

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

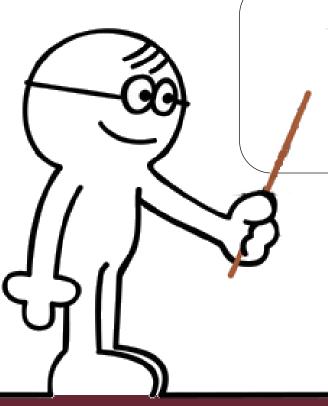
- OBSERVE EL VIDEO Y RESPONDA A LAS SIGUIENTES PREGUNTAS:
 - ¿Cuántos datos se requieren para entrenar un sistema de visión artificial?
 - ¿Es posible decir que los computadores ya sobrepasaron la capacidad humana?
 - ☑ ¿Qué problemas evidencian los sistemas de visión artificial, y en general de los sistemas de Reconocimiento de Patrones?



https://www.ted.com/talks/fei fei li how we re teaching computers to understand pictures?langua@







¿Qué son los Métodos de Aprendizaje de Máquina?



Métodos de Aprendizaje de Máquina

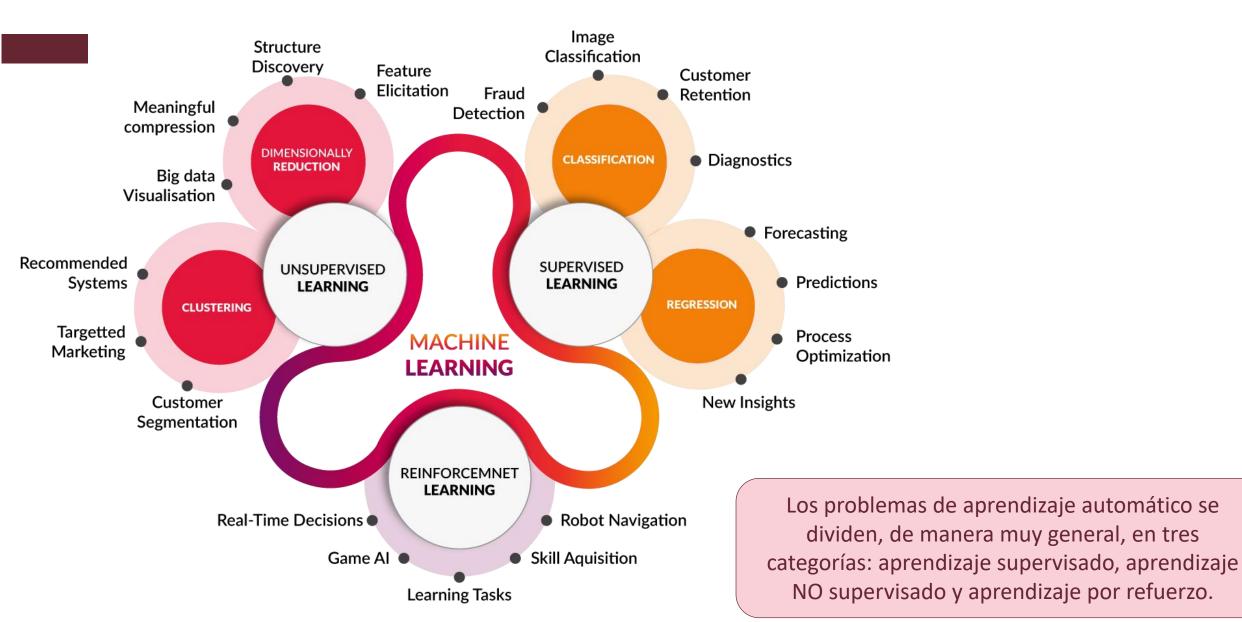
- Los métodos de aprendizaje automático tienen como objetivo crear un modelo de los datos optimizando un criterio de rendimiento, esto a partir de un conjunto de datos de entrenamiento.
- ¿Por qué hacer que las máquinas aprendan?
 - La experiencia humana no existe (navegar en Marte)
 - Los humanos no 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





