin Correlación

No hay una relación aparente entre las variables



¿CUÁL ES EL PRINCIPIO DE LA REGRESIÓN LINEAL?

Regresión Lineal



■ Generalidades:

- Los métodos de regresión buscan modelar la relación entre 2 variables.
- El modelo se ajusta usando una medida de error sobre las predicciones que éste hace.
- En la **Regresión Lineal** el modelo a ajustar es una línea recta:

$$\hat{y} = w_0 + w_1x$$

Puede haber múltiples líneas rectas dependiendo de los valores de intercepción y pendiente. Básicamente, lo que hace el algoritmo de regresión lineal es ajustar varias líneas y retornar la línea que produce el menor error.

Regresión Lineal

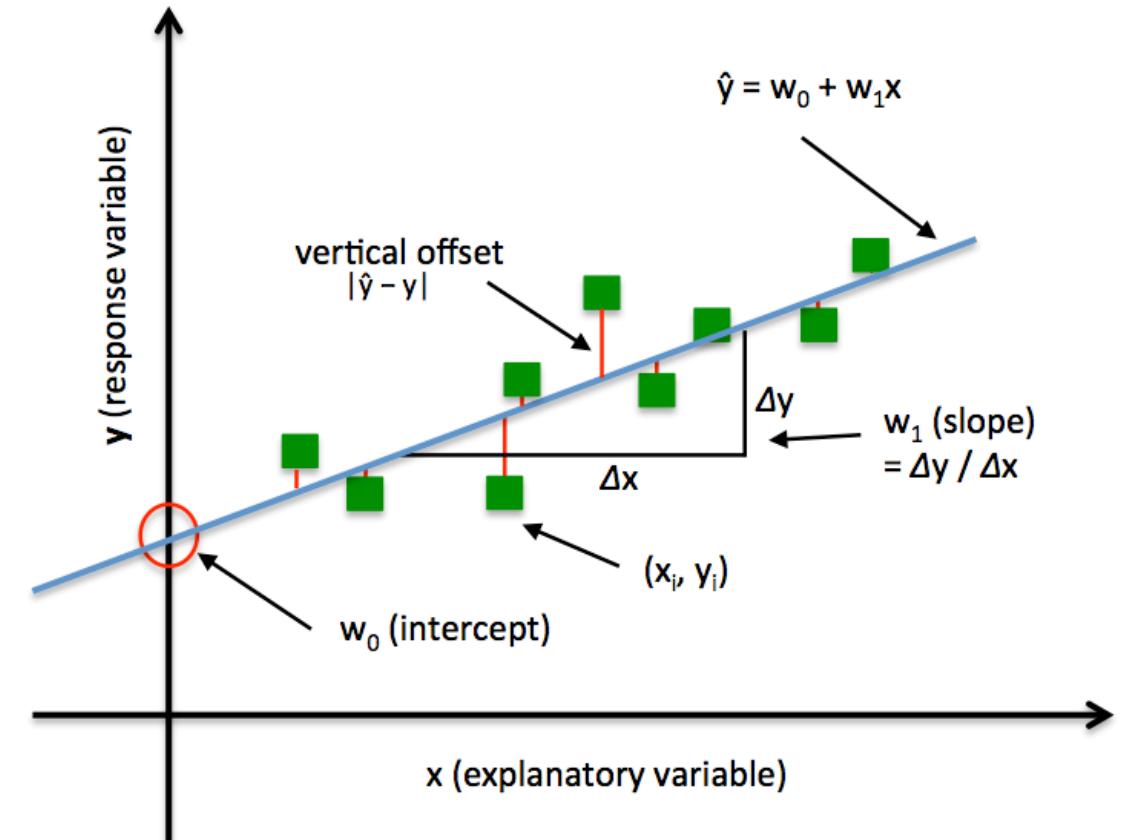
GENERALIDADES:

En la **REGRESIÓN LINEAL POR MÍNIMOS CUADRADOS**, el objetivo es encontrar la línea (o hiperplano) que minimiza la suma de los errores al cuadrado (SSE) o el error al cuadrado medio (MSE) entre la variable objetivo (y) y la salida del modelo sobre todas las muestras x_i . Entonces, buscamos aquella línea tal que minimice la función de costo:

Predicción
del Modelo

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

↓
Valor
Observado



Regresión Lineal



REGRESIÓN LINEAL MÚLTIPLE:

Cuando la **Regresión Lineal** se usa para predecir una variable y a partir de **más de una variable (x^i)**, el modelo debe ajustar un hiperplano a los datos dados:

$$\hat{y} = w_0 + w_1 x^1 + w_2 x^2 + \dots + w_d x^d$$

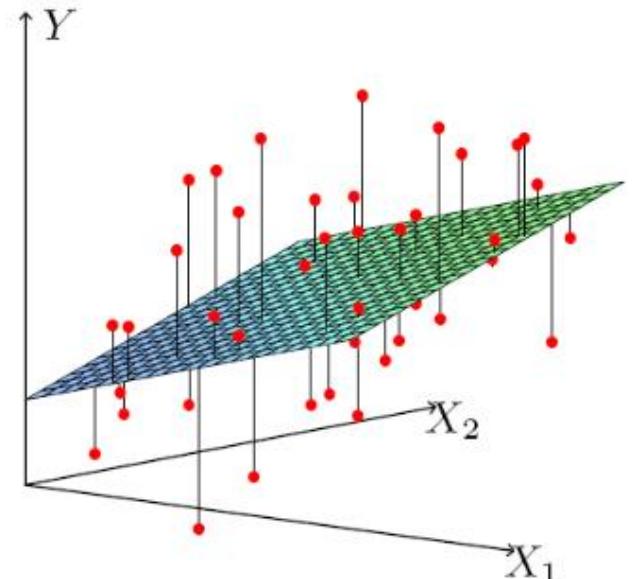
$$\hat{y} = h_{\mathbf{W}}(\mathbf{x}) = \sum_{k=0}^d w_k x^k = \mathbf{W}^T \cdot \mathbf{x}, \text{ con } x^0 = 1$$

Donde,

- $h_{\mathbf{W}}$ es la función de hiperplano
- \mathbf{W} es el vector de parámetros del modelo (a estimar)
- $\mathbf{x} = \{x^1, x^2, \dots, x^d\}$ es el vector de características d-dimensional

La función de costo a minimizar sigue siendo la suma de los errores al cuadrado:

$$J(\mathbf{W}) = \frac{1}{2} \sum_{i=1}^n (\mathbf{W}^T \cdot \mathbf{x}_i - y_i)^2$$



Regresión Lineal

SOLUCIÓN POR EL MÉTODO TEÓRICO CON ALGEBRA LINEAL:

Para encontrar el valor de W que minimiza la función de error, hay una solución de forma cerrada, en otras palabras, una ecuación matemática que da el resultado directamente.

Dicha ecuación es la siguiente:

$$W = (X^T X)^{-1} X^T y$$



Para más información se sugiere consultar:

- <https://www.oreilly.com/library/view/hands-on-machine-learning/9781491962282/ch04.html>
- <https://www.ritchieng.com/one-variable-linear-regression/>

Regresión Lineal

SOLUCIÓN POR EL MÉTODO ESTADÍSTICO:

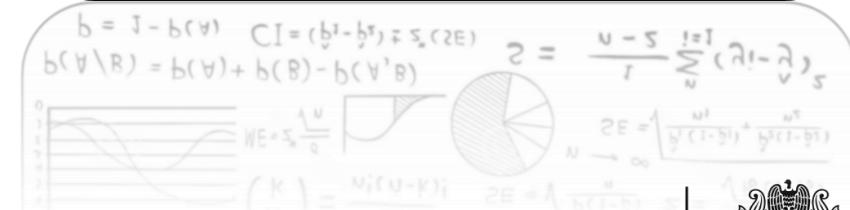
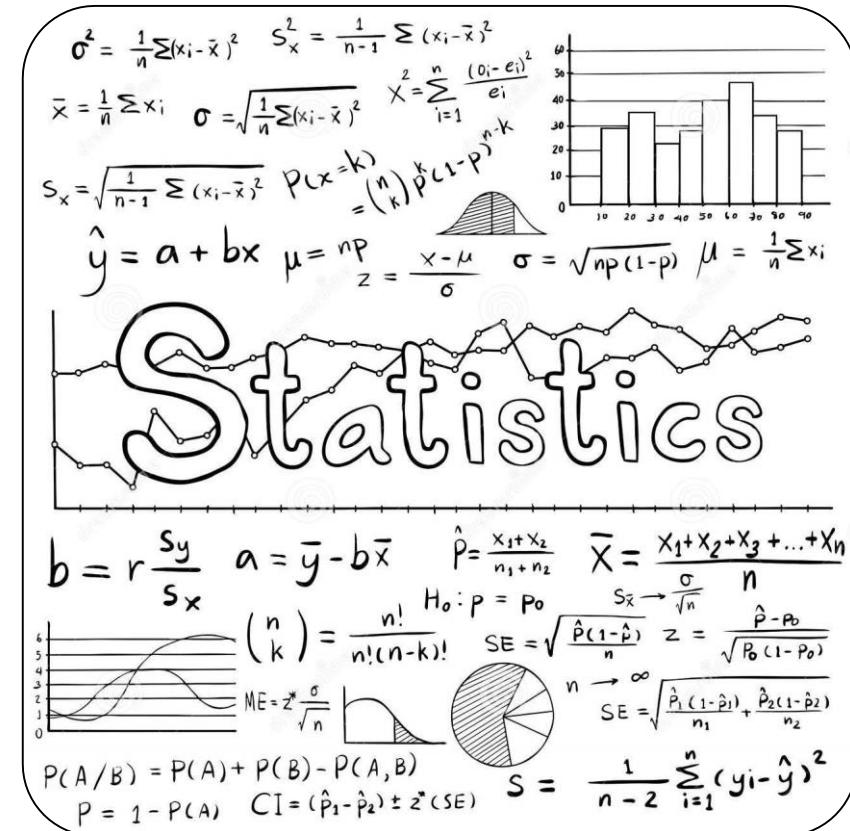
Para encontrar el valor de W , tenemos que:

$$w_0 = \bar{y} - w_1 \bar{x}$$

$$w_1 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}$$

Para más información se sugiere consultar:

- <https://www.oreilly.com/library/view/hands-on-machine-learning/9781491962282/ch04.html>
- <https://www.ritchieng.com/one-variable-linear-regression/>



Regresión Lineal

SOLUCIÓN POR EL MÉTODO DE OPTIMIZACIÓN:

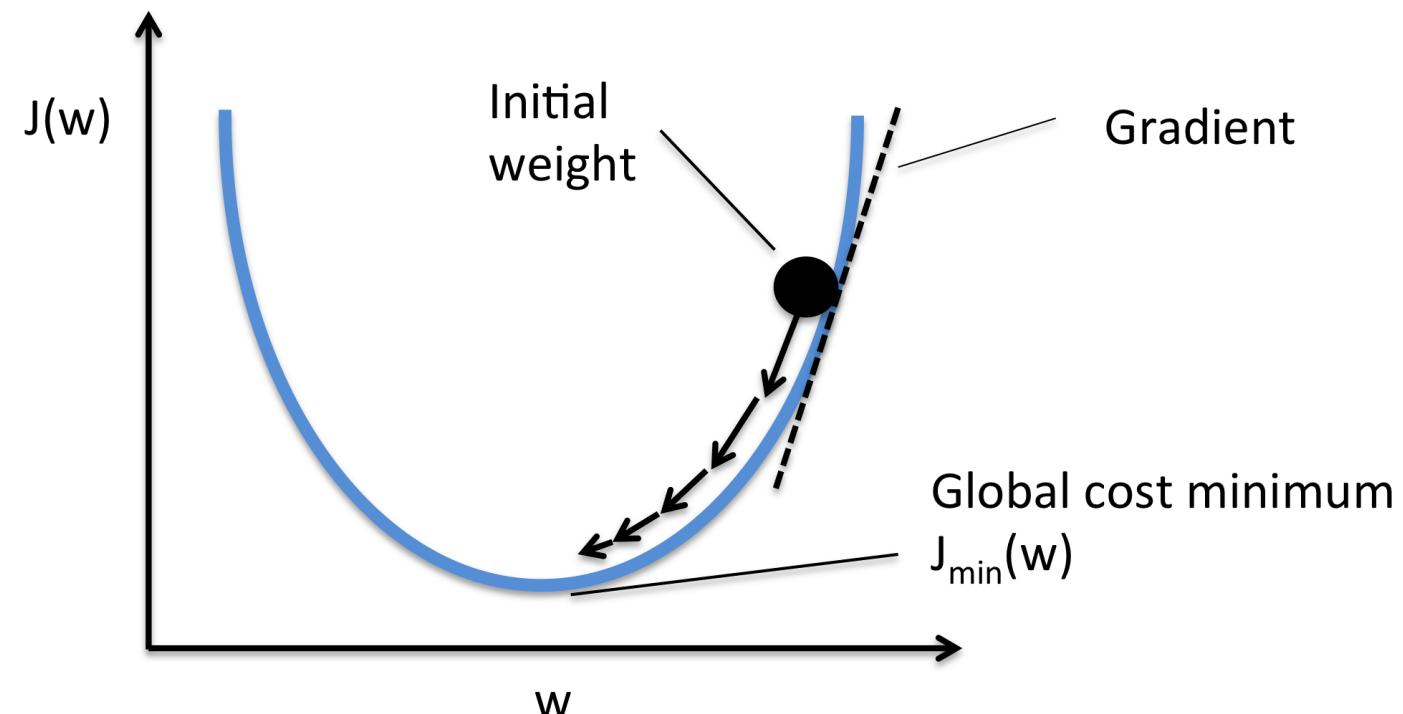
En este caso el valor del parámetro W se busca utilizando métodos de optimización, por ejemplo usando Gradiente Descendente.

