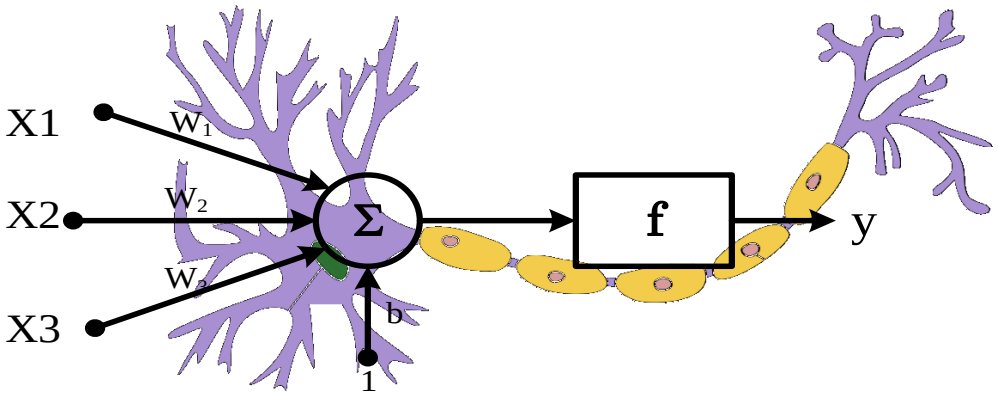
Redes Neuronales

Hoja de Ayuda

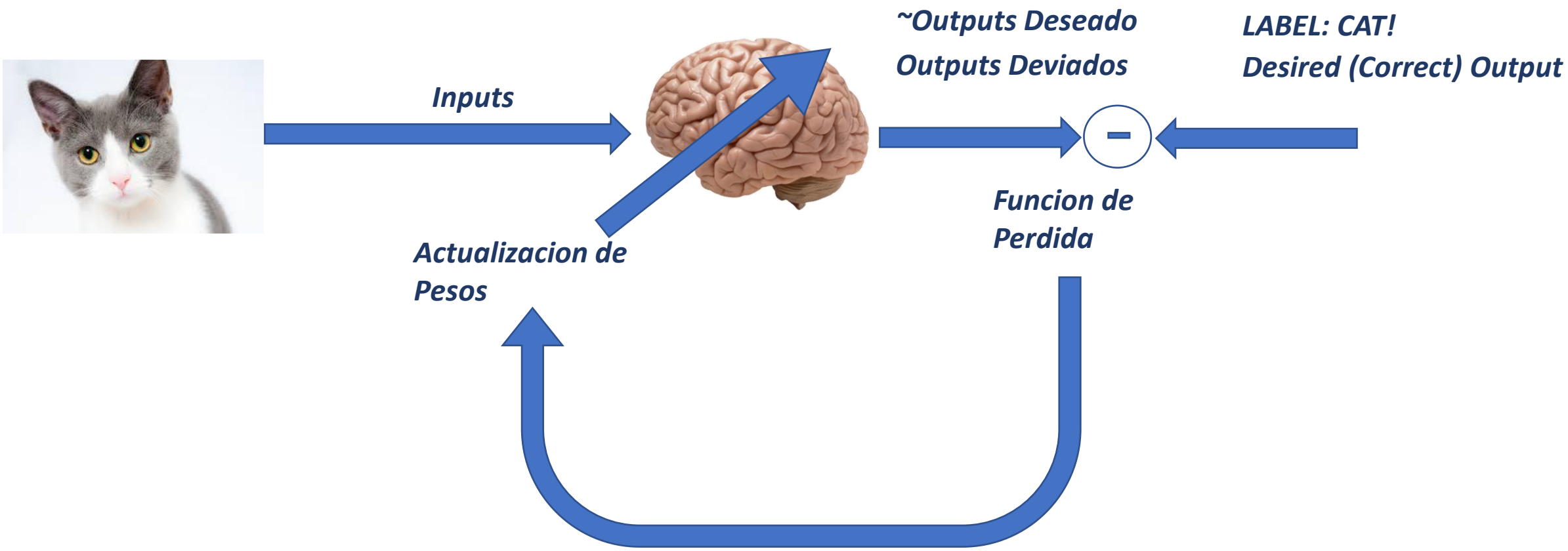
https://datadosis.com

Que son las Redes Neuronales y Como Aprenden?

- El cerebro tiene más de 100 mil millones de neuronas que se comunican a través de señales eléctricas y químicas.
- Las neuronas se comunican entre sí y nos ayudan a ver, pensar y generar ideas. El cerebro humano aprende creando conexiones entre estas neuronas.
- Las RNA son modelos de procesamiento de información inspirados en el cerebro humano.



Que son las RNA y Como Aprenden

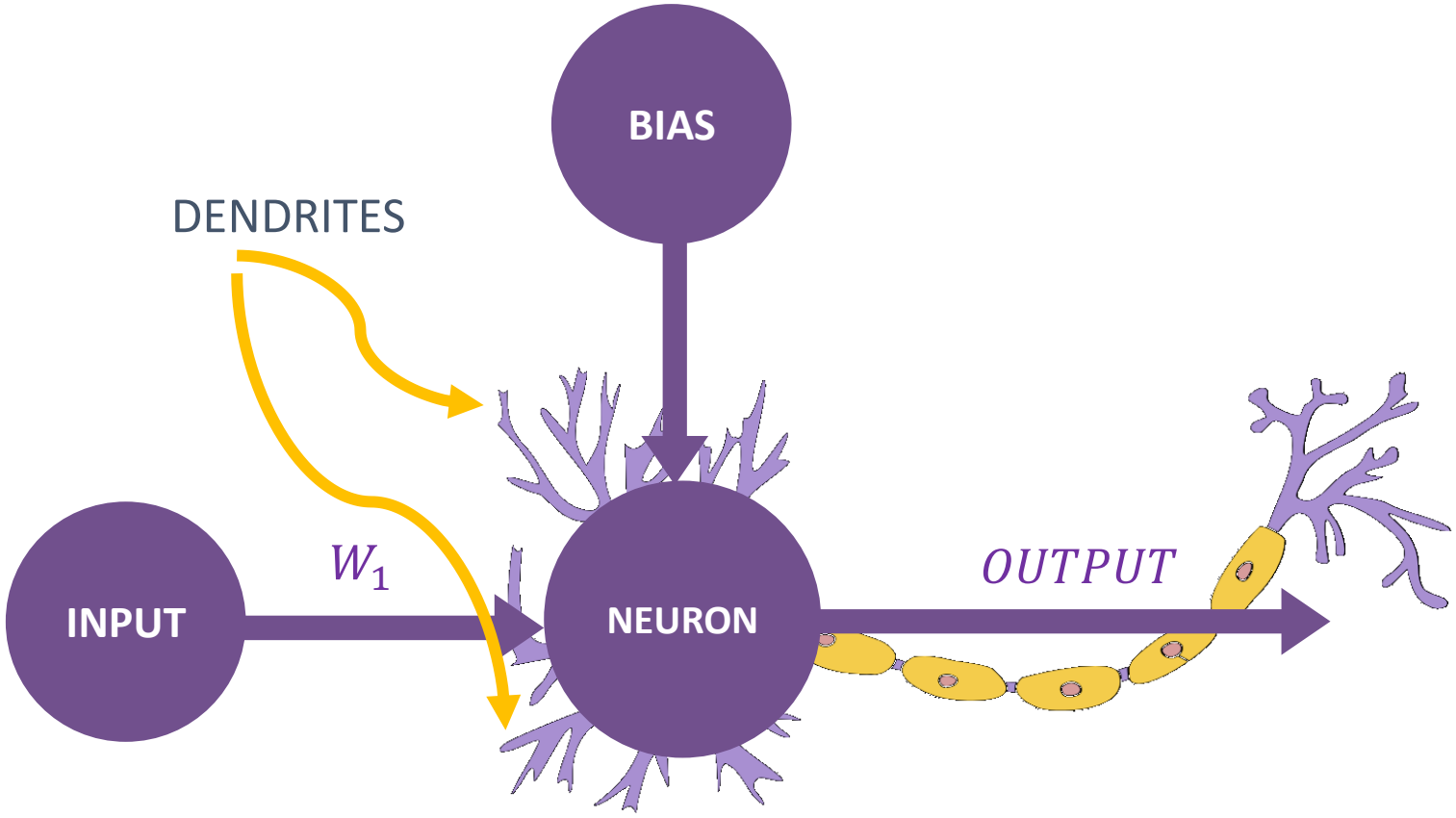


Redes Neuronales Artificiales en Accion!

Recuerdas la Primera Red?