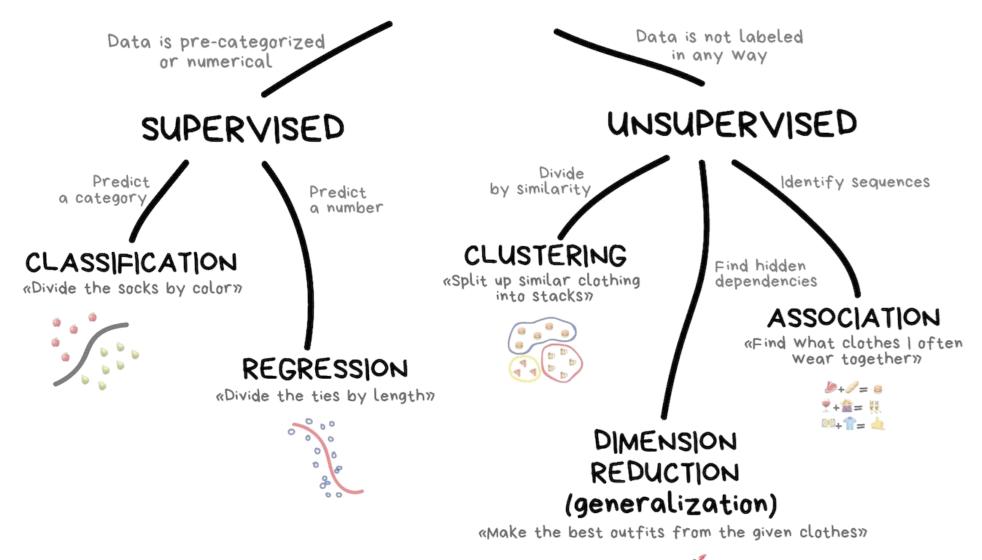


Tomada de: https://www.predictiveanalyticsworld.com/patimes/wp-content/uploads/2018/01/machinelearning-IMAGE.png





CLASSICAL MACHINE LEARNING

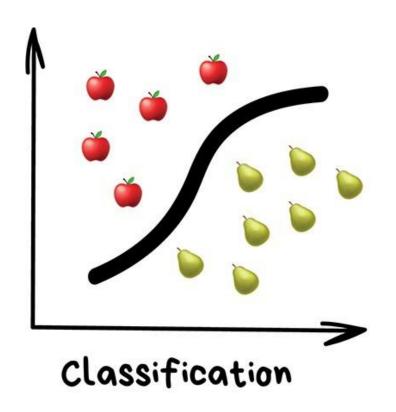


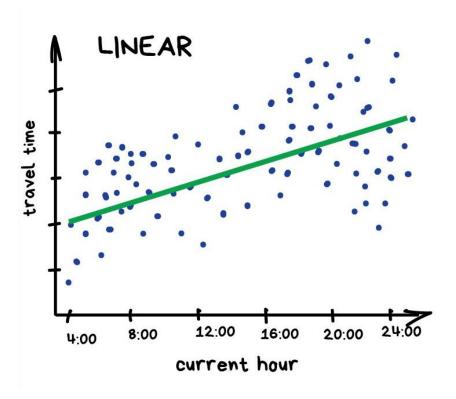
Tomada de: https://vas3k.com/blog/machine learning/

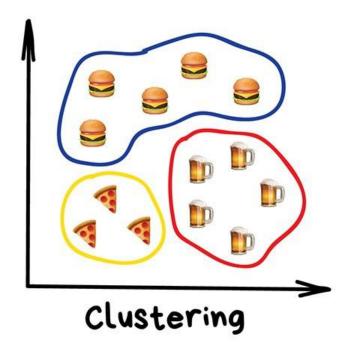




Aprendizaje Supervisado vs No Supervisado











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:

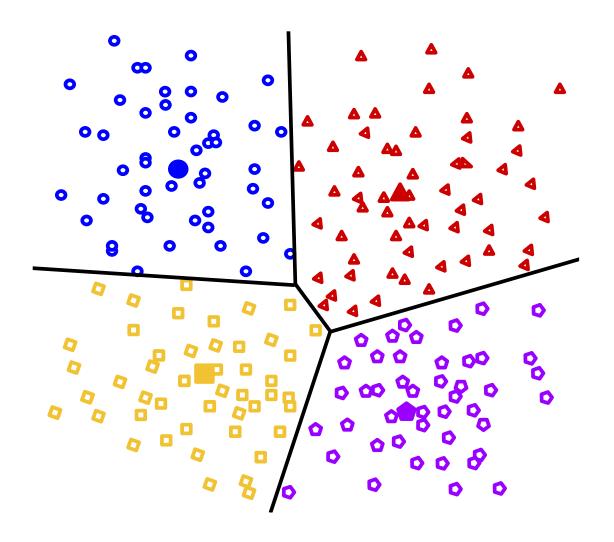
- Clasificación: en este caso la variable de salida es una etiqueta que determina la clase a la que pertenecen los datos de entrada, es decir, la variable de salida es una variable discreta.
- Regresión: en este caso los algoritmos de aprendizaje buscan predecir el valor de una variable continua a partir 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.





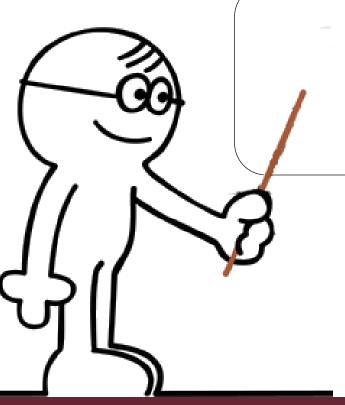
En el **Aprendizaje no supervisado** se cuenta con un conjunto de datos de entrenamiento, pero no hay una variable específica de salida (se desconocen las clases). En este sentido, el objetivo de los problemas del aprendizaje no supervisado es, por ejemplo, el de agrupar los datos de entrada con base 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