En este caso los pesos se actualizan gradualmente después con base en la función de costo, denominada $J(\cdot)$, que es la suma de los errores al cuadrado.

En este caso, la magnitud y la dirección de la actualización de W se calcula dando un paso en la dirección opuesta del gradiente de la función de costo:

$$\Delta w_j = -\eta \frac{\partial J}{\partial w_j}$$

Siendo η la razón (o tasa) de aprendizaje.

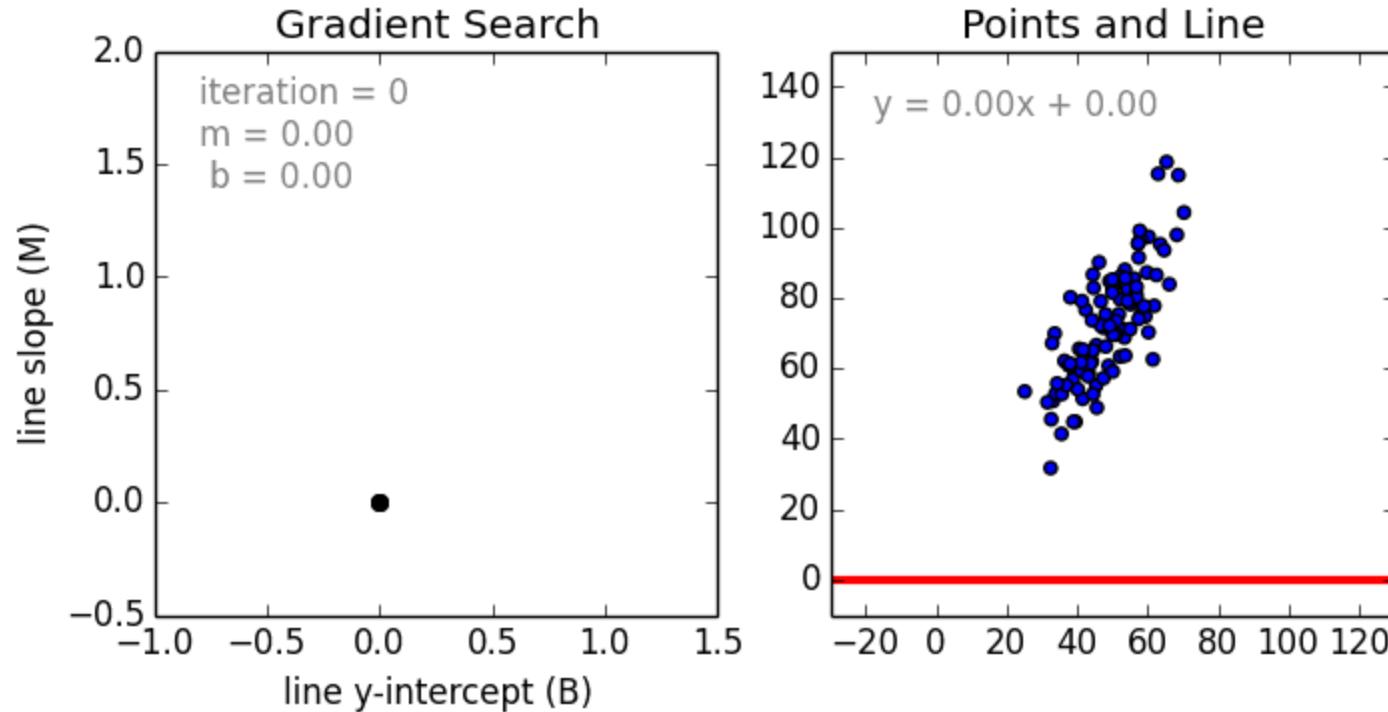


Regresión Lineal



SOLUCIÓN POR EL MÉTODO DE OPTIMIZACIÓN:

En este caso el valor del parámetro W se busca utilizando métodos de optimización, por ejemplo usando Gradiente Descendente.