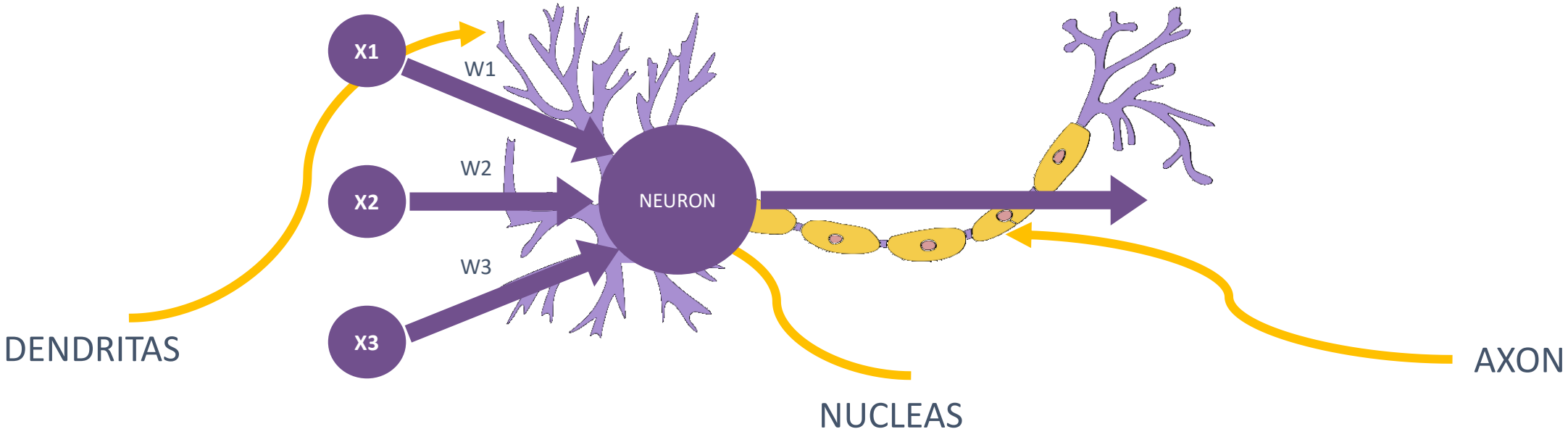
- El modelo era Muy Simple!
  - Sin function de activacion
  - Un Solo Input

$$\boldsymbol{Output = Input * W_1 + Bias}$$



Modelo Matematico de Neurona

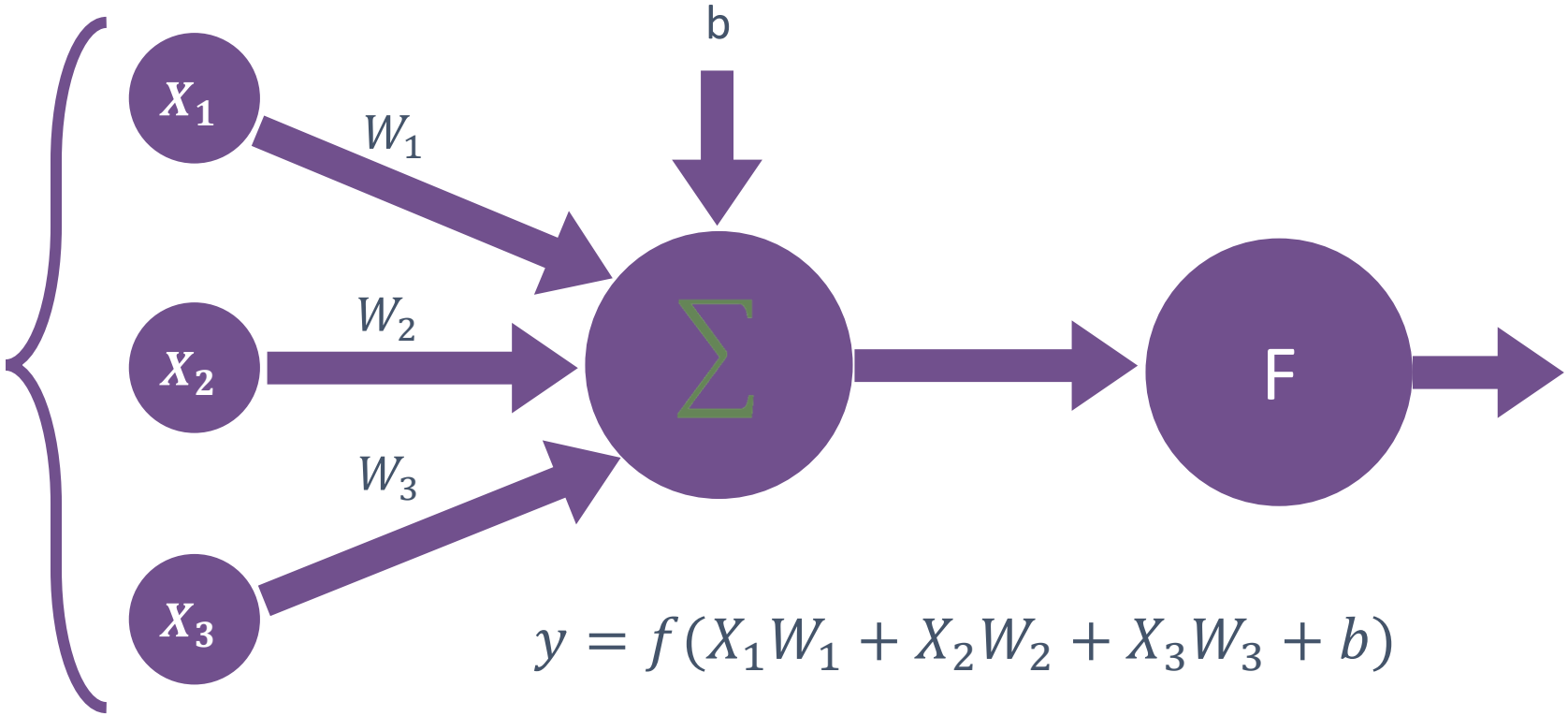
- La neurona recoge señales de los canales de entrada llamados dendritas, procesa la información en su núcleo y luego genera una salida en una larga y delgada rama llamada axón.



Recuerdas la Primera Red?

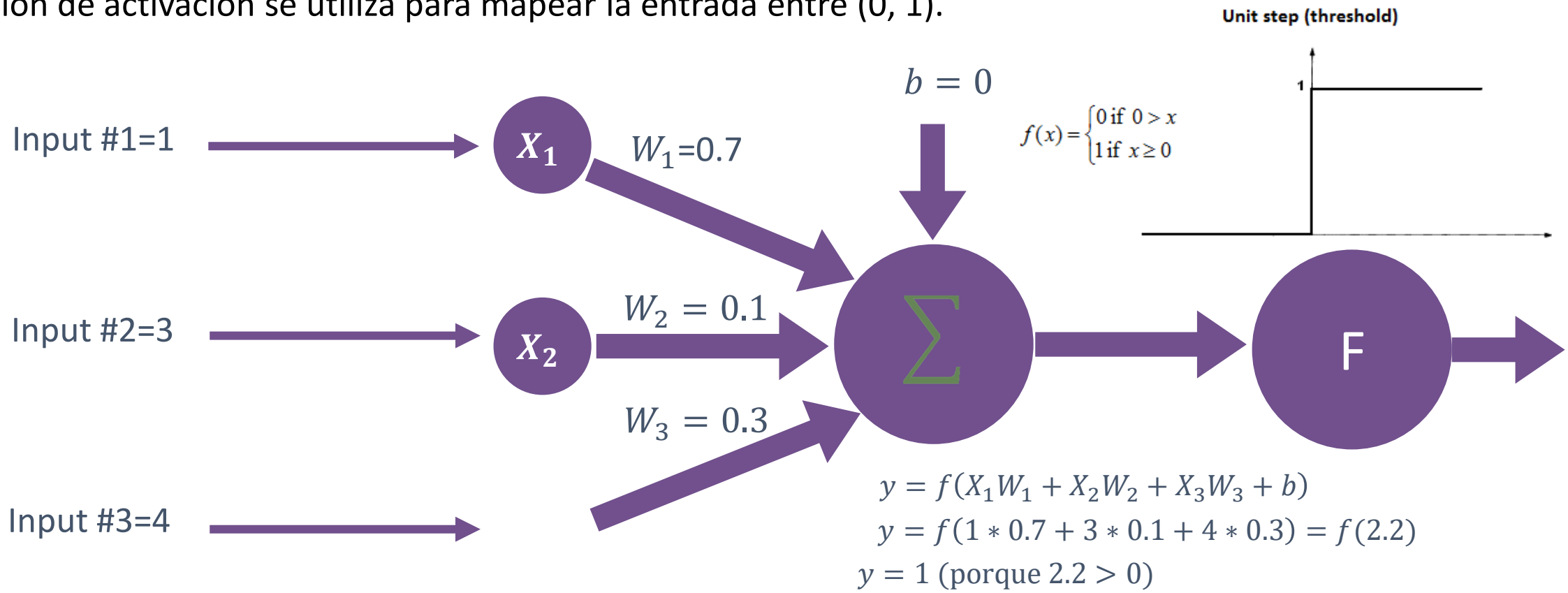
- El sesgo permite desplazar la curva de la función de activación hacia arriba o hacia abajo.
- Número de parámetros ajustables = 4 (3 pesos y 1 sesgo).
- Función de activación "F".

INPUTS/INDEPENDENT VARIABLES



Una Neurona en Accion

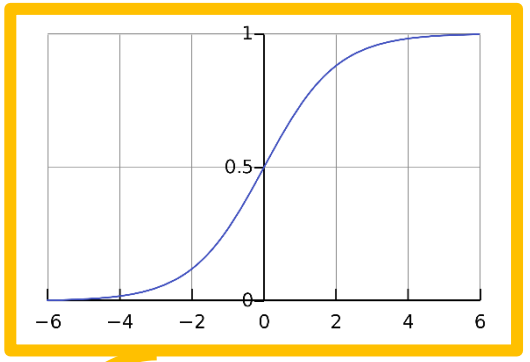
- Asumamos una función de activación de la Unidad de Paso.
- La función de activación se utiliza para mapear la entrada entre (0, 1).