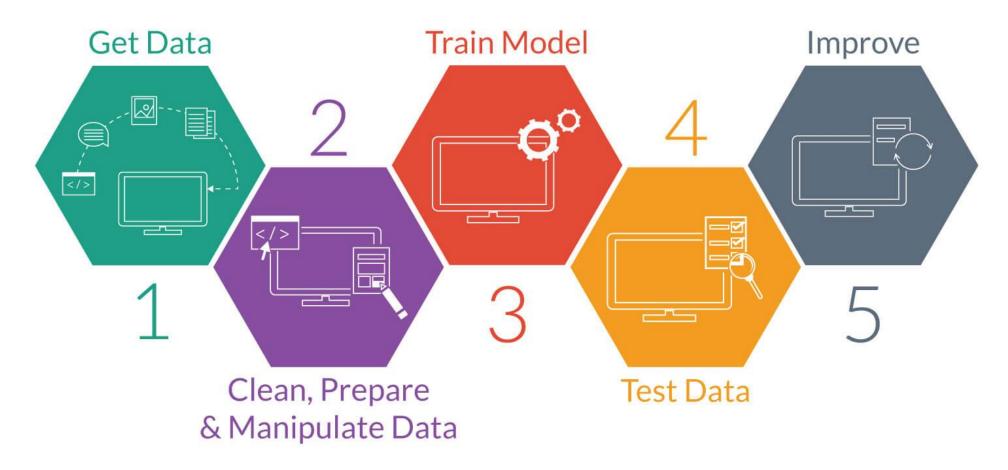


Flujo de Trabajo de los Métodos de Aprendizaje Automático





Flujo de Trabajo

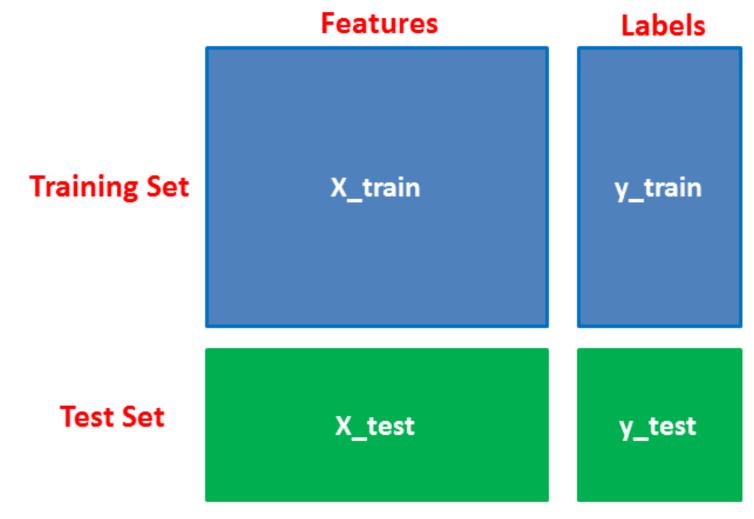


Tomada de: https://newtiummedia.blob.core.windows.net/images/Steps-to-Predictive-Modelling.jpg



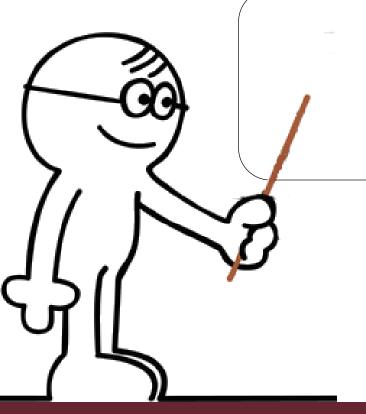


Partición del Conjunto de Datos









¿Qué es la Regresión Lineal?





CORRELACIÓN ENTRE DOS VARIABLES:

Se considera que dos variables cuantitativas ($x \in y$) 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?

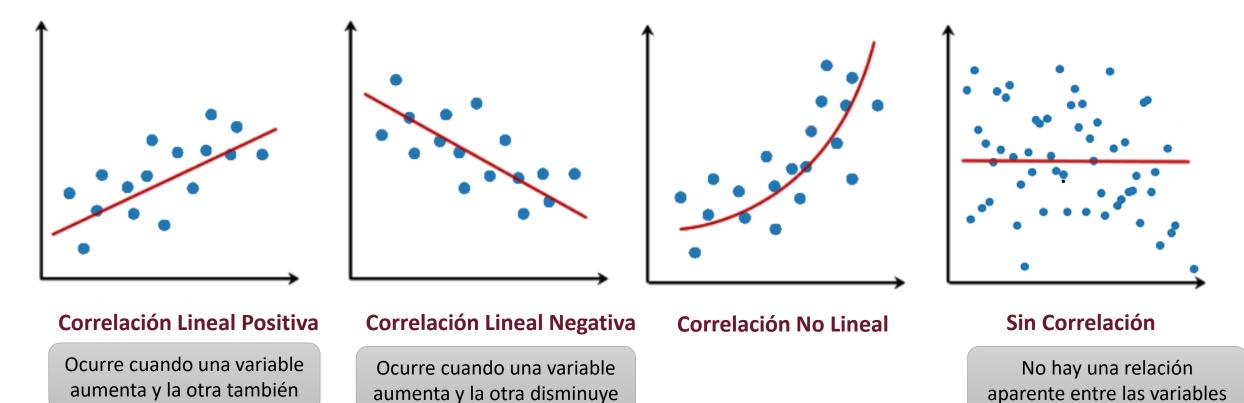
Claro está, si sabemos que la variable x está correlacionada con y, quiere decir que podemos **predecir** la viarble y a partir de x.

Estamos en el terreno de la PREDICCIÓN!