Regresión Lineal

- A modo de comparación en Sklearn:

Algorithm	Large dataset (m)	Out-of-core support	Large features (n)	Hyperparams	Scaling Required	Scikit-Learn
Normal Equation	Fast	No	Slow	0	No	<code>LinearRegression</code>
Batch Gradient Descent	Slow	No	Fast	2	Yes (Hadoop MapReduce)	n/a
Stochastic Gradient Descent	Fast	Yes	Fast	≥ 2	Yes	<code>SGDRegressor</code>
Mini Gradient Descent	Fast	Yes	Fast	≥ 2	Yes	n/a

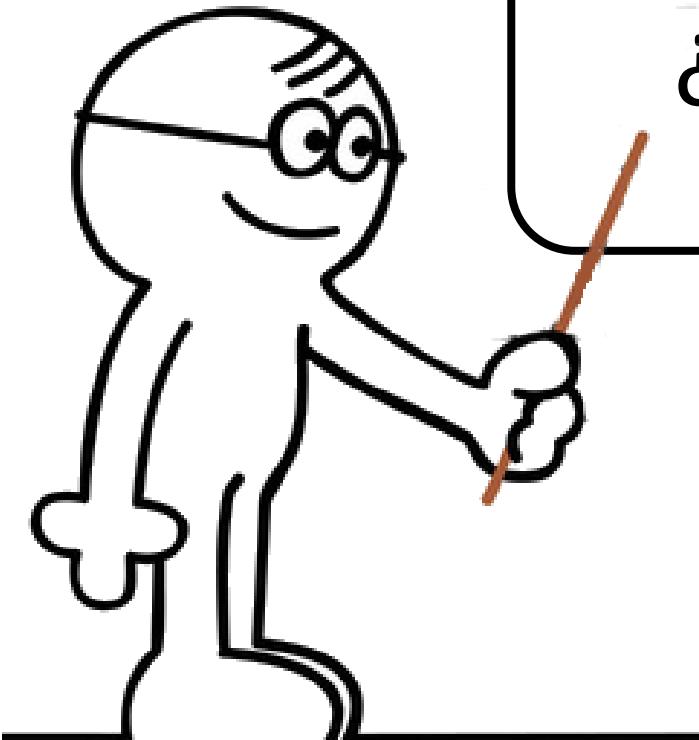
m: The size of the dataset

n: The number of features in the dataset

Tomada de:

<http://www.mostafa.rocks/2017/04/linear-regression-algorithms-in-scikit.html>





¿QUÉ ES LA REGRESIÓN LOGÍSTICA?

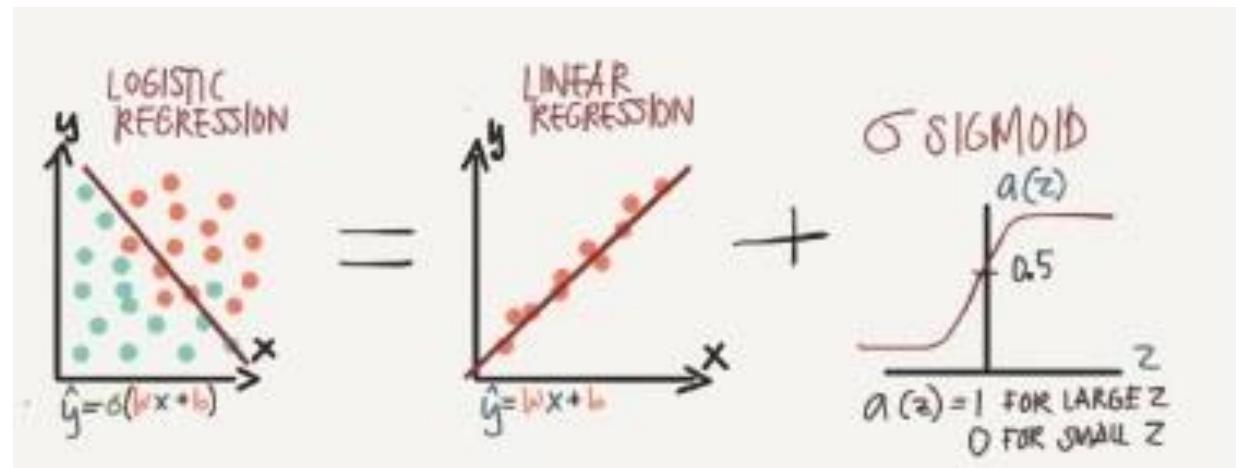
Regresión Logística

DEFINICIÓN:

La regresión logística es un modelo de clasificación que se utiliza para predecir la probabilidad $P(y = 1)$ de una variable dependiente categórica en función de x . Así, la variable y es una variable binaria codificada como 1 (positivo, éxito, etc.) o 0 (negativo, falla, etc.).