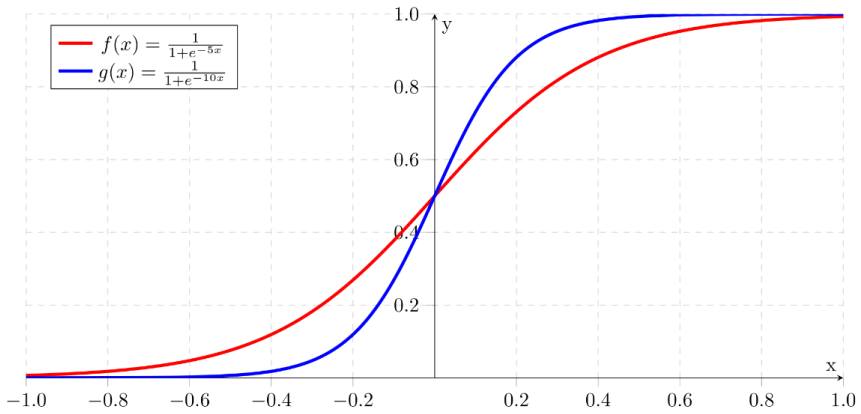


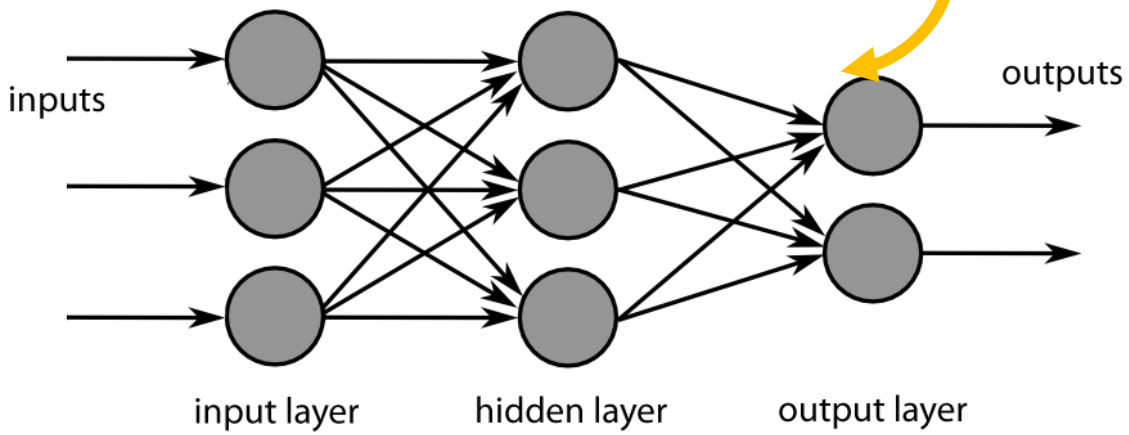
Funcion de Activacion

Funciones de Activacion

- SIGMOID:
  - Toma un número y lo pone entre 0 y 1
  - Convierte los grandes números negativos en 0 y los grandes números positivos en 1.
  - Generalmente se usa en la capa de salida.

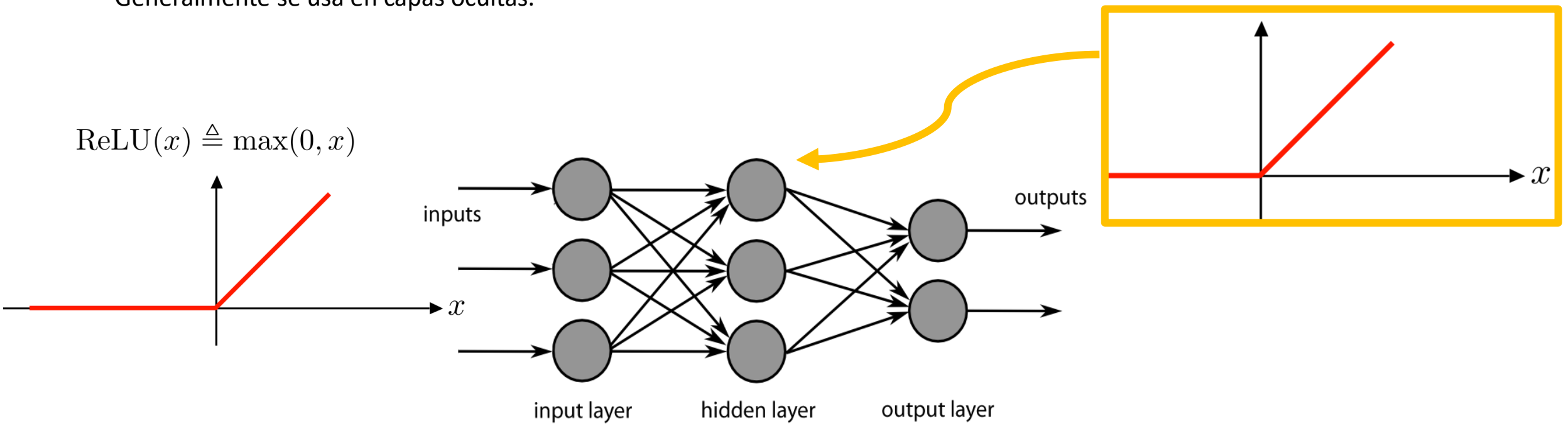






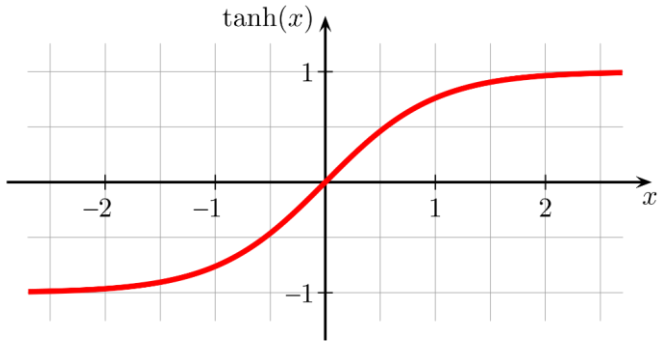
Funciones de Activacion

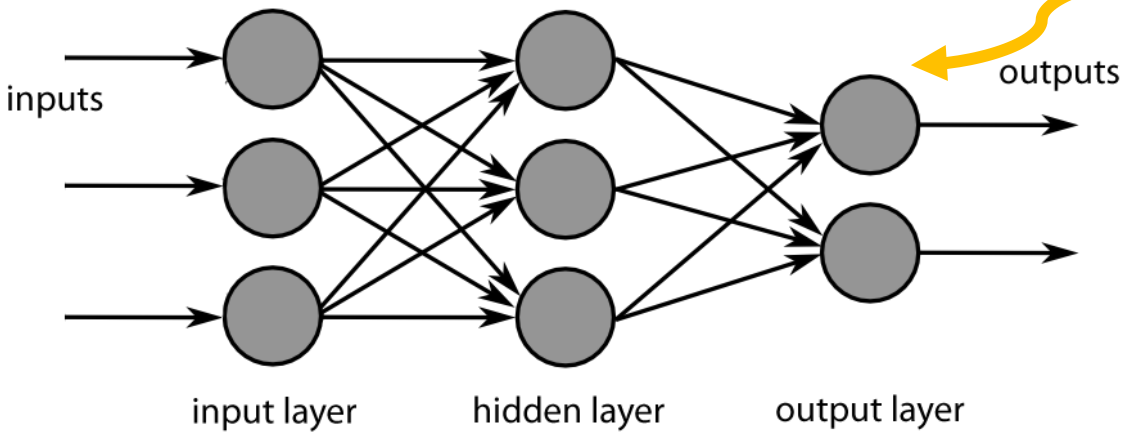
- RELU (UNIDADES LINEALES RECTIFICADAS):
  - si la entrada x < 0, la salida es 0 y si x > 0 la salida es x.
  - El relu no se satura, por lo que evita el problema del gradiente de desaparición.
  - Utiliza un umbral simple, por lo que es eficiente desde el punto de vista computacional.
  - Generalmente se usa en capas ocultas.

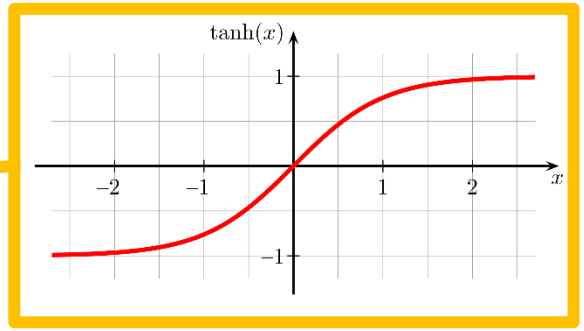


Funciones de Activacion

- FUNCIÓN DE ACTIVACIÓN DE LA TANGENTE HIPERBÓLICA:
  - "Tanh" es similar al sigmoide, convierte el número entre -1 y 1.
  - A diferencia de la sigmoide, las salidas de tanh están centradas en cero (rango: -1 y 1).
  - Tanh sufre el problema del gradiente de desaparición, por lo que mata los gradientes cuando está saturado.
  - En la práctica, el tanh es preferible al sigmoide.