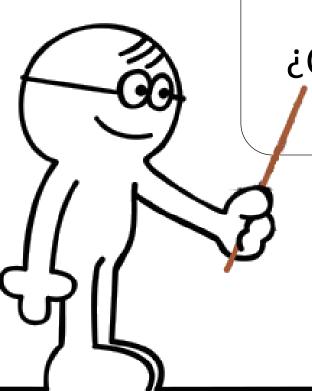
TIPOS DE CORRELACIÓN:

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



Grupo de I+D en Inteligencia Artificial

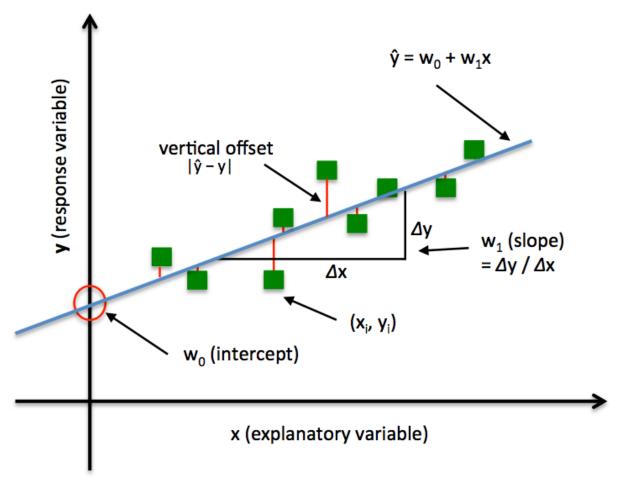




¿CUÁL ES EL PRINCIPIO DE LA REGRESIÓN LINEAL?







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:

$$\hat{y} = w_0 + w_1 x$$

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.

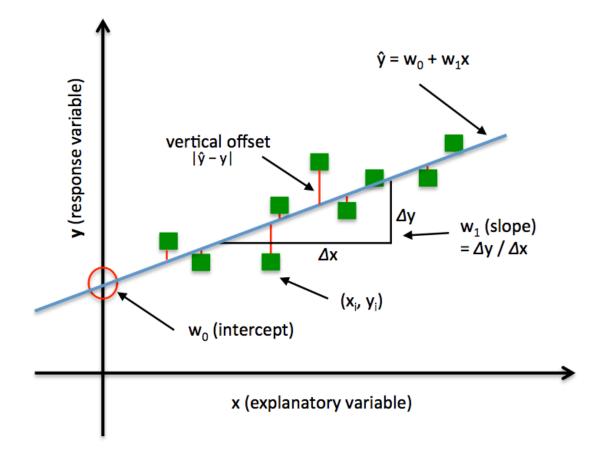




GENERALIDADES:

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 \mathbf{x}_i . Entonces, buscamos aquella línea tal que minimice la función de costo:

Predicción del Modelo
$$J(w) = rac{1}{2} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$
 Valor Observado







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:

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

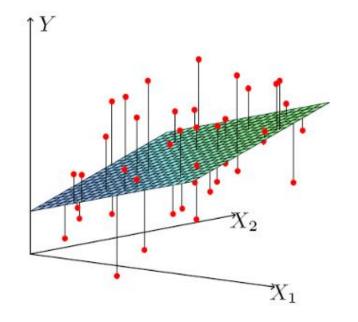
$$\hat{y} = h_{\mathbf{W}}(\mathbf{x}) = \sum_{k=0}^{d} w_k x^k = \mathbf{W}^T \cdot \mathbf{x}$$
, con $x^0 = 1$

Donde,

- h_W es la función de hiperplano
- 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:

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



SOLUCIÓN POR EL MÉTODO TEÓRICO CON ALGEBRA LINEAL:

Para encontrar el valor de \boldsymbol{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:

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



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/





SOLUCIÓN POR EL MÉTODO ESTADÍSTICO:

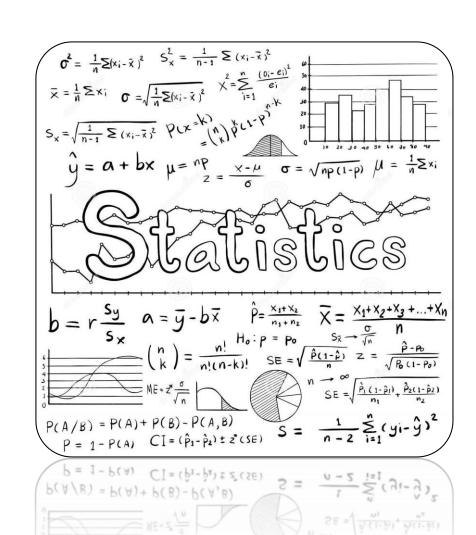
Para encontrar el valor de W, tenemos que:

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

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

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/







SOLUCIÓN POR EL MÉTODO DE OPTIMIZACIÓN:

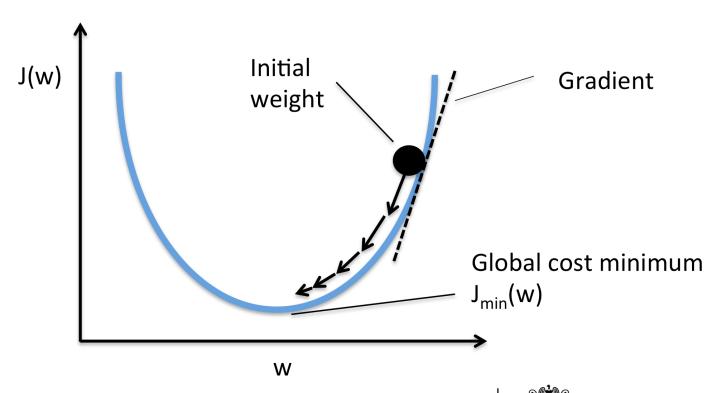
En este caso el valor del parámetro W se busca utilizando métodos de optimización, por ejemplo usando Gradiente Descendente.

En este caso los pesos se actualizan gradualmente después con base en la función de costo, denominada J (·), 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:

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

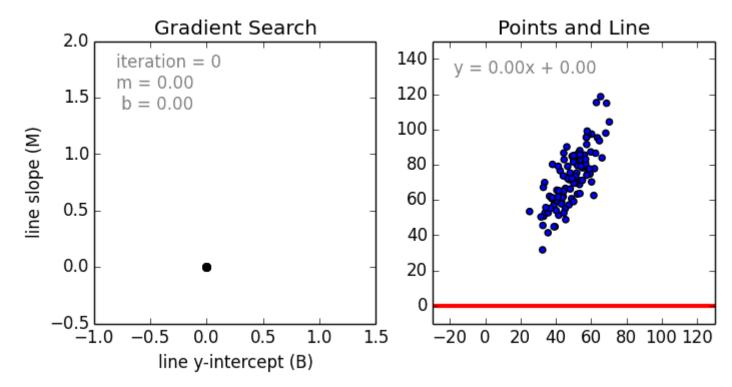
Siendo η la razón (o tasa) de aprendizaje.





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.







■ 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	LinearRegression
Batch Gradient Descent	Slow	No	Fast	2	Yes (Hadoop MapReduce)	n/a
Stochastic Gradient Descent	Fast	Yes	Fast	>=2	Yes	SGDRegressor
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-algorthims-in-scikit.html

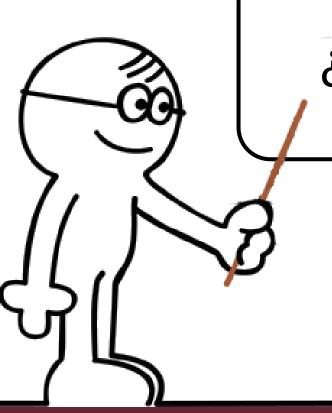












¿Qué es la Regresión Logística?



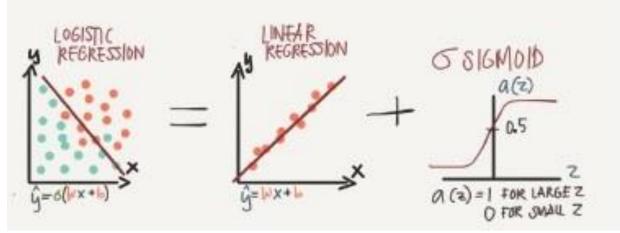


DEFINICIÓN:

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.).

