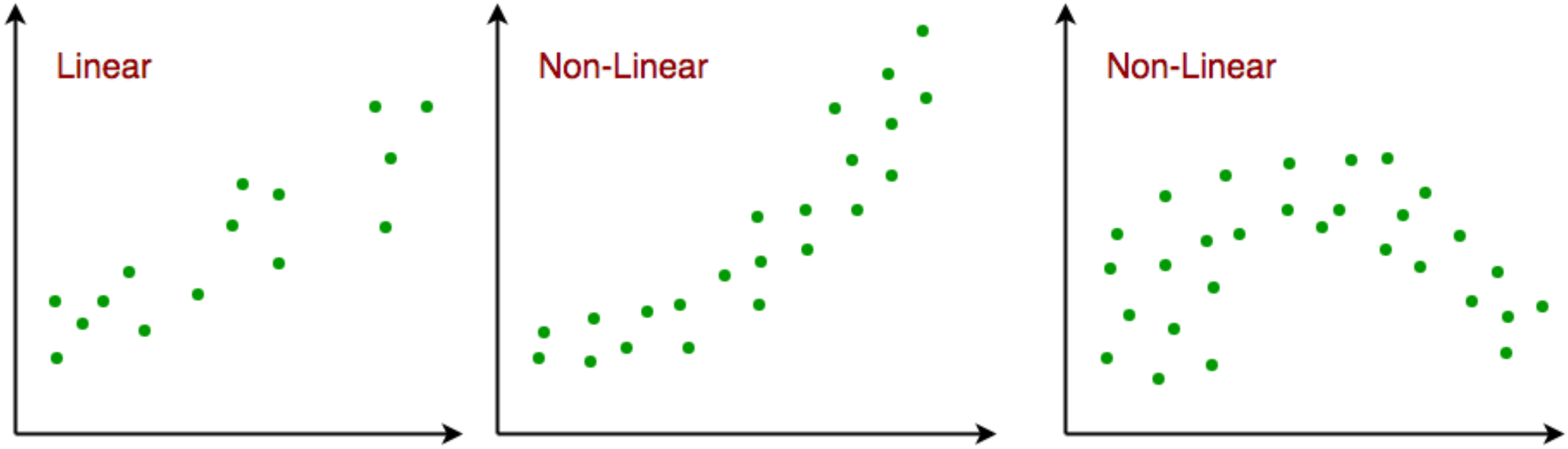
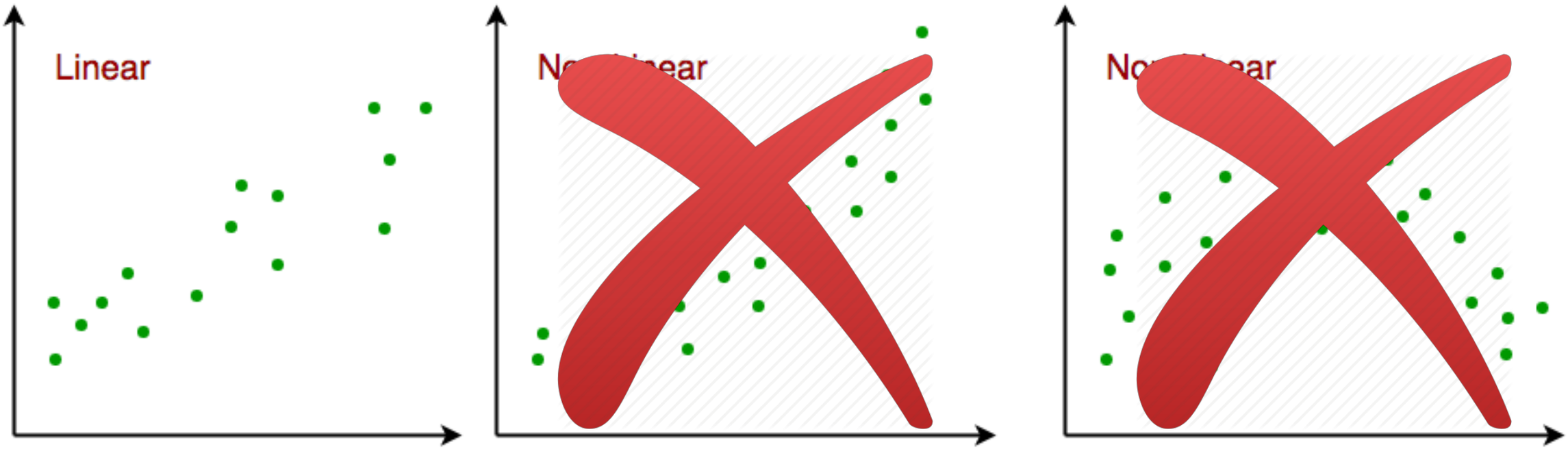
Repaso Regresión