MODELO MULTI-NEURÓNICO (MODELO PERCEPTRÓNICO MULTICAPA)

- La red está representada por una matriz de pesos, entradas y salidas.
- Número total de parámetros ajustables = 8:
- Pesos = 6
- Sesgos = 2

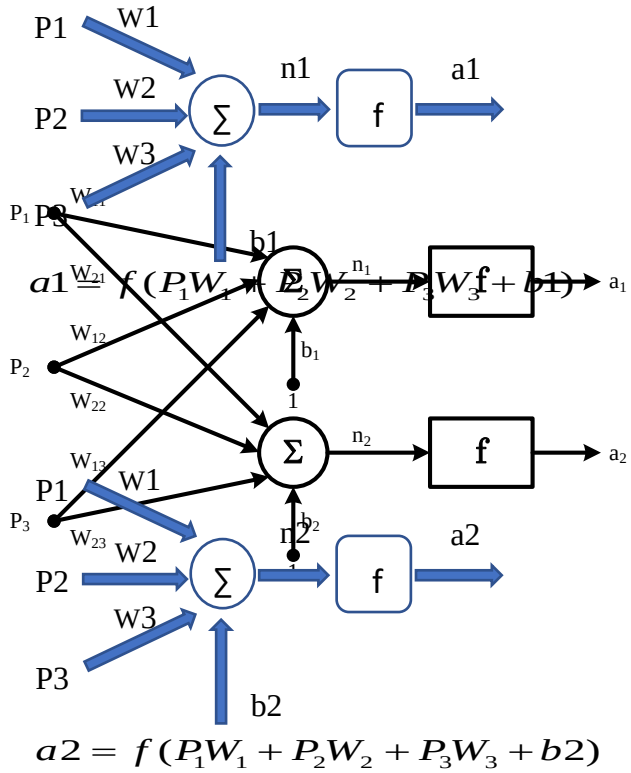
Matrix Representation

$$P = \begin{bmatrix} P_1 \\ P_2 \\ P_3 \end{bmatrix}$$

$$W = \begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \end{bmatrix}$$

$$b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

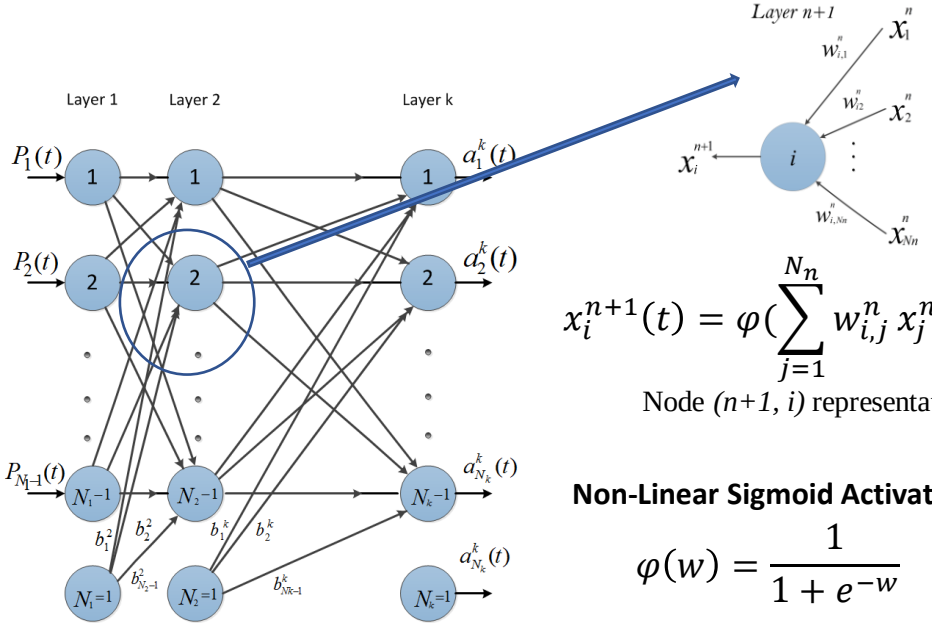
$$a = f(W \times P + b)$$



MODELO MULTI-NEURÓNICO (MODELO PERCEPTRÓNICO MULTICAPA)

- Conectemos varias de estas neuronas de forma multicapa.
- Cuantas más capas ocultas, más "profunda" será la red.

$$P = \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_{N_1} \end{bmatrix}$$



$$x_i^{n+1}(t) = \varphi\left(\sum_{j=1}^{N_n} w_{i,j}^n \, x_j^n(t)\right)$$

Node *(n+1, i)* representation

$$\begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1,N_1} \\ W_{21} & W_{22} & & W_{2,N_1} \\ \vdots & & \ddots & \vdots \\ W_{m-1,1} & W_{m-1,2} & \cdots & W_{m-1,N_1} \\ W_{m,1} & W_{m,2} & & W_{m,N_1} \end{bmatrix}$$

Non-Linear Sigmoid Activation function

$$\varphi(w) = \frac{1}{1 + e^{-w}}$$

m: numero neuronas en capa oculta

N_1*: numero de inputs*

Como Entrenan las Redes Neuronales

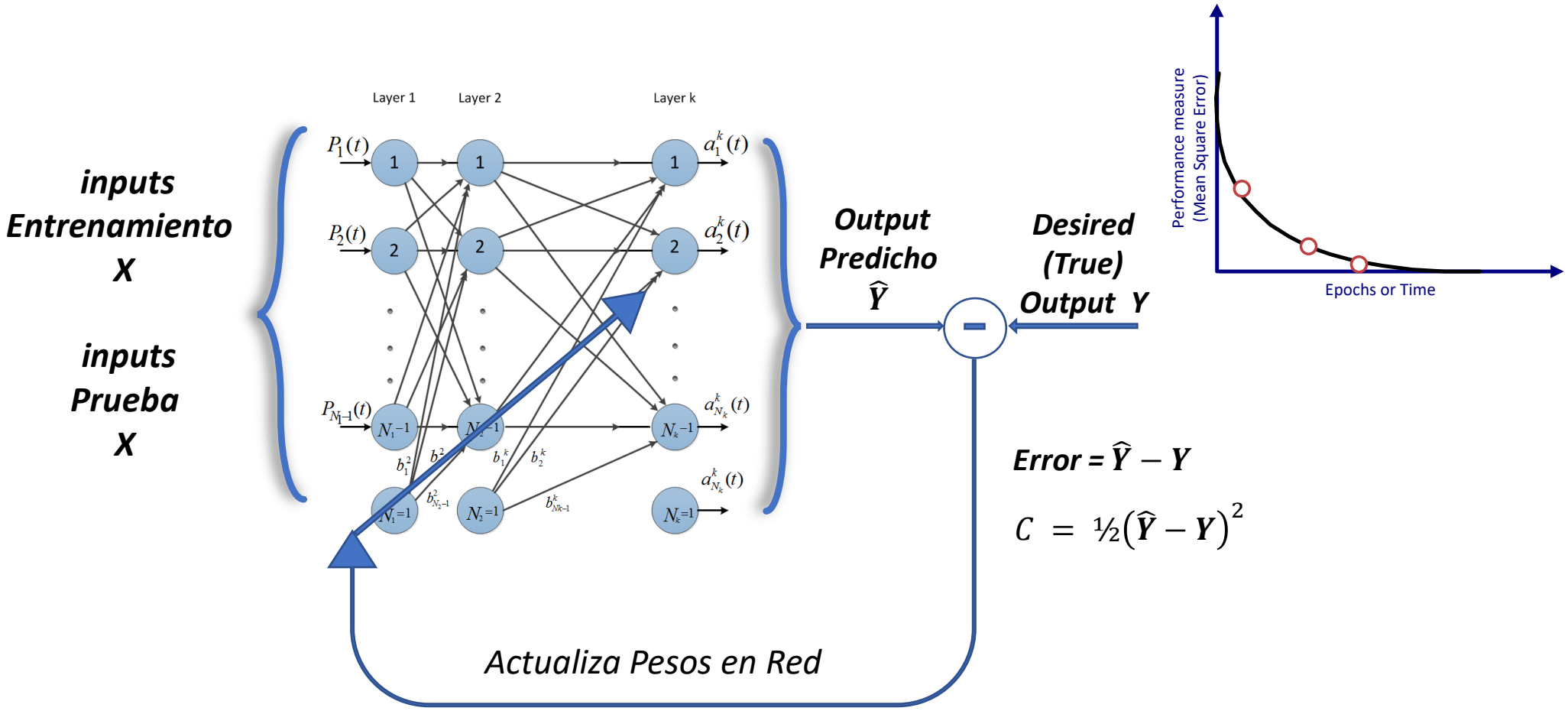
Estrategias de Entrenamiento

- Aprendizaje supervisado
  - Se utiliza si hay un gran conjunto de datos de prueba con etiquetas conocidas (salidas).
  - El algoritmo de aprendizaje evalúa la salida (es decir, hace predicciones), compara la salida con la etiqueta y ajusta la red y la repetición.

- Aprendizaje no supervisado
  - Se utiliza con datos "no etiquetados" (no categorizados) (Ej: agrupación de k-means).
  - Dado que el algoritmo de aprendizaje funciona con datos no etiquetados, no hay forma de evaluar la exactitud de la estructura sugerida por el algoritmo

- Aprendizaje reforzado
  - El algoritmo de aprendizaje toma acciones que maximizan alguna noción de recompensa acumulativa.
  - Con el tiempo, la red aprende a preferir el tipo de acción correcta y a evitar la incorrecta.