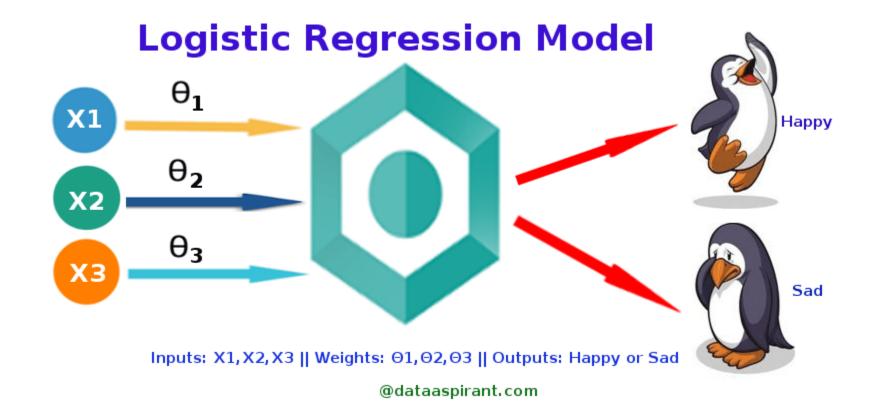
Algunos ejemplos de aplicación:

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













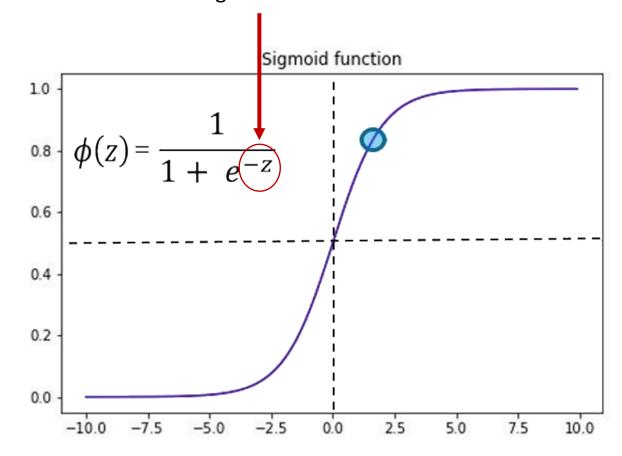
Funcionamiento:

La probabilidad de que **cierta muestra** pertenezca a una clase particular es la inversa de la función *logit*, es decir la función logística o sigmoide, la cual está definida como:

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

donde z es la combinación lineal entre los pesos (w_i) y las características (x^i) , es decir, $z = \mathbf{W}^T \mathbf{x}$.

Esta es la salida del modelo de regresión lineal







Funcionamiento:

La salida de la función sigmoide puede ser interpretada como la **probabilidad** de que una observación particular pertenezca a la Clase 1, dadas sus características x, ponderadas por los pesos W.

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

Con base en esto, a través de la función *logit* obtenemos el modelo lineal:

$$\lograc{p}{1-p}=eta_0+eta_1x_1+eta_2x_2+\cdots+eta_mx_m$$

Ejemplo: Diagnóstico de cáncer

$$\boldsymbol{x} = \left[\begin{array}{c} x_0 \\ x_1 \end{array} \right] = \left[\begin{array}{c} 1 \\ \text{tumorSize} \end{array} \right]$$

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



Funcionamiento:

Para estimar los coeficientes W se debe minimizar una función de costo. En nuestro caso la función de costo, interpretada como la suma de los errores, corresponde al logaritmo de la verosimilitud que es:

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

Es la predicción hecha por el modelo

$$P(y=1|\mathbf{x};\mathbf{W})$$

Este error se puede ver como la diferencia que hay entre la clase estimada y la etiqueta real de cada observación.

VENTAJAS:

- Es un modelo de clasificación eficiente y simple.
- No es necesario disponer de grandes recursos computacionales.
- Los resultados son altamente interpretables.



DESVENTAJAS:

- Imposibilidad de resolver directamente problemas no lineales.
- La variable objetivo esta ha de ser linealmente separable.
- La regresión logística no es uno de los algoritmos más potentes que existen.