(lineal y logística)



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



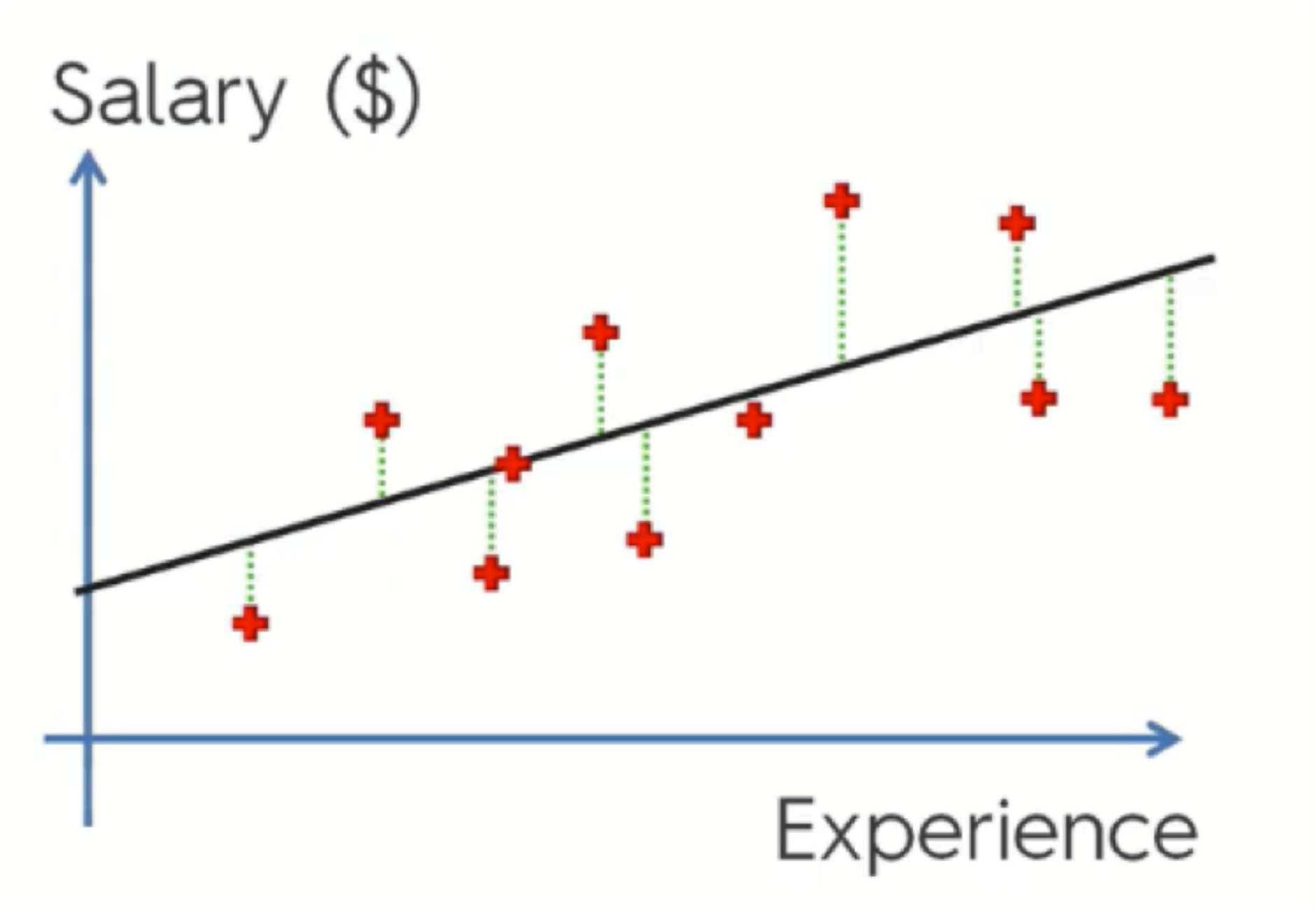
https://cdncontribute.geeksforgeeks.org/wp-content/uploads/python-linear-regression-4.png

Regresión lineal con una variable.