Epochs


inputs
Entrenamiento
X
inputs
Prueba
X
Layer 1
Layer 2
Layer k
Output
Predicho
$\widehat{Y}$
Desired
(True)
Output Y
Performance measure
(Mean Square Error)
Epochs or Time
Error = $\widehat{Y} - Y$
$C = \frac{1}{2}(\widehat{Y} - Y)^2$
Actualiza Pesos en Red

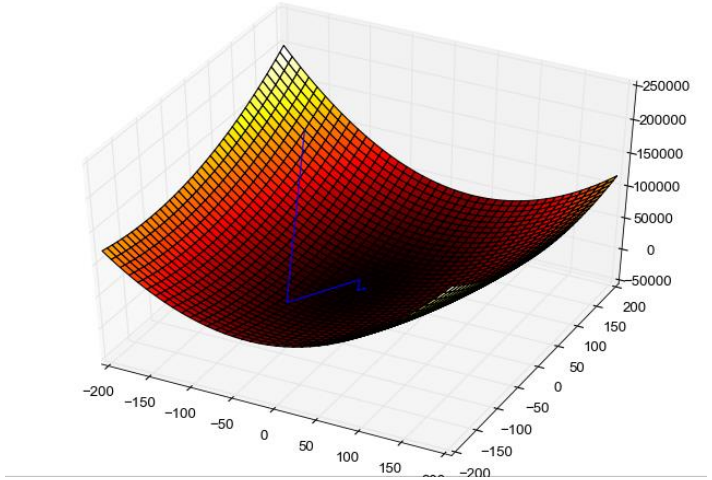
Divide Datos en Sets de Entrenamiento y Prueba

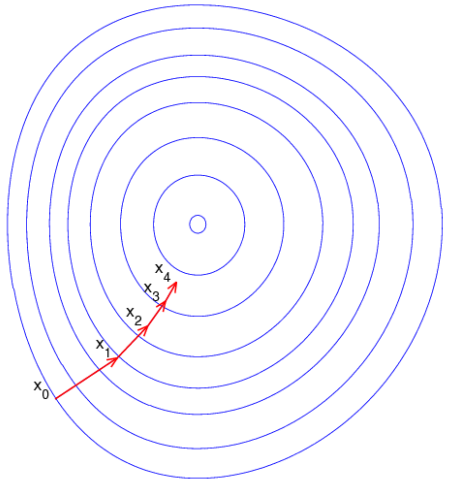
- El conjunto de datos se divide generalmente en segmentos de 50%, 25%, 25% para entrenamiento, validación y pruebas, respectivamente.
  - Conjunto de entrenamiento: se utiliza para el cálculo de gradientes y la actualización del peso.
  - Conjunto de validación:
    - utilizado para la validación cruzada que se realiza para evaluar la calidad del entrenamiento a medida que éste avanza.
    - La validación cruzada se implementa para superar el sobreajuste (sobreentrenamiento). El sobreajuste se produce cuando el algoritmo se centra en los detalles del conjunto de entrenamiento a costa de perder la capacidad de generalización.
    - La red entrenada MSE puede ser pequeña durante el entrenamiento, pero durante las pruebas, la red puede mostrar un rendimiento de generalización deficiente.
  - Conjunto de pruebas: se utiliza para probar la red entrenada.

Descenso de Gradiente

Descenso de Gradiente

- El descenso de gradiente es un algoritmo de optimización que se utiliza para obtener los valores optimizados de peso y sesgo de la red
- Funciona tratando iterativamente de minimizar la función de costo
- Funciona calculando el gradiente de la función de costo y moviéndose en dirección negativa hasta que se alcanza el mínimo local/global
- Si se toma el positivo del gradiente, se alcanza el máximo local/global





Rango de Aprendizaje

El descenso gradual funciona de la siguiente manera:

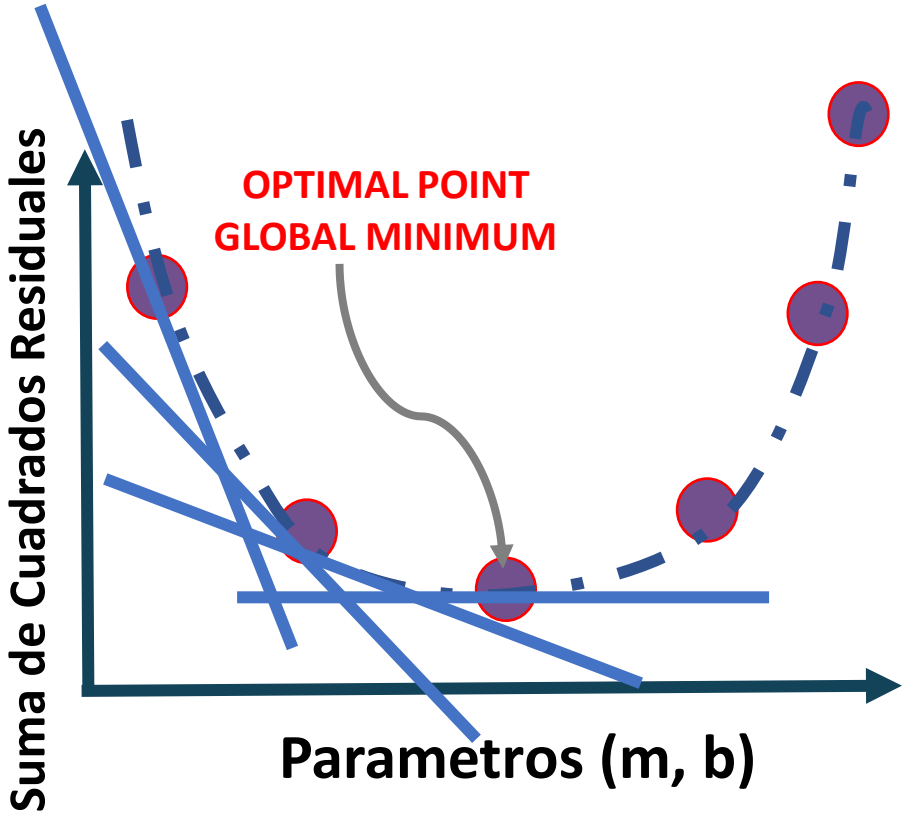
1. Calcular la derivada (gradiente) de la función de pérdida
2. Escoge valores aleatorios para los parámetros m, b y sustituto
3. Calcular el tamaño del paso (¿cuánto vamos a actualizar los parámetros?)

$$\boldsymbol{Step\ size} = \boldsymbol{Slope} * \boldsymbol{learning\ rate}$$

4. Actualice los parámetros y repita

$$y = b + m * x$$

GOAL IS TO FIND BEST PARAMETERS



*Nota: en realidad, este gráfico es 3D y tiene tres ejes, uno para m, b y la suma de los residuos cuadrados

Descenso de Gradiente con Matematicas

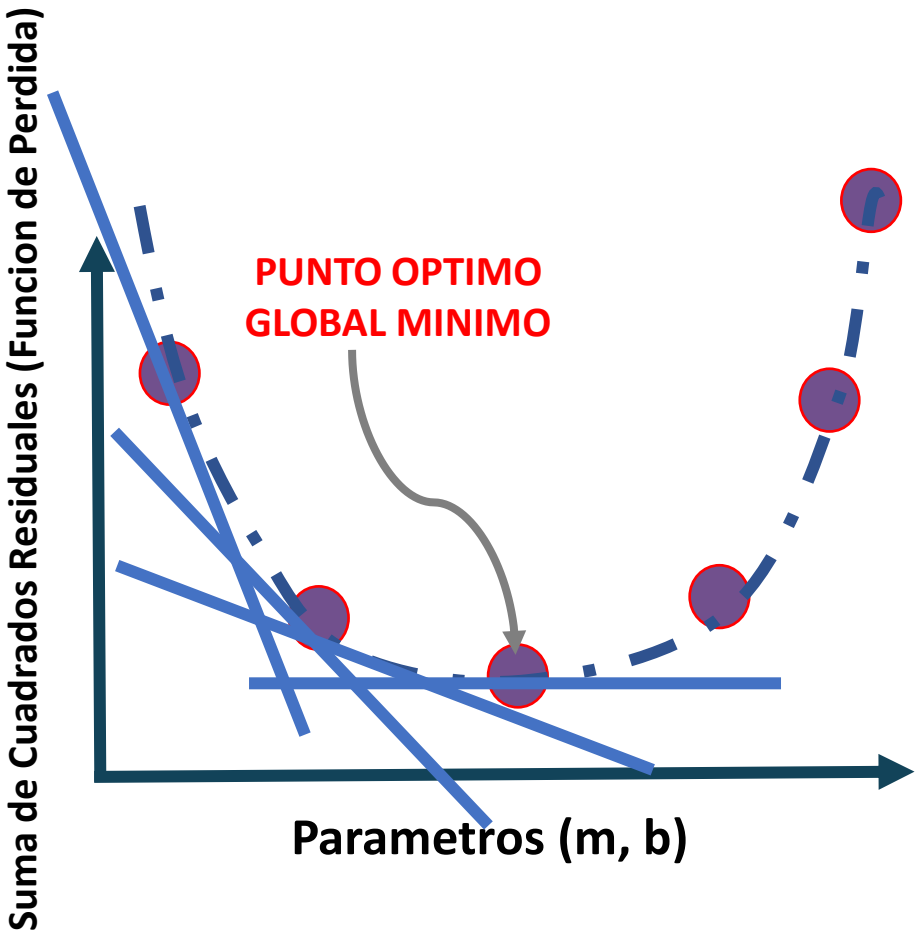
$$y = b + m * x$$

LA META ES ENCONTRAR LOS MEJORES PARAMETROS

$$Funcion\ de\ Perdida\ f(m,b) = \frac{1}{N}\sum_{i=1}^{n}\left(y_i - (b + m * x_i)\right)^2$$

$$gradiente\ f'(m,b) = \begin{bmatrix} \frac{df}{dm} \\ \frac{df}{db} \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{n} -2x_i\left(y_i - (b + m * x_i)\right)^2 \\ \frac{1}{N}\sum_{i=1}^{n} -2\left(y_i - (b + m * x_i)\right)^2 \end{bmatrix}$$

