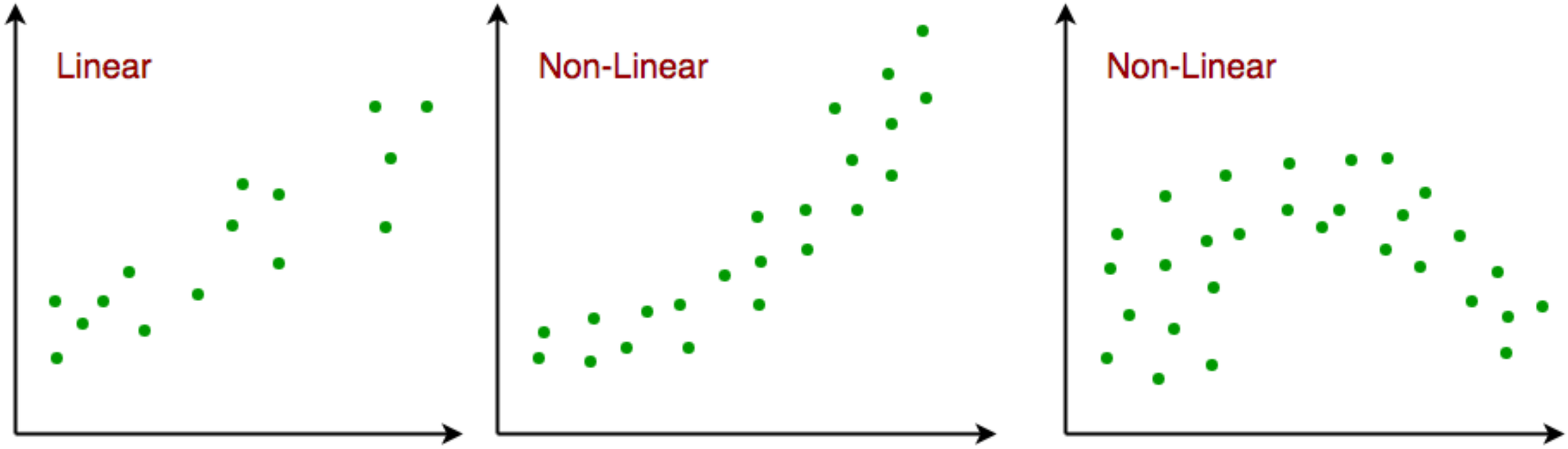
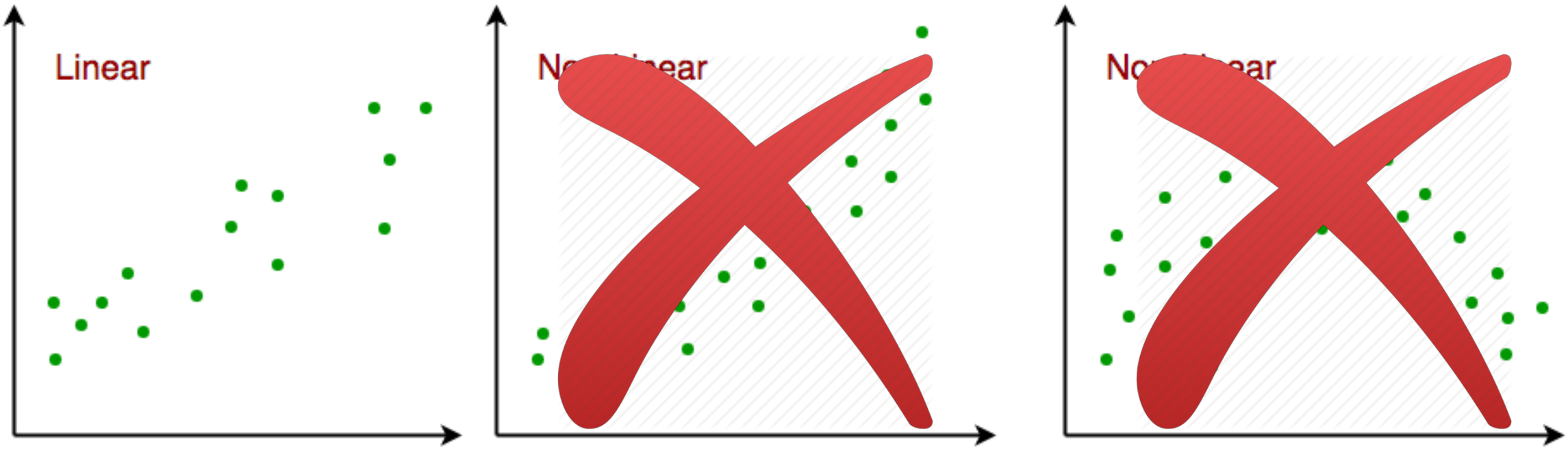
Repaso Regresión