$$y = mx + n$$

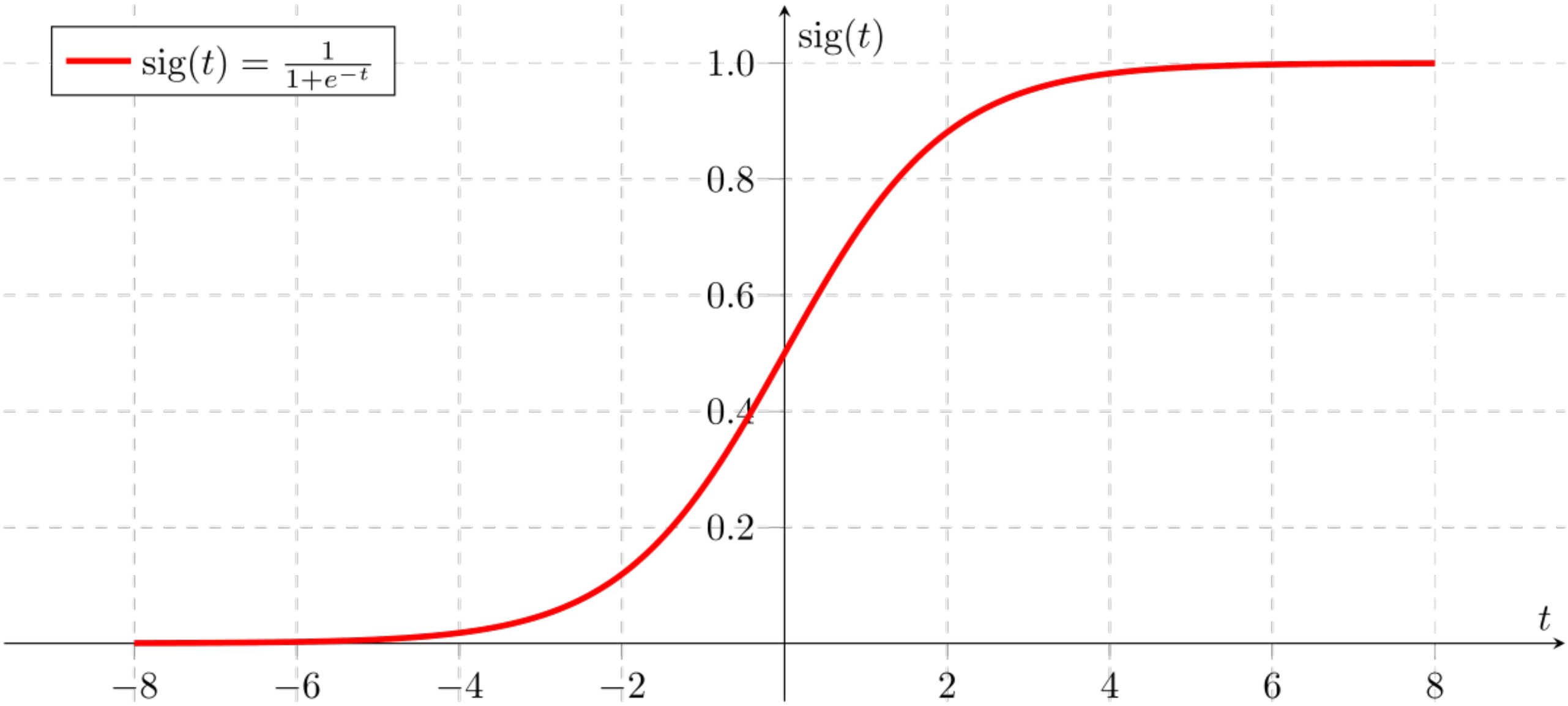
Regresión lineal múltiple.

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



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

Función logística (sigmoide)