Algunos ejemplos de aplicación:

- E-mail: spam/no spam
- Transacciones en línea: fraude/no fraude
- Tumores: maligno/no maligno



Regresión Logística

■ Funcionamiento:

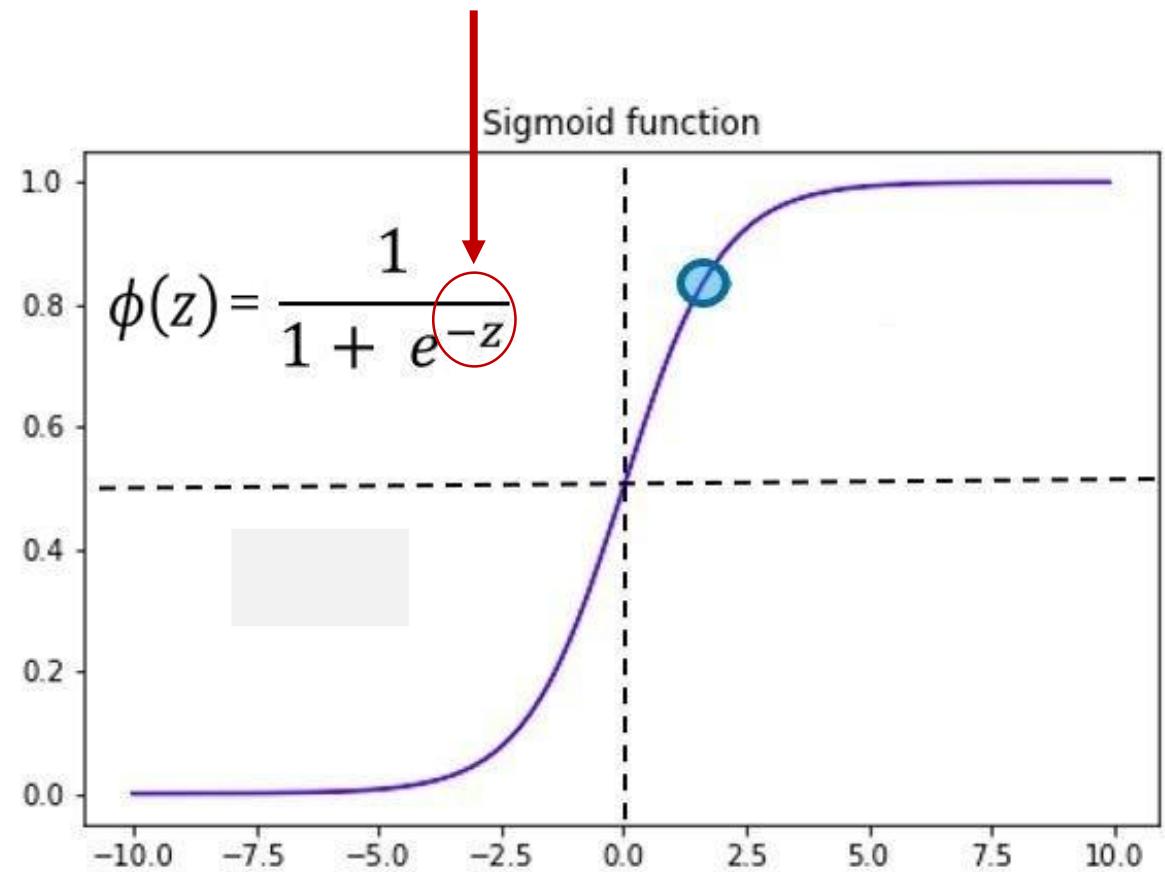
La probabilidad de que cierta muestra pertenezca a una clase particular, que es el objeto de nuestro interés, es la forma inversa de la función *logit*.

Esta función es llamada función logística o sigmoide, y se define como:

$$\phi(z) = \frac{1}{1 + e^{-z}}$$

donde z es la combinación lineal de los pesos y las características, es decir, $z = W^T x$.

Esta es la salida del modelo de regresión lineal



Regresión Logística

- Funcionamiento:

La salida de la función sigmoide puede ser interpretada como la **probabilidad** de una observación particular de pertenecer a la clase 1, dadas sus características **x** parametrizadas por los pesos **W**.

$$\phi(z) = P(y = 1|x; W)$$

- Ejemplo: Diagnóstico de cáncer

$$x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumorSize} \end{bmatrix}$$

Para un tamaño de tumor definido el modelo tiene como salida:

$$\phi(z) = 0,7$$

Esto nos indica que el paciente tiene 70% de probabilidad de que el tumor sea maligno por su tamaño

Regresión Logística

■ Funcionamiento:

Para estimar los coeficientes \mathbf{W} se debe minimizar una función de costo. En nuestro caso la función de costo es el error es la suma de errores al cuadrado:

$$J(\mathbf{W}) = \frac{1}{2} \sum_{i=1}^n (\phi(z) - y_i)^2$$

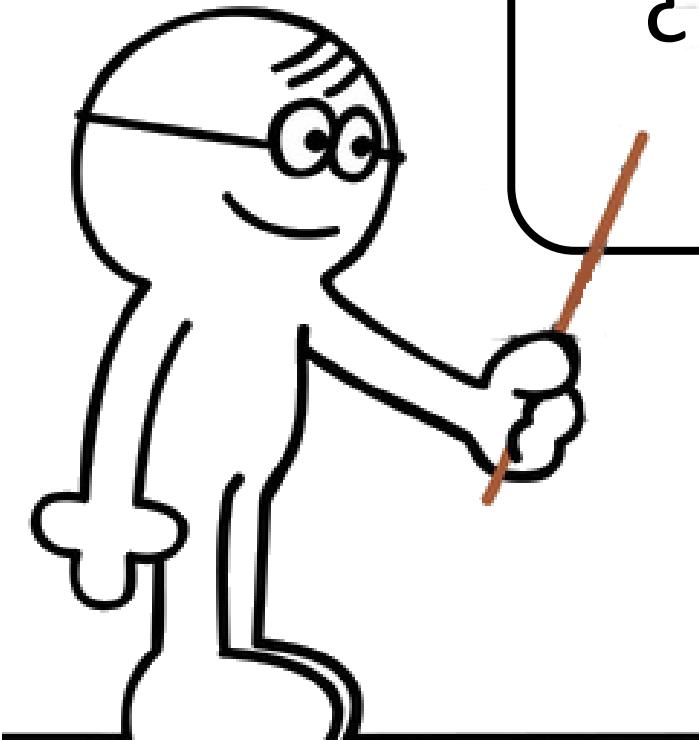
Este error corresponde a la diferencia que hay entre la clase estimada y la etiqueta real de cada observación.

Ahora bien, para la regresión logística la función a minimizar queda definida como:

$$J(\mathbf{W}) = \sum_{i=1}^n -y_i \log(\phi(z_i)) - (1 - y_i) \log(1 - \phi(z_i))$$

Es la predicción hecha por el modelo





¿EN QUÉ CONSISTE UN CLASIFICADOR
INGENUO DE BAYES?

Teorema de Bayes

LA REGLA DE BAYES

El **Teorema de Bayes** expresa la probabilidad *a posteriori* de un evento aleatorio A (que es una clase c_i) dado un evento B (que es el vector de características, x) en términos de la distribución de probabilidad condicional y la probabilidad marginal.

GAUSSIAN NAIVE BAYES CLASSIFIER

"Gaussian" because this is a normal distribution

This is our prior belief

$$P(\text{class} \mid \text{data}) = \frac{P(\text{data} \mid \text{class}) \times P(\text{class})}{P(\text{data})}$$

We don't calculate this in naive bayes classifiers

ChrisAlbon

Teorema de Bayes

$$P(c_i|x) = \frac{p(x|c_i)P(c_i)}{p(x)}$$

LA REGLA DE BAYES

- Para un problema de 2 clases (A y B):

$$p(A|x) > p(B|x) \rightarrow A \text{ else } B$$

$$\frac{p(x|A) p(A)}{p(x)} > \frac{p(x|B) p(B)}{p(x)} \rightarrow A \text{ else } B$$

$$p(x|A) p(A) > p(x|B) p(B) \rightarrow A \text{ else } B$$

- Así un clasificador g se define como: $g(x) = \begin{cases} 1 & \text{si } P(c_1|x) > P(c_2|x) \\ 2 & \text{en otro caso} \end{cases}$

- Para múltiples clases el clasificador se define como $g(x) = \arg \max_{c_i} (P(x|c_i)P(c_i))$

Naïve Bayes

- **Características:**

- Es un clasificador lineal, simple y eficiente.
- Su modelo probabilístico se basa en el Teorema de Bayes
- El adjetivo de “ingenuo” viene de la suposición de que las características son mutuamente independientes (i.i.d.).
- En la práctica, la suposición de independencia se viola frecuentemente, pero este clasificador tiene un buen desempeño aún bajo esta suposición, especialmente para tamaños pequeños de muestra.



Por ejemplo: una fruta puede ser considerada como una manzana si es roja, redonda y de alrededor de 7 cm de diámetro.

Este clasificador considera que cada característica contribuye de manera independiente a la probabilidad de que esta fruta sea una manzana, independientemente de la presencia o ausencia de las otras características.