(lineal y logística)



https://cdncontribute.geeksforgeeks.org/wp-content/uploads/python-linear-regression-4.png



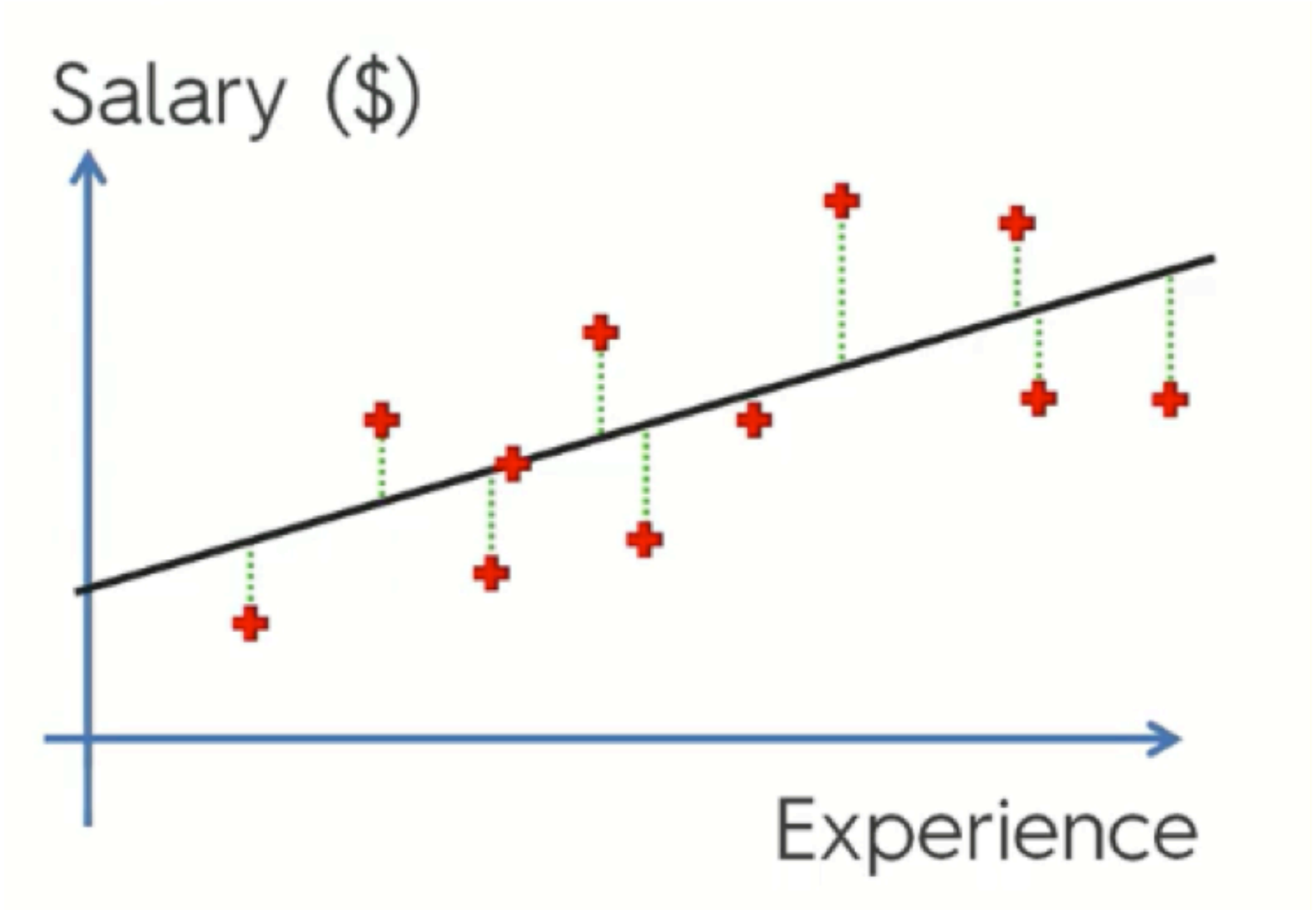
https://cdncontribute.geeksforgeeks.org/wp-content/uploads/python-linear-regression-4.png