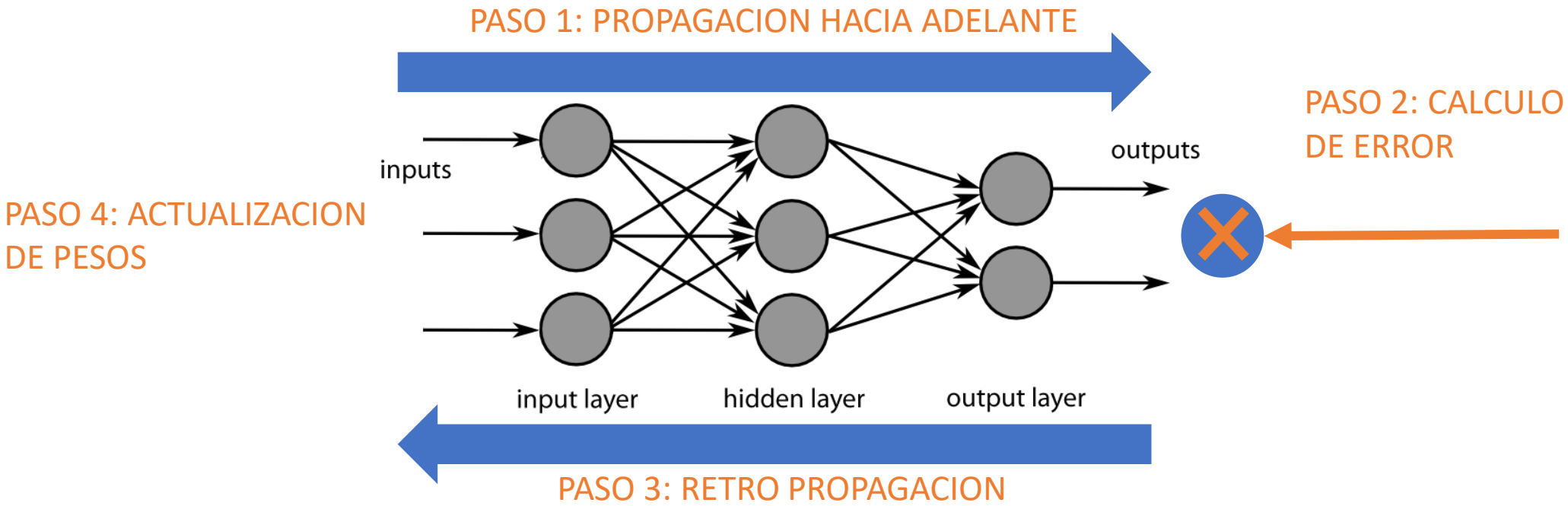
*Nota: en realidad, este gráfico es 3D y tiene tres ejes, uno para m, b y la suma de los residuos cuadrados



Retropropagacion

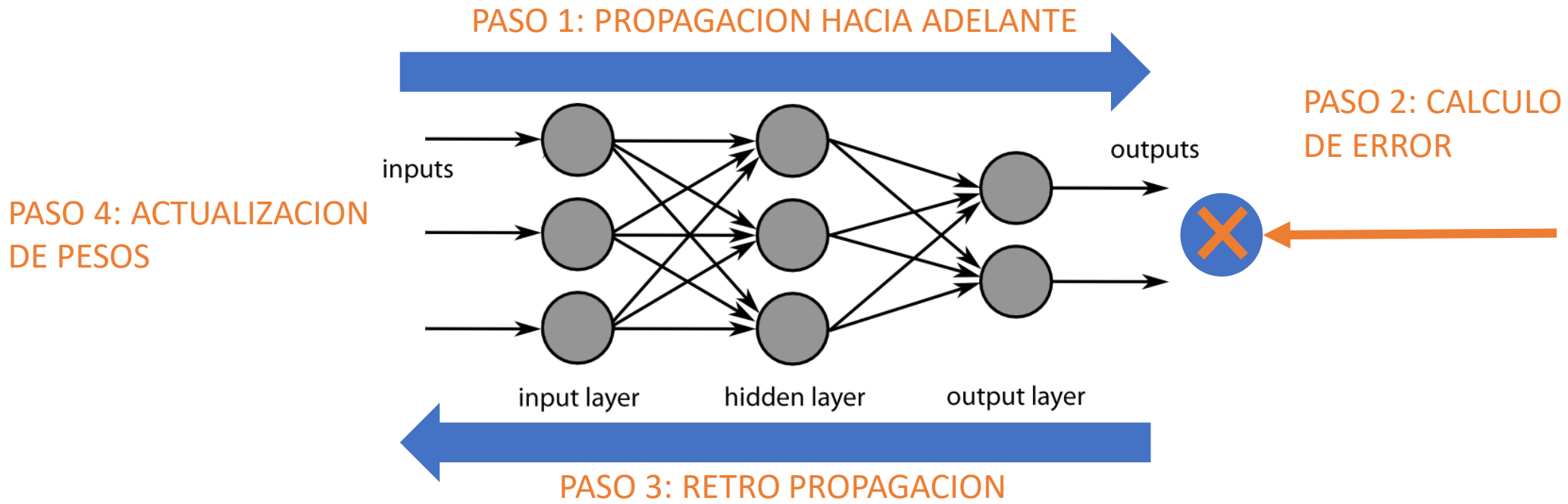
Retro-Propagacion

- La retro-propagación es un método utilizado para entrenar a las RNA mediante el cálculo del gradiente necesario para actualizar los pesos de la red.
- Se utiliza comúnmente por el algoritmo de optimización de descenso de gradiente para ajustar el peso de las neuronas mediante el cálculo del gradiente de la función de pérdida.



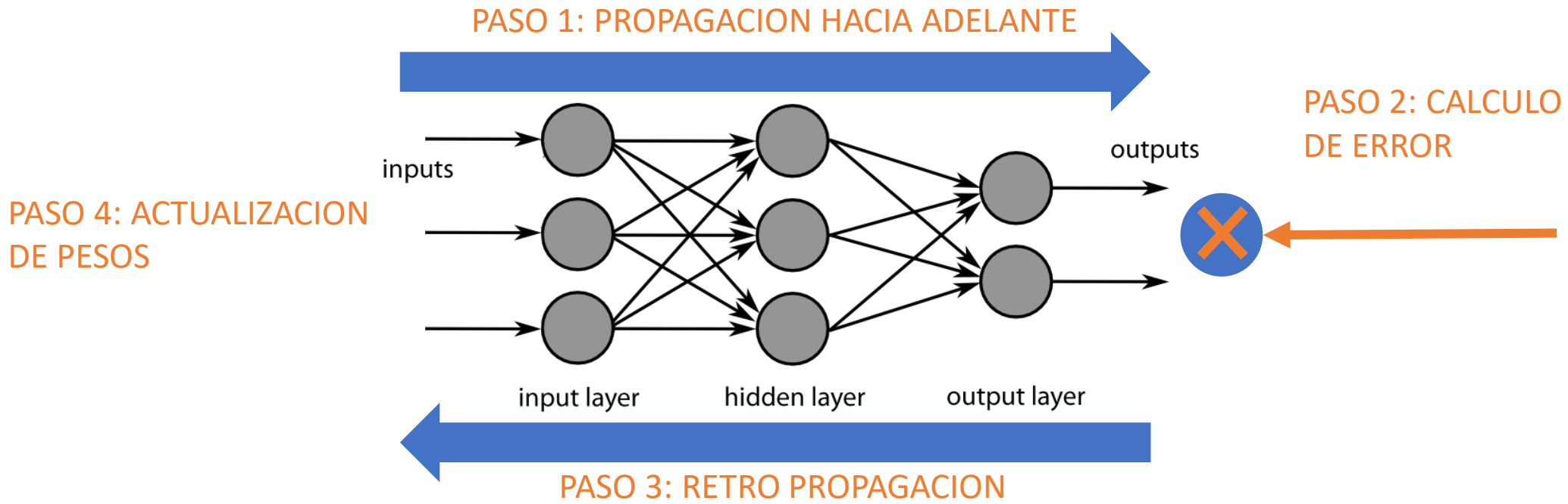
Retro-Propagacion

- Fase 1 de retropropagación: propagación
  - Propagación hacia adelante a través de la red para generar el valor o valores de salida
  - Cálculo del costo (término de error)
  - Propagación de las activaciones de salida de vuelta a través de la red utilizando el patrón de entrenamiento objetivo para generar los deltas (diferencia entre los valores de salida objetivo y los reales)



Retro-Propagacion

- Fase 2: actualización del peso
  - Calcula el gradiente de peso.
  - Se resta del peso una proporción (porcentaje) del gradiente del peso.
  - Esta proporción influye en la velocidad y la calidad del aprendizaje y se denomina tasa de aprendizaje. Cuanto mayor es la proporción, más rápido es el entrenamiento de las neuronas, pero menor es la proporción, más preciso es el entrenamiento.