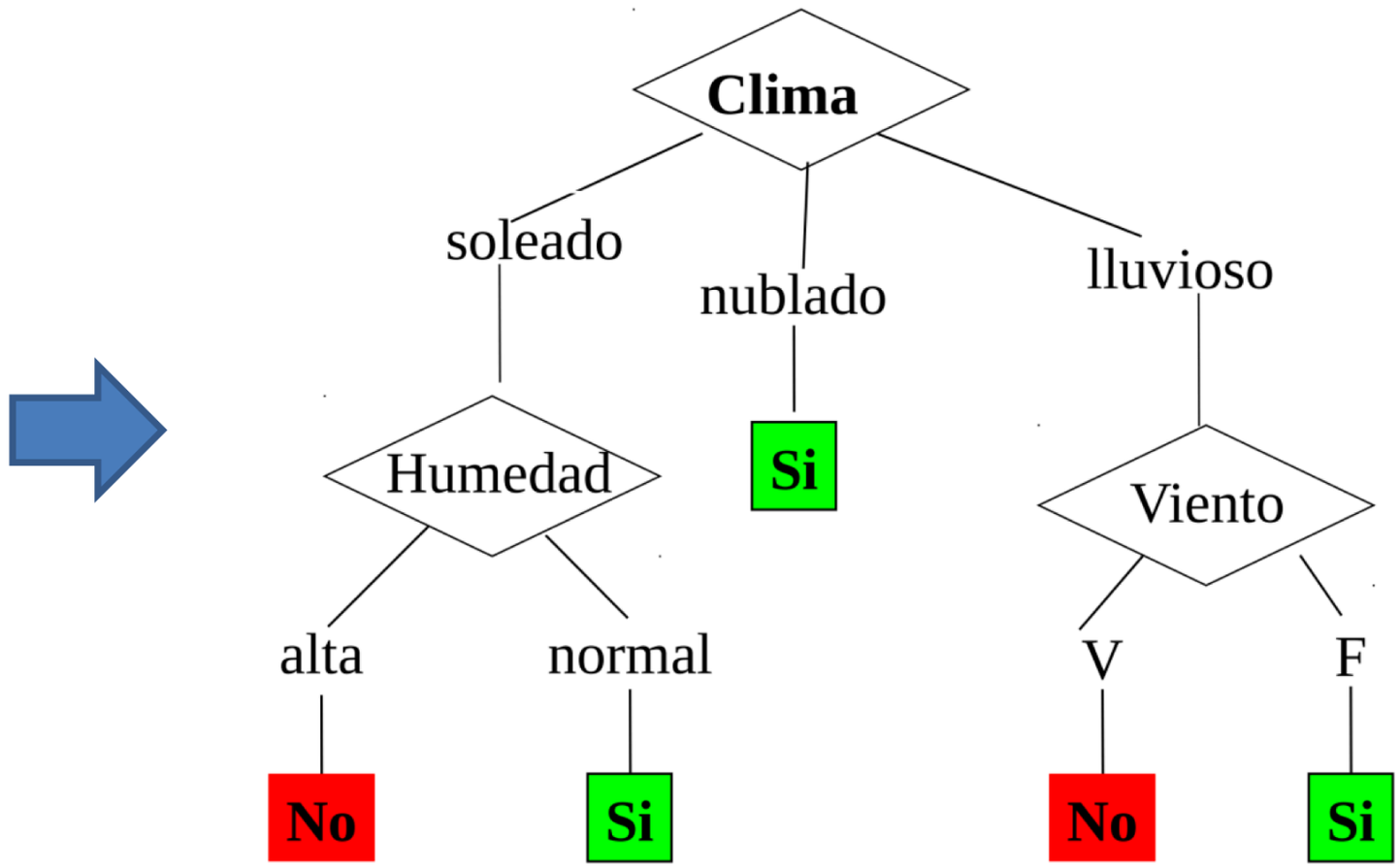
https://cdn-images-1.medium.com/max/1600/1*RqXFpiNGwdiKBWyLJc_E7g.png