Regresión lineal con una variable.

$$y = mx + n$$

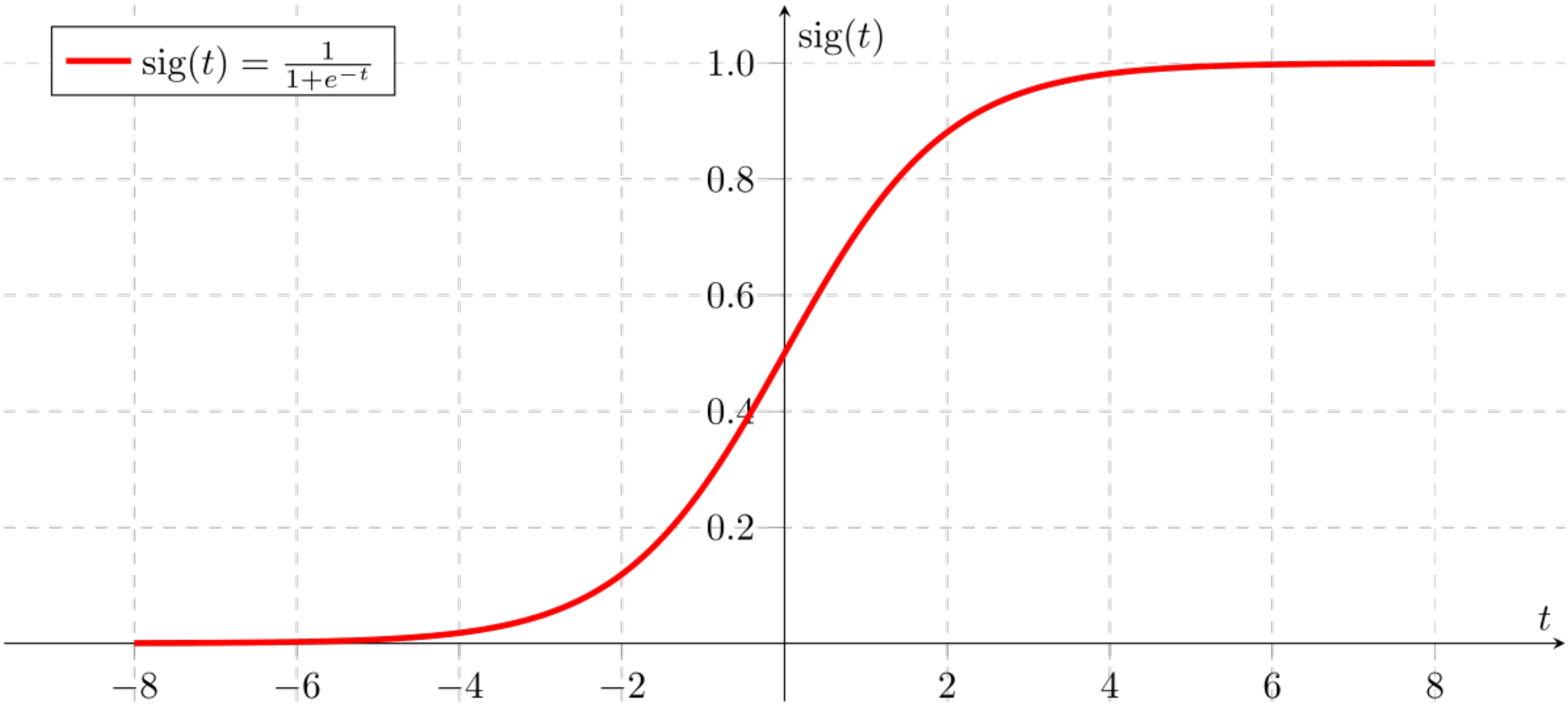
Regresión lineal múltiple.