Naïve Bayes

■ Definiciones:

- Para el clasificador de Bayes ingenuo, la independencia significa que la probabilidad de una observación no afecta la probabilidad de otra observación. Esto lleva a la probabilidad de clase condicional puede calcularse como:

$$P(x|c_j) = \prod_{k=1}^d P(x_k|c_j)$$

Donde $P(x|c_j)$ significa ¿Qué tan probable es observar este patrón particular x dado que pertenece a la clase c_j ?

■ Definiciones:

- Las verosimilitudes individuales para cada característica pueden estimarse vía estimación de máxima verosimilitud, lo que es una simple frecuencia en el caso de datos categóricos:

$$\hat{P}(x_i|c_j) = \frac{N_{x_i,c_j}}{N_{c_j}}$$

donde N_{x_i,c_j} es el número de veces que la característica x_i aparece en las observaciones de clase c_j , y N_{c_j} es el conteo total de todas las características en la clase c_j

Naïve Bayes

■ Definiciones:

- La probabilidad *a priori* también es llamada *prior* de clase y describe la probabilidad general de encontrar una clase particular. Si los *priors* siguen una distribución uniforme, las probabilidades posteriores estarán determinadas por completo por la probabilidad de clase condicional y por la evidencia.

■ Definiciones:

- Eventualmente, el conocimiento *a priori* puede ser determinado a través del conjunto de entrenamiento, si se asume que los datos de entrenamiento son i.i.d y que son una muestra representativa de toda la población.

$$\hat{P}(c_j) = \frac{N_{c_j}}{N}$$

donde N_{c_j} es el número de muestras de la clase c_j y N es el número total de muestras.

Naïve Bayes

■ Definiciones:

- La evidencia $P(x)$ puede entenderse como la probabilidad de encontrar un patrón particular x independientemente de la etiqueta de clase. Usualmente se puede eliminar de la regla de decisión, debido a que suele ser común a todos los términos.

■ Funcionamiento del Clasificador:

- **Paso 1:** Calcule la probabilidad a priori para las cada una de las clases
- **Paso 2:** Estime la probabilidad $\hat{P}(x_i|c_j)$ para cada característica dada cada clase usando un estimador de máxima verosimilitud
- **Paso 3:** Use la fórmula de Bayes y calcule la probabilidad posterior.
- **Paso 4:** Determine cuál de las clases tiene la probabilidad más alta, y asigne esa clase a un dato particular x .

Naïve Bayes

■ Ejemplo:

- Necesitamos determinar la probabilidad si un jugador jugará un partido con base en las condiciones climáticas. Para ello tenemos los siguientes datos.
- Lo primero que debemos hacer es crear tres tablas: la tabla de frecuencias por condición climática y las tablas de verosimilitud.

Whether	Play
Sunny	No
Sunny	No
Overcast	Yes
Rainy	Yes
Rainy	Yes
Rainy	No
Overcast	Yes
Sunny	No
Sunny	Yes
Rainy	Yes
Sunny	Yes
Overcast	Yes
Overcast	Yes
Rainy	No

Naïve Bayes

Whether	Play
Sunny	No
Sunny	No
Overcast	Yes
Rainy	Yes
Rainy	Yes
Rainy	No
Overcast	Yes
Sunny	No
Sunny	Yes
Rainy	Yes
Sunny	Yes
Overcast	Yes
Overcast	Yes
Rainy	No



Frequency Table

Whether	No	Yes
Overcast		4
Sunny	2	3
Rainy	3	2
Total	5	9

Likelihood Table 1

Whether	No	Yes		
Overcast		4	=4/14	0.29
Sunny	2	3	=5/14	0.36
Rainy	3	2	=5/14	0.36
Total	5	9		
	=5/14	=9/14		
	0.36	0.64		



$\hat{P}(c_j)$

Likelihood Table 2

Whether	No	Yes	Posterior Probability for No	Posterior Probability for Yes
Overcast		4	0/5=0	4/9=0.44
Sunny	2	3	2/5=0.4	3/9=0.33
Rainy	3	2	3/5=0.6	2/9=0.22
Total	5	9		

$\hat{P}(x_i | c_j)$

Naïve Bayes

- Ejemplo:

- Ahora suponga que desea calcular la probabilidad de jugar cuando el clima está nublado, entonces se utiliza el teorema de Bayes con base en las probabilidades calculadas.

$$P(Yes|Overcast) = \frac{P(Overcast|Yes)P(Yes)}{P(Overcast)}$$

De las tablas tenemos:

$$P(Overcast) = 4/14 = 0.29$$

$$P(Yes) = 9/14 = 0.64$$

$$P(Overcast|Yes) = 4/9 = 0.44$$

$$P(Yes | Overcast) = 0.44 * 0.64 / 0.29 = 0.98$$