Repaso 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

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

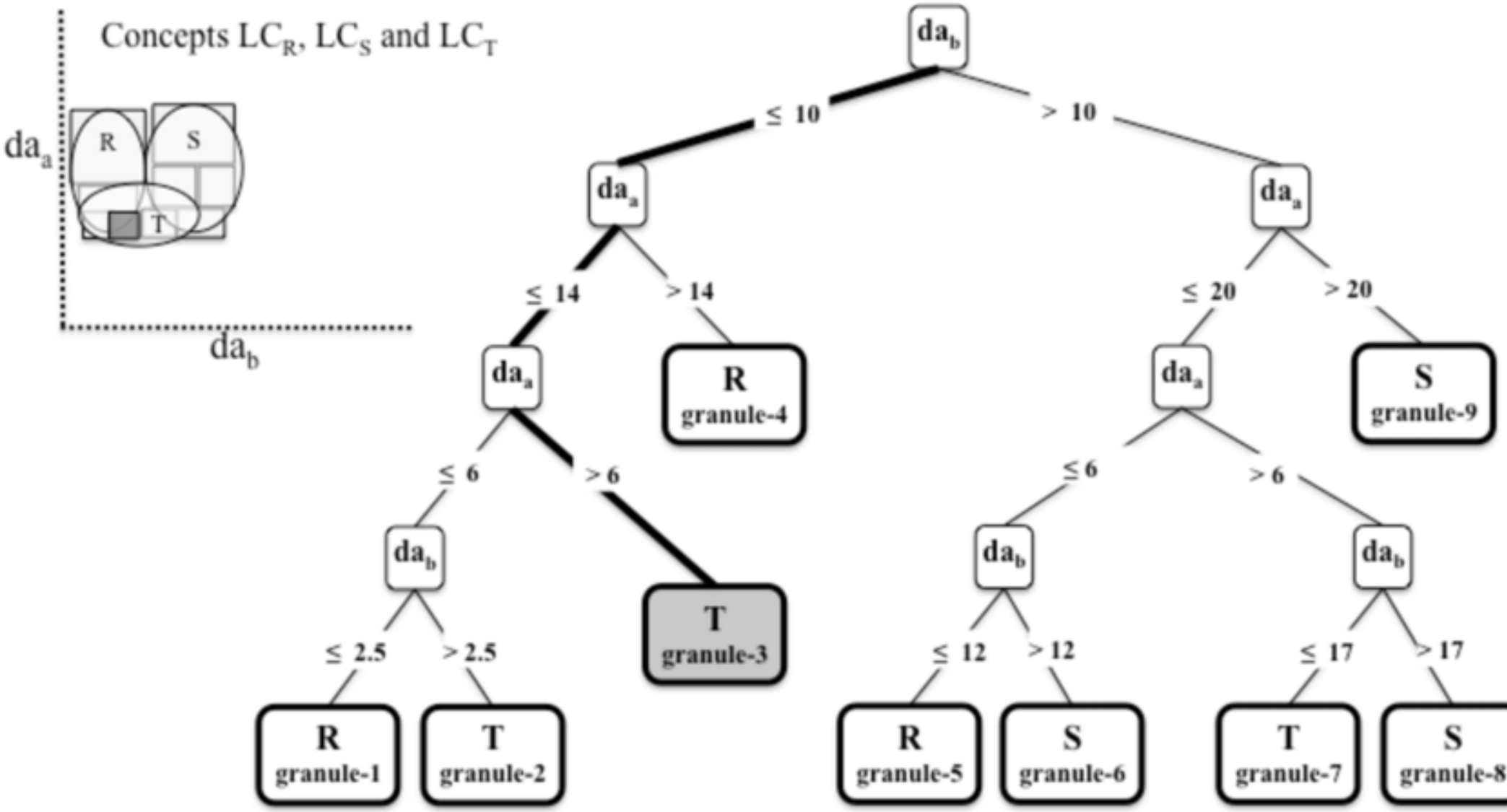
Ganancia de información

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

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