$$y = n + \sum_{x_i \in x} m_i x_i$$



https://sds-platform-private.s3-us-east-2.amazonaws.com/uploads/37_blog_image_1.png

Función logística (sigmoide)



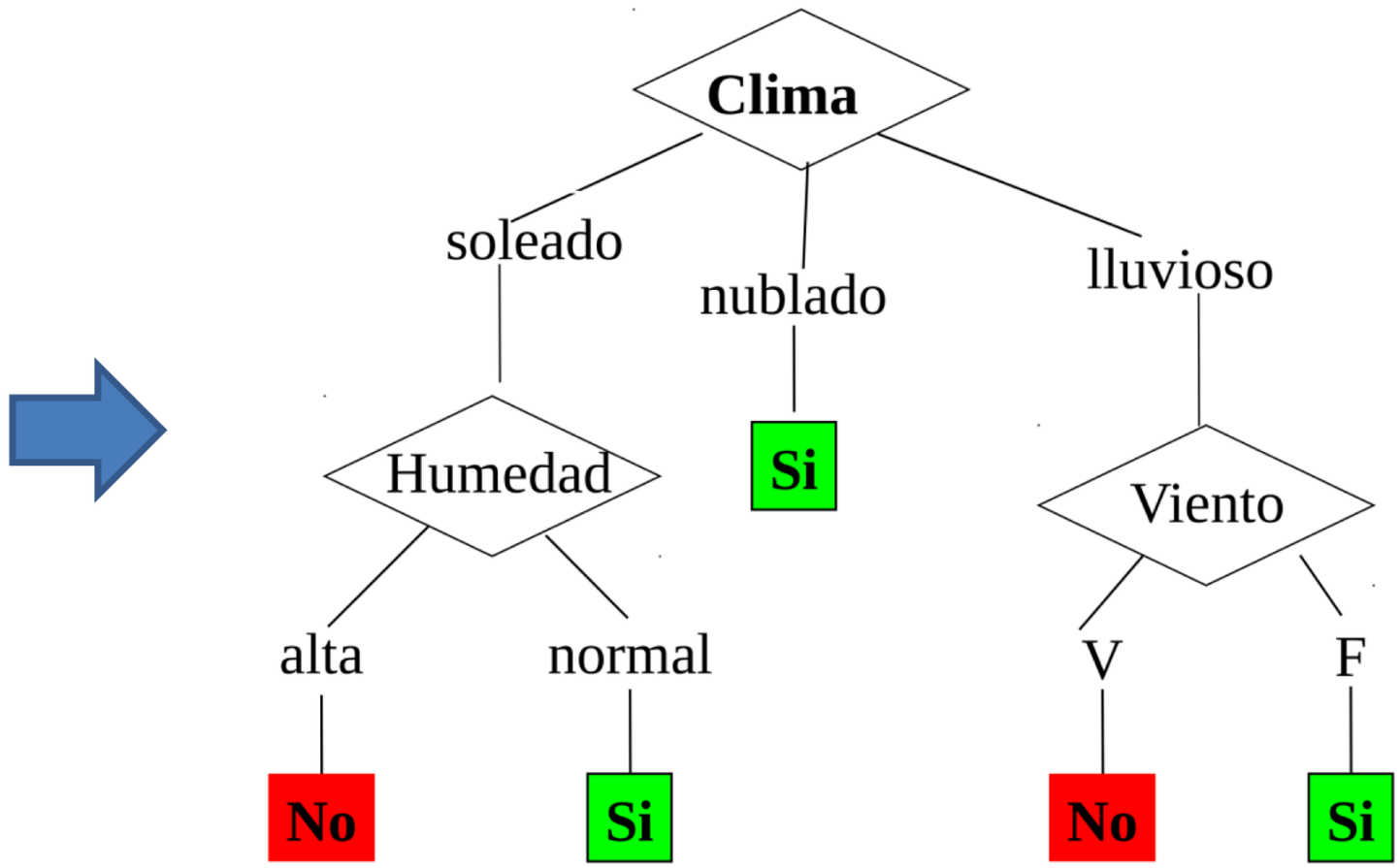
https://cdn-images-1.medium.com/max/1600/1*RqXFpiNGwdiKBWyLJc_E7g.png