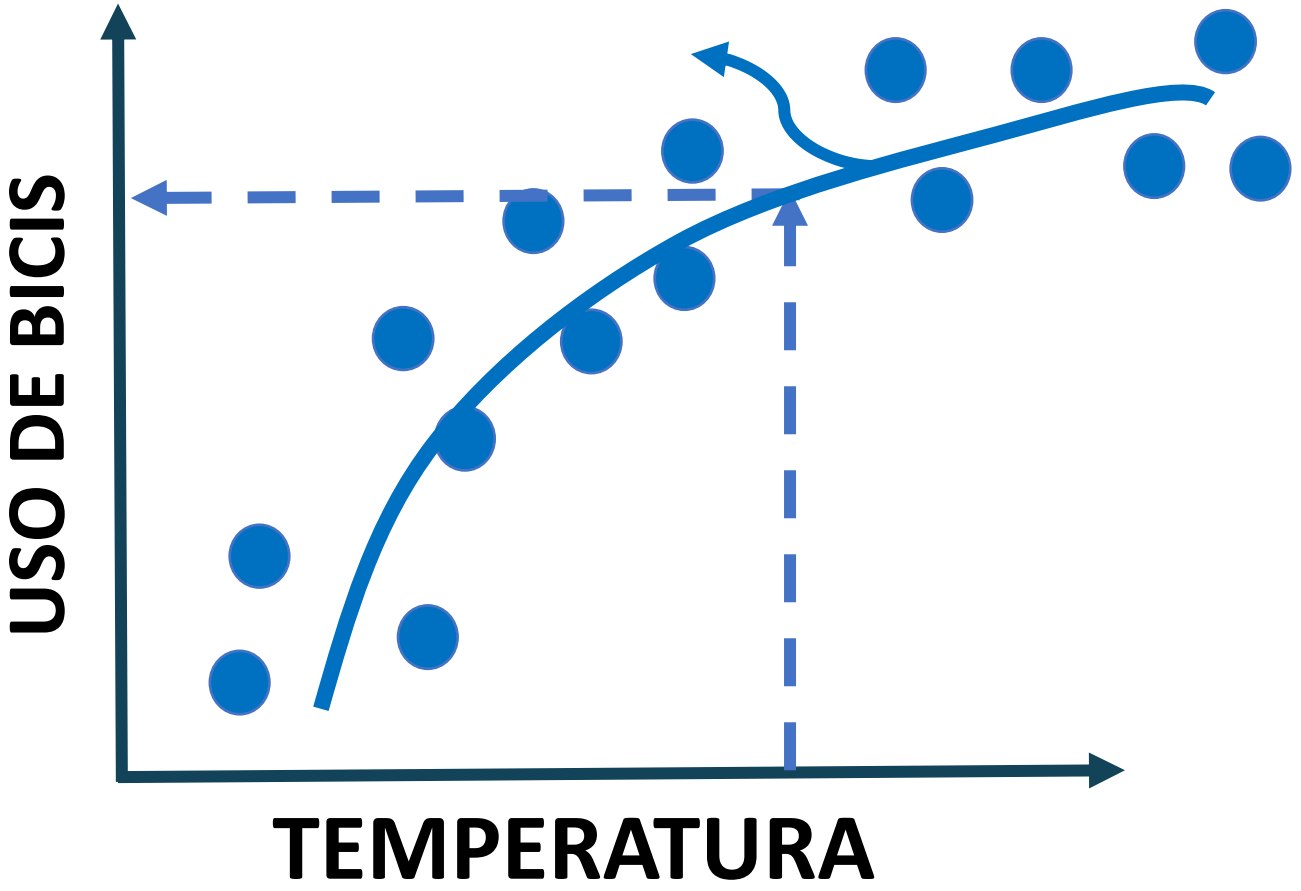
Lectura Extra sobre Retro-Propagacion

- “Backpropagation neural networks: A tutorial” by Barry J.Wythoff
- “Improved backpropagation learning in neural networks with windowed momentum”, International Journal of Neural Systems, vol. 12, no.3&4, pp. 303-318.

COMPENSACIÓN DE LA VARIACIÓN DEL SESGO

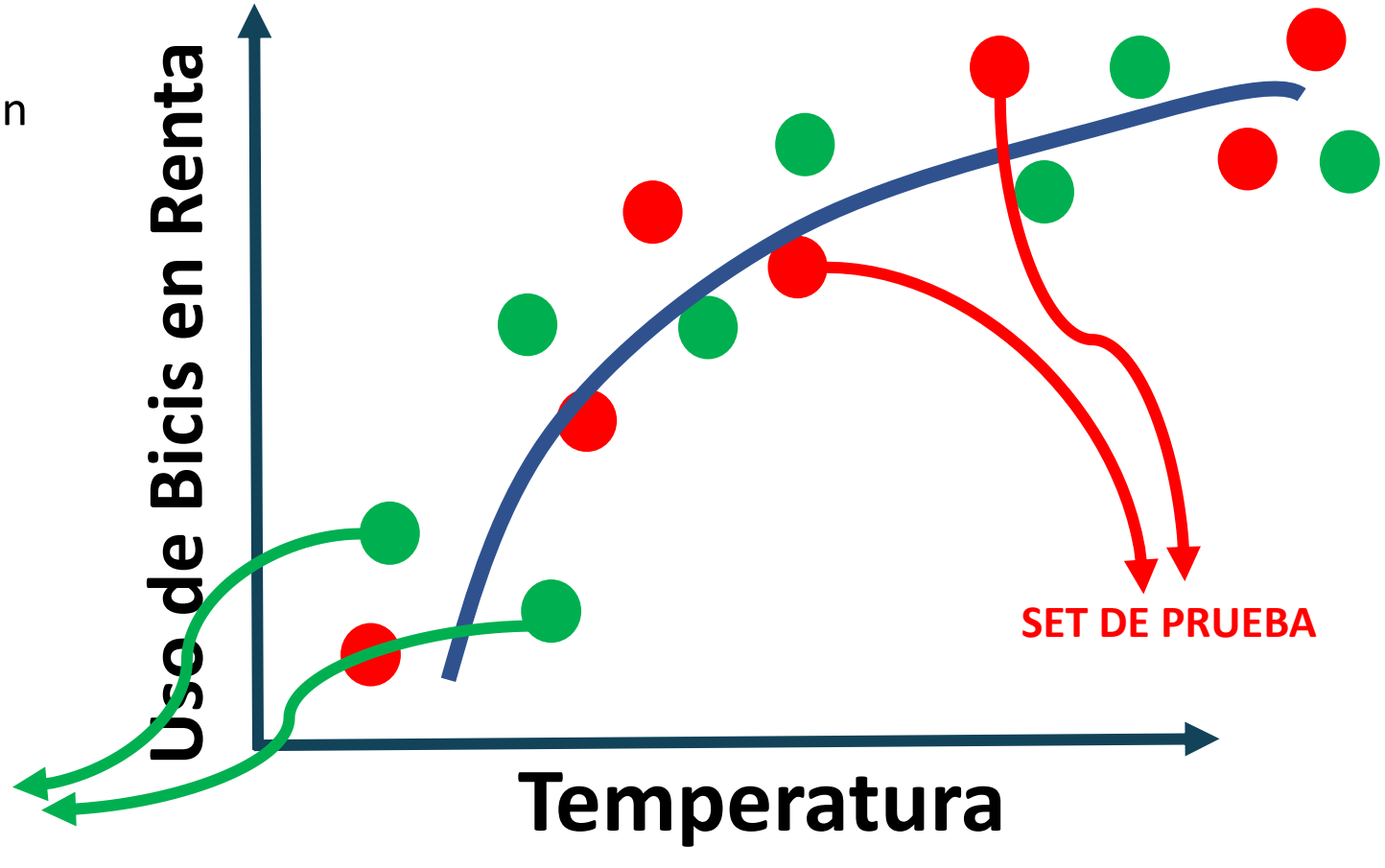
Intuición de la Variación del Sesgo

- Asumamos que queremos obtener la relación entre la temperatura y el uso de la bicicleta de alquiler.
- A medida que la experiencia de la temperatura aumenta, el uso del alquiler de bicicletas tiende a aumentar también.
- A medida que la temperatura supera un cierto límite, el uso tiende a estancarse y no aumenta más.



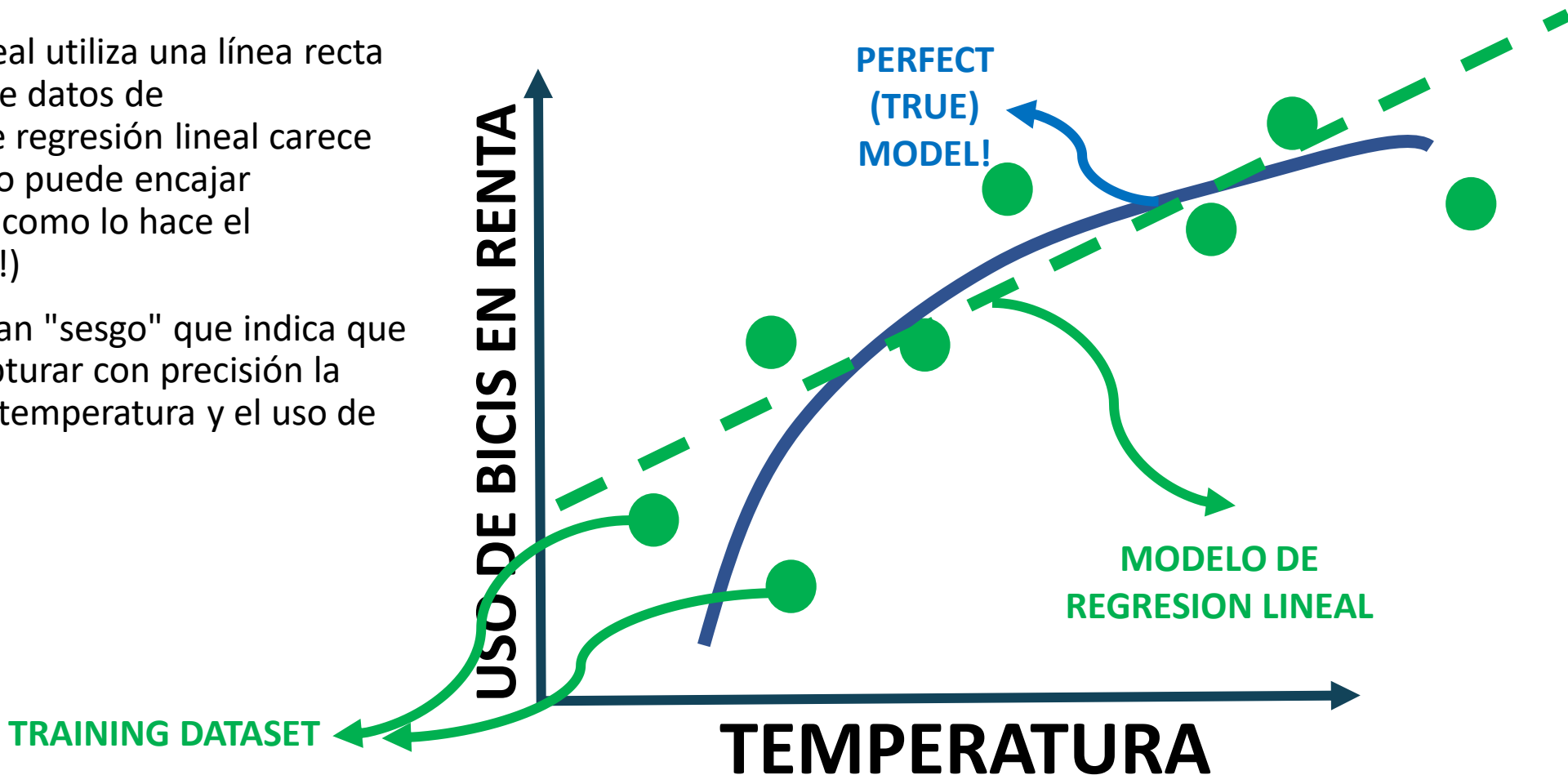
Varianza de Sesgo en Sets Entrenamiento y Prueba