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

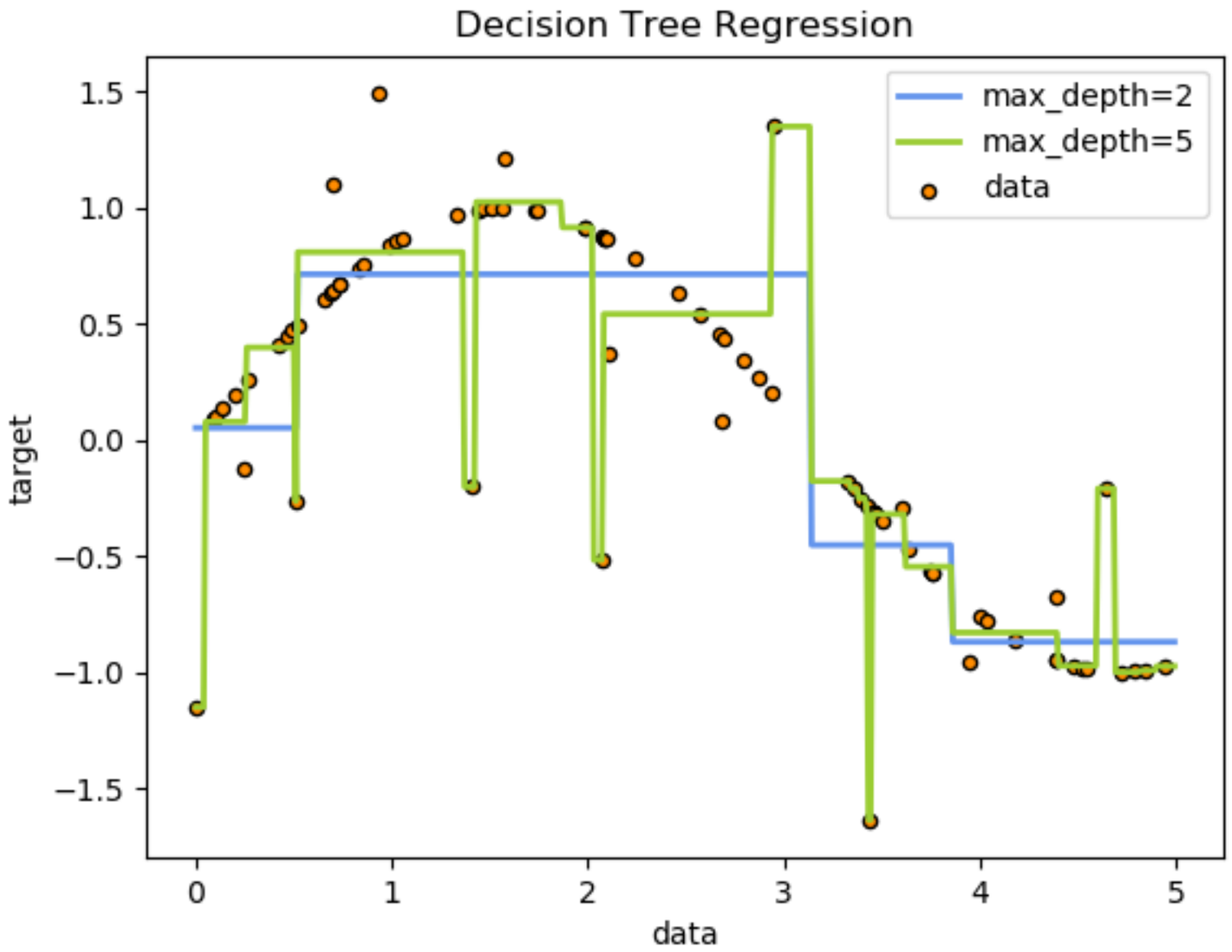
Variables numéricas



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

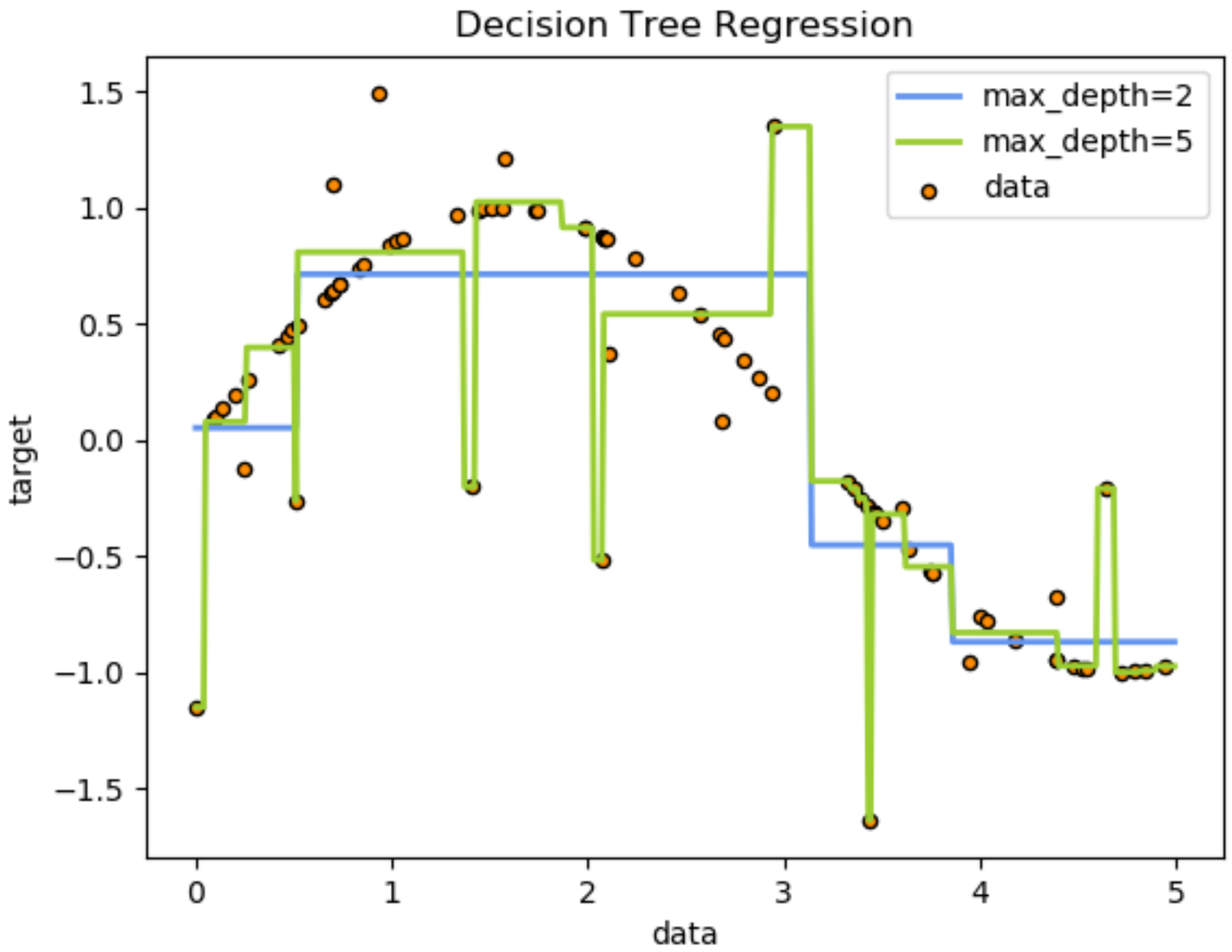
Árboles para regresión

Regresión



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

Elegir variable split