Árboles de decisión

(clasificación y regresión)

Clima	Temperatura	Humedad	Viento	Jugar?
soleado	alta	alta	F	No
soleado	alta	alta	V	No
nublado	alta	alta	F	Si
lluvioso	Agradable	alta	F	Si
lluvioso	frio	normal	F	Si
lluvioso	frio	normal	V	No
nublado	frio	normal	V	Si
soleado	Agradable	alta	F	No
soleado	frio	normal	F	Si
lluvioso	Agradable	normal	F	Si
soleado	Agradable	normal	V	Si
nublado	Agradable	alta	V	Si
nublado	alta	normal	F	Si
lluvioso	Agradable	alta	V	No

Clima	Temperatura	Humedad	Viento	Jugar?
soleado	alta	alta	F	No
soleado	alta	alta	V	No
nublado	alta	alta	F	Si
lluvioso	Agradable	alta	F	Si
lluvioso	frio	normal	F	Si
lluvioso	frio	normal	V	No
nublado	frio	normal	V	Si
soleado	Agradable	alta	F	No
soleado	frio	normal	F	Si
lluvioso	Agradable	normal	F	Si
soleado	Agradable	normal	V	Si
nublado	Agradable	alta	V	Si
nublado	alta	normal	F	Si
lluvioso	Agradable	alta	V	No



¿Cómo elijo la variable para el “corte”?

Árboles para clasificación

Entropía

$$H(S) = -\sum_{clases} p_i log_2(p_i)$$

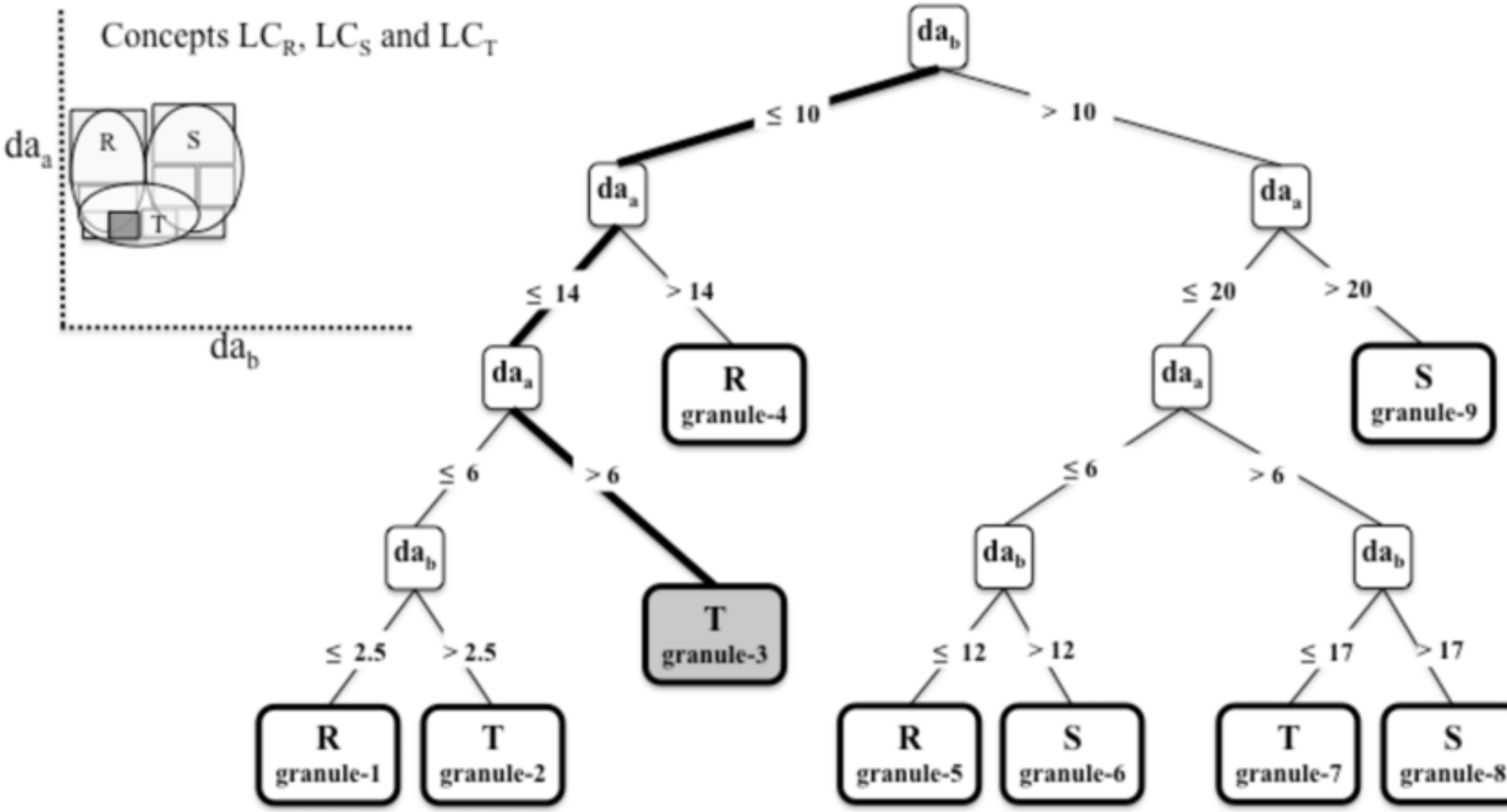
Ganancia de información

$$IG = H(S) - \sum \frac{|S_v|}{|S|} H(s_v)$$

$$GainRatio(S,A) \equiv \frac{Gain(S,A)}{SplitInformation(S,A)}$$

$$SplitInformation(S,A) \equiv -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