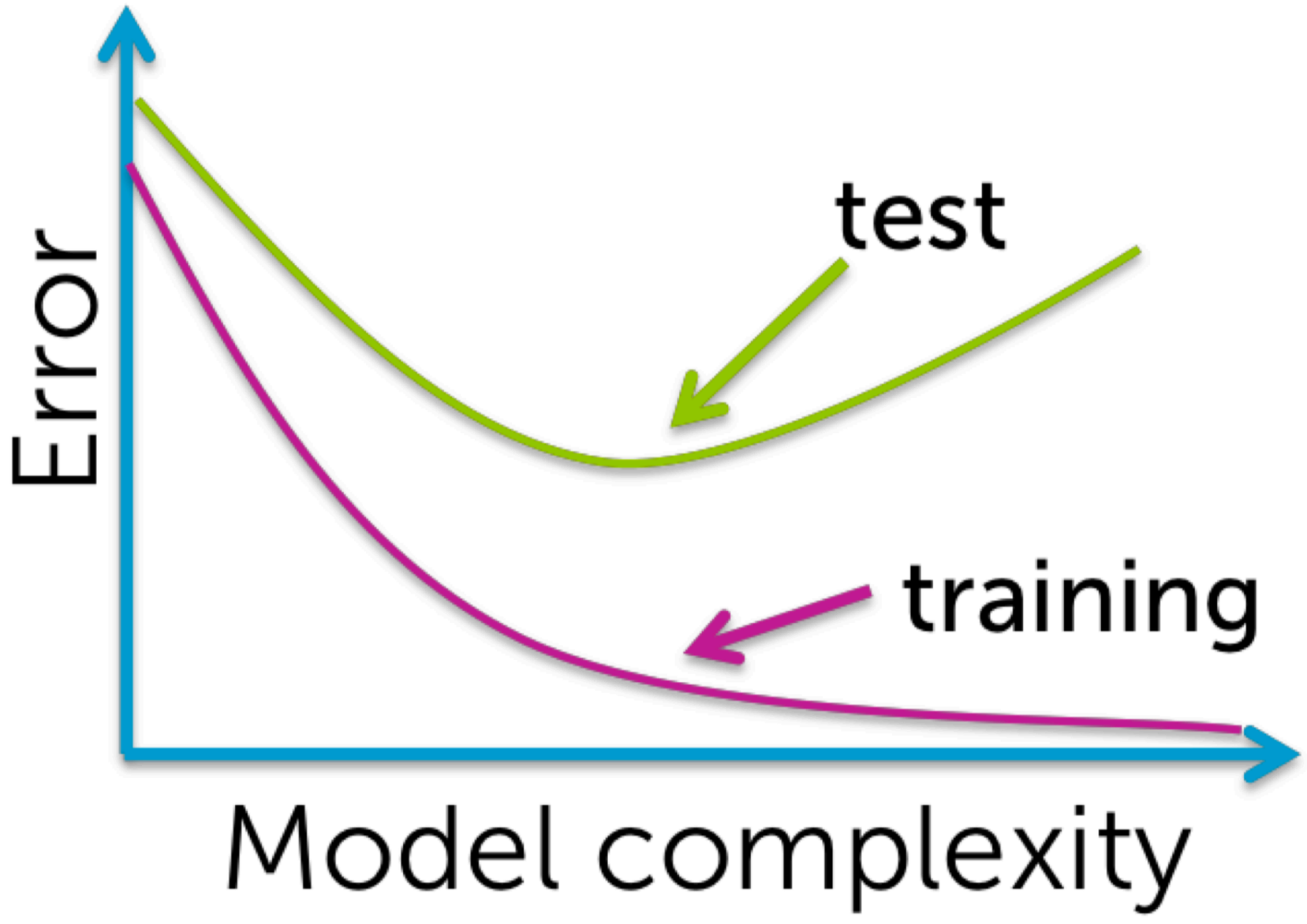


https://scikit-learn.org/stable/_images/sphx_glr_plot_tree_regression_0011.png

Overfitting en Arboles

¿Podemos prevenir overfitting?

¿Podemos prevenir overfitting?



¿Podemos prevenir overfitting?

(ahora después de entrenar)

¿Podemos prevenir overfitting?