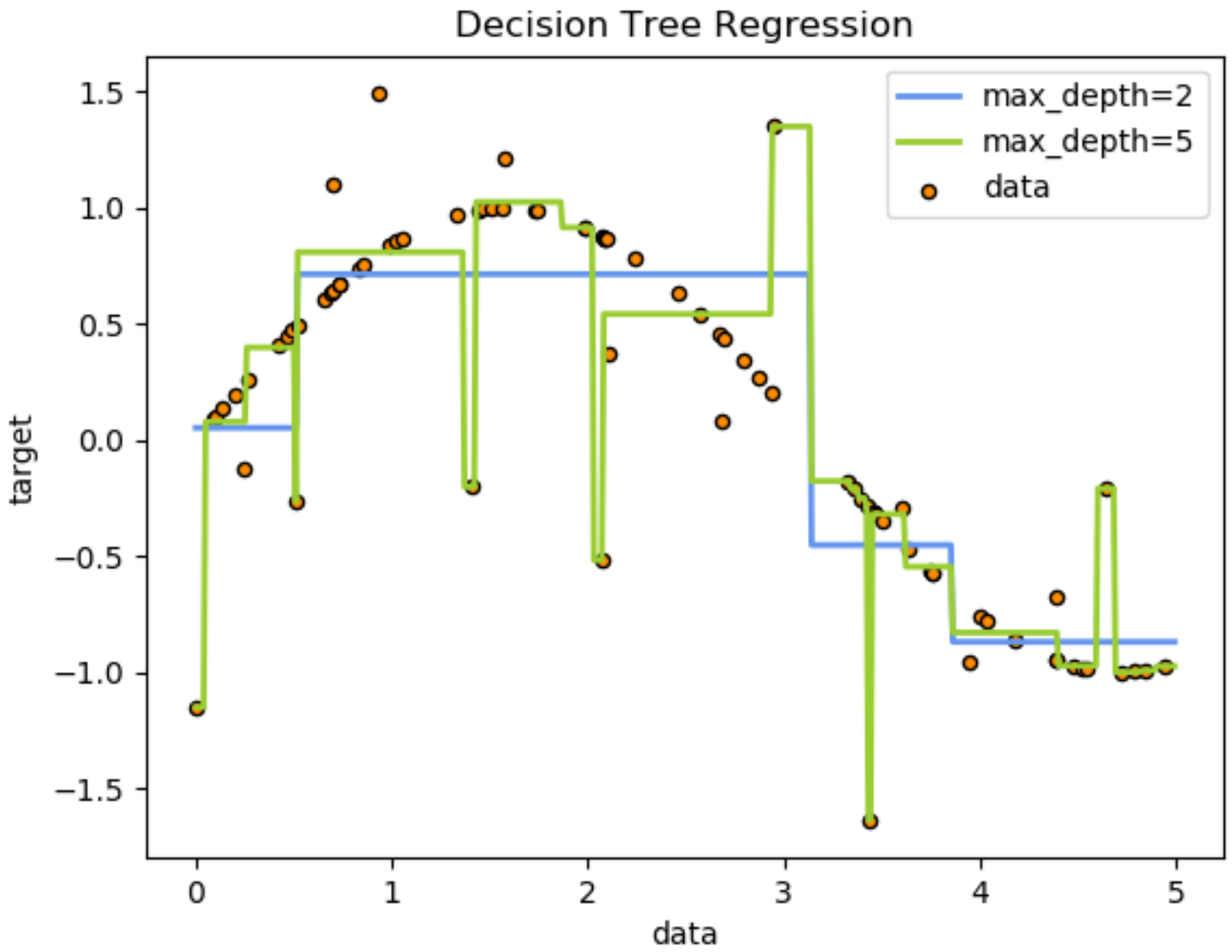
Variables numéricas



https://www.researchgate.net/profile/Bart_Gajderowicz/publication/248703533/figure/fig8/AS:644673399975938@1530713524260/Decision-tree-classification-with-2-numeric-data-attributes-for-sub-classes-of-LC-A.png

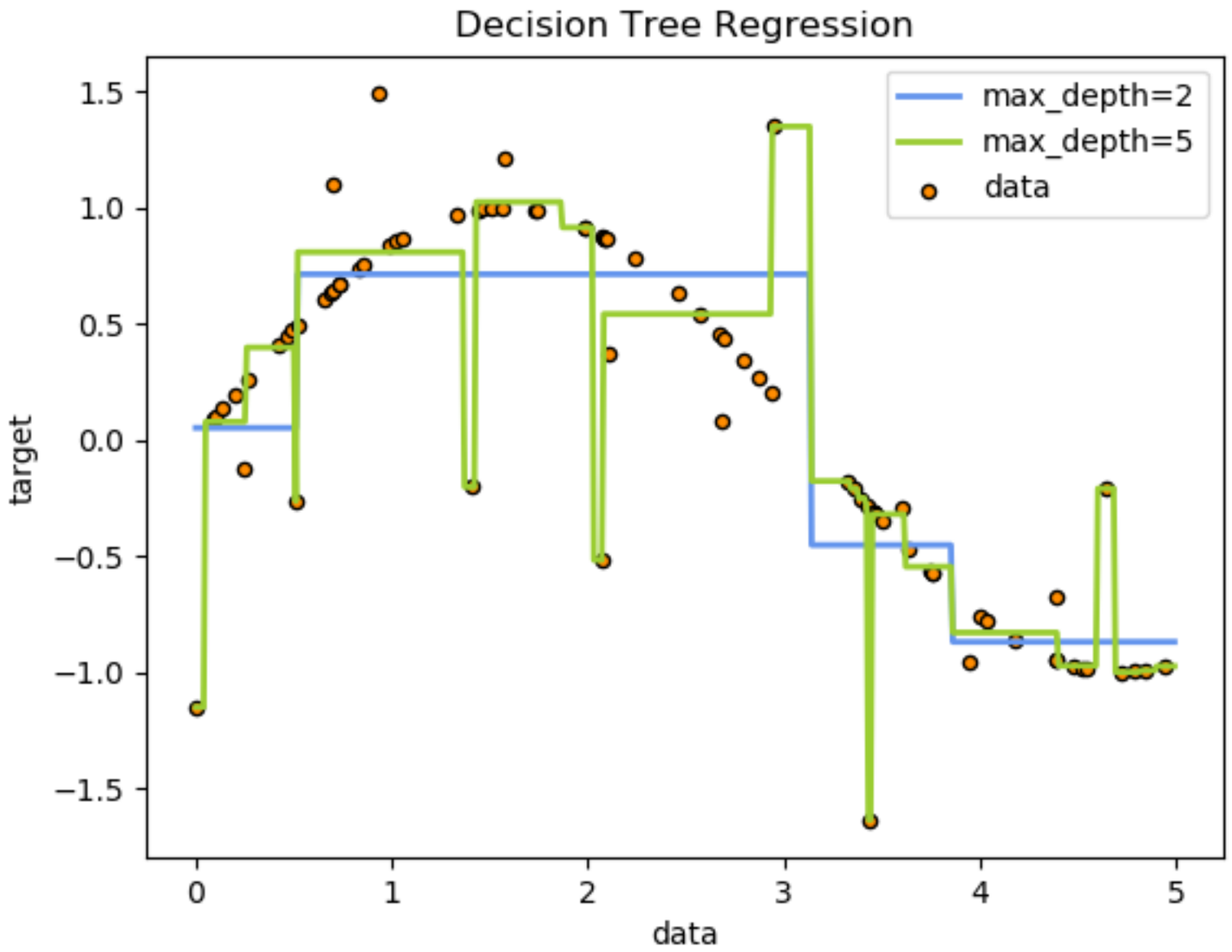
Árboles para regresión

Regresión



https://scikit-learn.org/stable/_images/sphx_glr_plot_tree_regression_0011.png

Elegir variable split



https://scikit-learn.org/stable/_images/sphx_glr_plot_tree_regression_0011.png

Ejercicios

1.4. Si un algoritmo alcanza un 100% de efectividad en el set de entrenamiento, entonces se garantiza que la hipótesis respectiva tendrá un alto grado de generalidad para clasificar nuevas instancias.

1.9. Si un modelo lineal y otro cuadrático modelan igualmente bien los datos, uno debería preferir el cuadrático.

1.4. El algoritmo visto en clases para explorar el espacio de hipótesis de un árbol de decisión garantiza encontrar el árbol que presenta el mejor rendimiento posible en el set de entrenamiento.

1.11. Sin contar los nodos hoja, un árbol de decisión que es entrenado con un set de N atributos puede tener como máximo N nodos en el árbol resultante.

1.3. En un set de datos S, los atributos A, B y C son usados para construir un árbol de decisión que permite predecir la clase de un atributo binario D. Después de calcular la ganancia de información, G(S,Atributo), se obtiene que G(S,A)=0.1, G(S,B)=0.3 y G(S,C)=0.25, por tanto el atributo B es usado como nodo raíz del árbol.

1.10. En cierto problema de aprendizaje de máquina con árboles de decisión, se decide extender una de las ramas de un árbol D1 agregando un nodo adicional que aplica un nuevo test sobre uno de los atributos originales del problema (obs: se extiende sólo una rama), con lo cual se obtiene un árbol D2. De acuerdo a lo anterior, podemos decir que el árbol de decisión D2 tiene un espacio de hipótesis mayor al árbol D1.

a) ¿Como se podría realizar clasificación con un árbol de decisión, si falta el valor de alguna de las dimensiones del vector de entrada? **(1 pto.)**

e) ¿En qué situaciones es preferible utilizar el radio de ganancia por sobre la ganancia de información?

a) Considere un problema de clasificación sobre variables categóricas. Extienda el algoritmo de construcción de los árboles de decisión basado en la ganancia de información, para que se puedan realizar tests sobre dos variables de manera simultánea **(2 ptos.)**.

b) En general, al momento de decidir el valor a testear en un nodo de un árbol de decisión, se toma la ganancia de información como métrica. Una de las desventajas de esta, es que no considera la nueva estructura del árbol en el cálculo (la resultante de seleccionar ese test, con distinta profundidad y número de nodos), lo que puede derivar en problemas de sobreentrenamiento. Extienda la métrica de la ganancia de información, agregando un nuevo término aditivo, de manera que ahora, para tomar la decisión, se considere información sobre la posible nueva estructura del árbol. **Hint:** considere la decisión en un nodo como la minimización del riesgo estructural empírico. **(2 ptos.)**

c) Considere una competencia, donde se debe resolver un problema de regresión en base a variables categóricas. Dado que en este problema el riesgo de sobreentrenamiento es alto, sólo se permite utilizar árboles de regresión sobre una (1) de las variables disponibles, con el fin de limitar la profundidad del árbol. Utilizando múltiples árboles de **manera secuencial** (cada árbol puede utilizar la variable que quiera), indique como es posible construir un sistema de regresión que estime de mejor manera la función buscada. **(2 ptos.)**