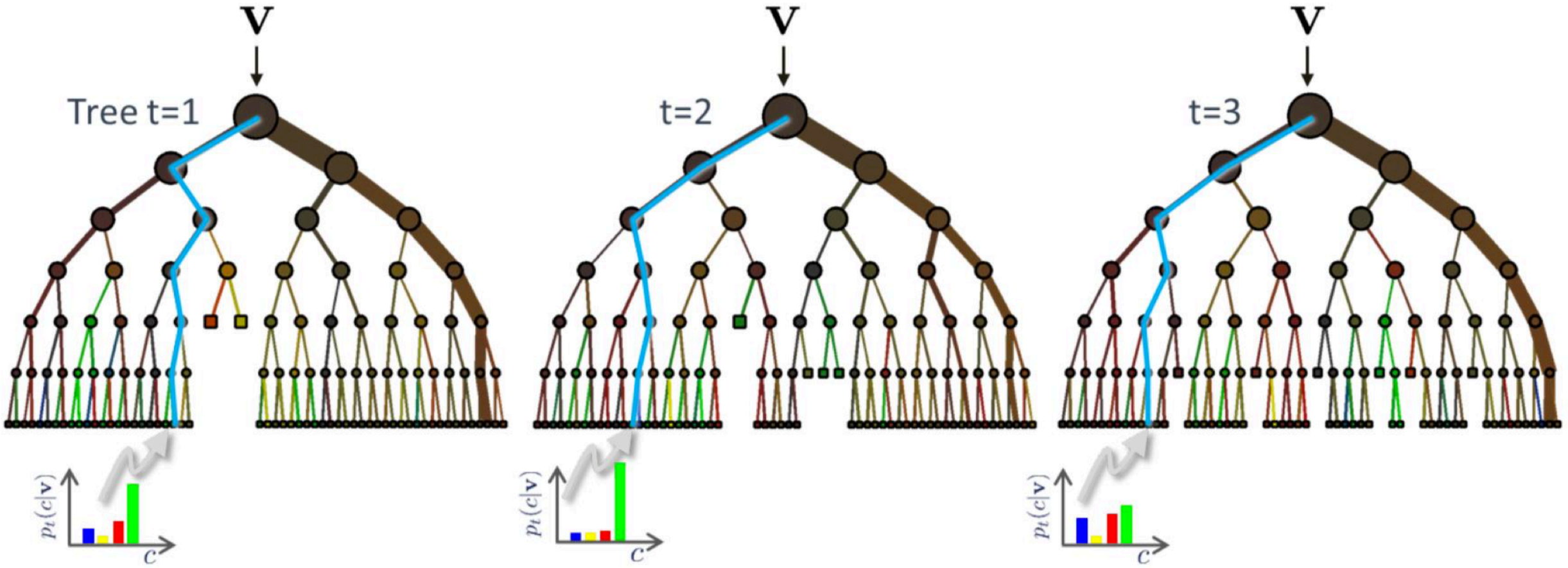
$$\frac{\text{err}(\text{prune}(T, t), S) - \text{err}(T, S)}{|\text{leaves}(T)| - |\text{leaves}(\text{prune}(T, t))|}$$

Ensambles

“The reason that model averaging works is that different models will usually not make all the same errors on the test set.”

— Page 256, Deep Learning, 2016.

Modelos no correlacionados



Variar datos “train”

“… a natural way to reduce the variance and hence increase the prediction accuracy of a statistical learning method is to take many training sets from the population, build a separate prediction model using each training set, and average the resulting predictions. […] Of course, this is not practical because we generally do not have access to multiple training sets. Instead, we can bootstrap, by taking repeated samples from the (single) training data set.”

— *Pages 216-317, An Introduction to Statistical Learning with Applications in R, 2013.*

Pregunta:

Estime error que tendrá el modelo en el set de test usando datos de entrenamiento.

Variar modelos

“Training the same under-constrained model on the same data with different initial conditions will result in different models given the difficulty of the problem, and the stochastic nature of the learning algorithm.”

https://machinelearningmastery.com/ensemble-methods-for-deep-learning-neural-networks/

Pregunta:

Variar forma de combinar

“average these points in weight space, and use a network with these averaged weights, instead of forming an ensemble by averaging the outputs of networks in model space”

— *Averaging Weights Leads to Wider Optima and Better Generalization, 2018.*

promedio,
suma ponderada,
modelo elige cual es el mejor para cada caso, (ej. datos lineales?)
boosting,
etc.

Motivación ensambles de árboles

Gradient boosted machines and deep neural nets have dominated recent Kaggle competitions

Competition	Type	Winning ML Library/Algorithm
Liberty Mutual	Regression	**XGBoost**
Caterpillar Tubes	Regression	**Keras** + **XGBoost** + Reg. Forest
Diabetic Retinopathy	Image	SparseConvNet + RF
Avito	CTR	**XGBoost**
Taxi Trajectory 2	Geostats	Classic neural net
Grasp and Lift	EEG	**Keras** + **XGBoost** + other CNN
Otto Group	Classification	Stacked ensemble of 35 models
Facebook IV	Classification	sklearn GBM

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.

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

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.)**

a) Considere un árbol de decisión para resolver un problema de clasificación con 2 categorías. En un nodo C del árbol hay 100 ejemplos; 30 positivos y 70 negativos. Si el atributo binario A se selecciona en C, su primer hijo obtendrá 18 ejemplos positivos y 22 negativos, y su segundo hijo obtendrá 12 ejemplos positivos y 48 negativos. ¿Cuál es la entropía del nodo C? **(3 ptos.)**

j. Explique por qué se considera a un árbol de decisión como una técnica de subespacios de clasificación.

l. Explique por qué es importante usar un set de validación al podar un árbol de decisión.

2. (*16 puntos*) **Árboles de decisión**

Considere el siguiente set de entrenamiento:

A	B	C	Class
1	1	0	0
1	0	1	1
0	1	1	1
0	0	1	0

a. (*4 pts*) Encuentre un árbol de decisión de profundidad mínima que permita clasificar estos registros sin error.