- El conjunto de datos se divide en conjuntos de datos de entrenamiento y pruebas
- Los conjuntos de datos de prueba nunca han sido vistos por el modelo antes



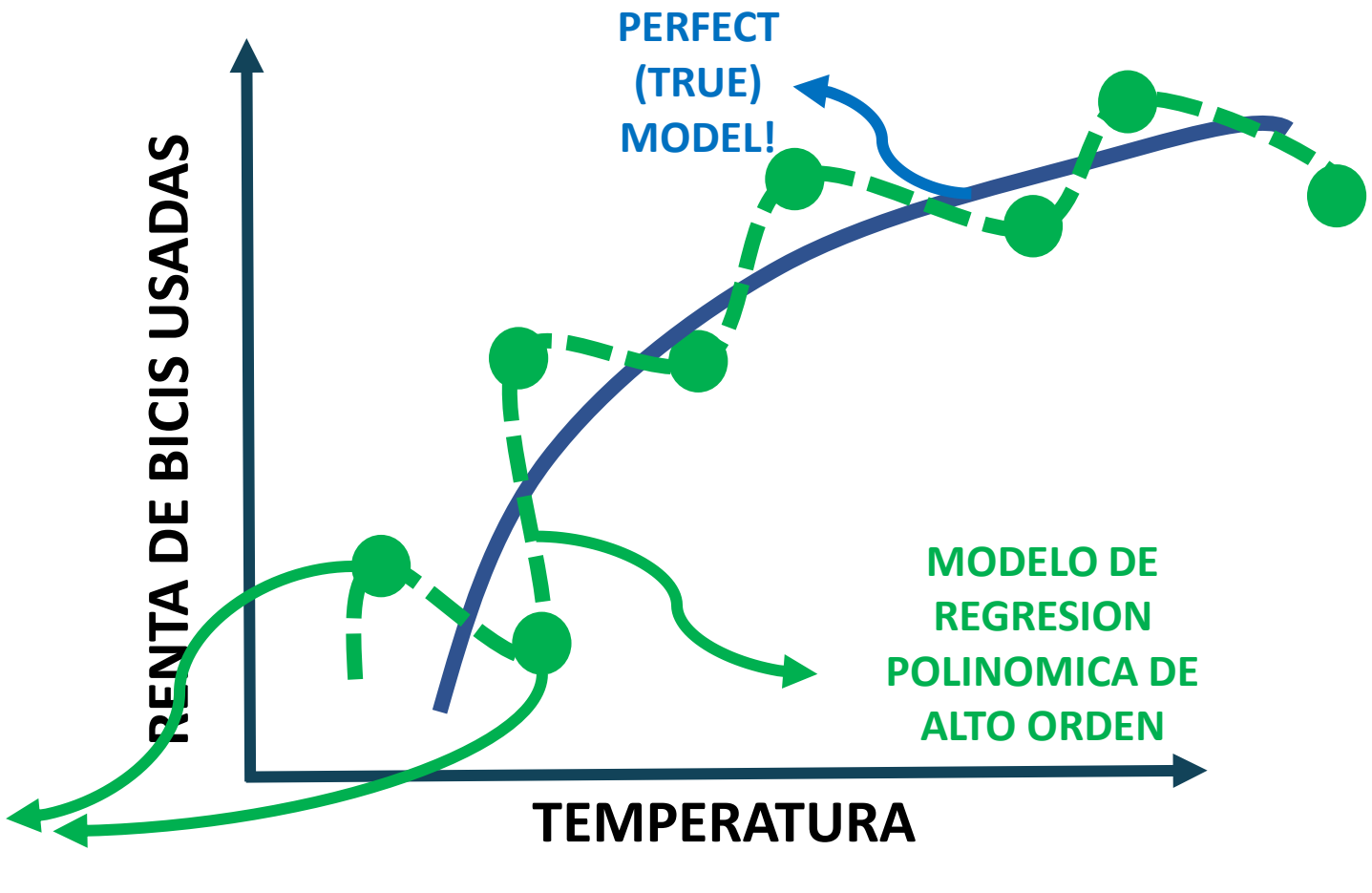
Varianza de Sesgos: Modelo #1 – Regresion Lineal Simple

- El modelo de Regresión Lineal utiliza una línea recta para ajustarse al conjunto de datos de entrenamientoEl modelo de regresión lineal carece de flexibilidad, por lo que no puede encajar adecuadamente los datos (¡como lo hace el verdadero modelo perfecto!)
- El modelo lineal tiene un gran "sesgo" que indica que el modelo es incapaz de capturar con precisión la verdadera relación entre la temperatura y el uso de la renta.

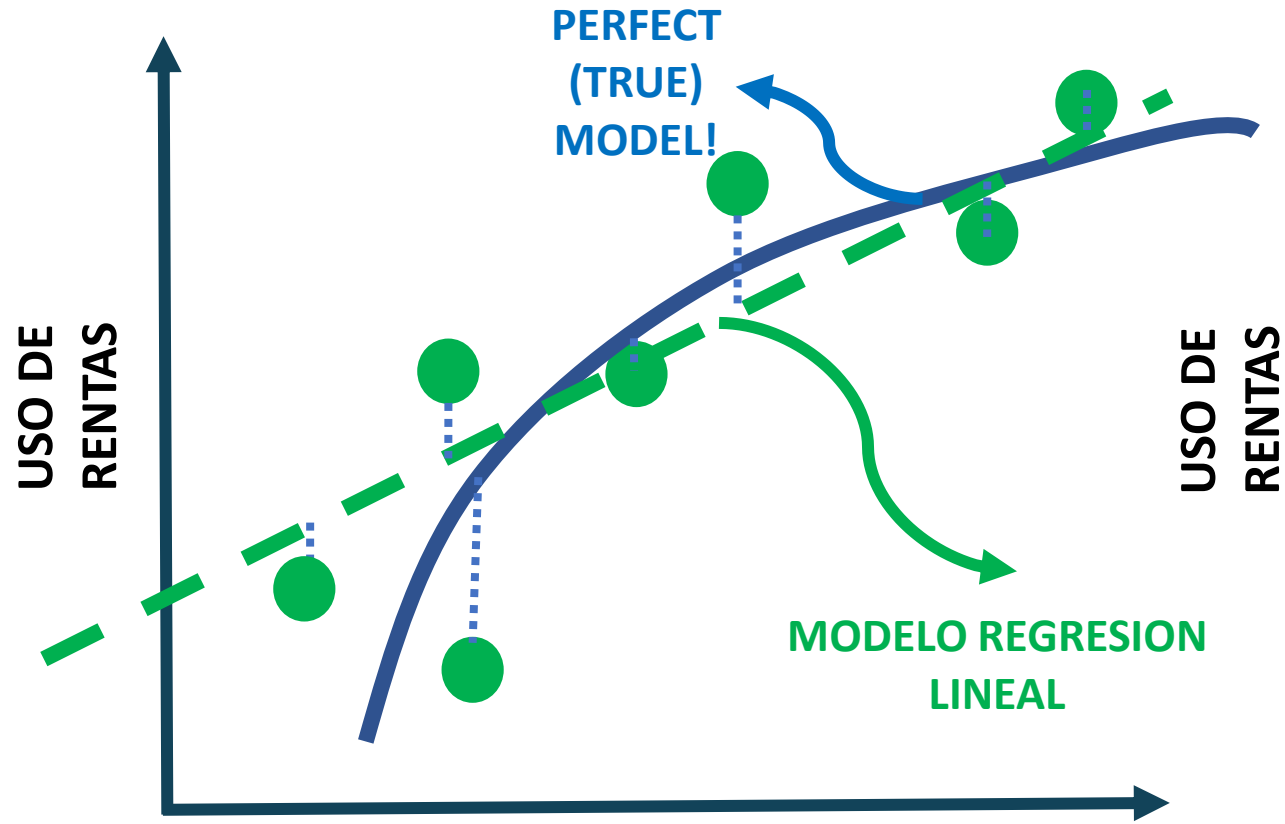


Sesgo y Varianza: Modelo #2 – Regresion Polinomica de Alto Orden

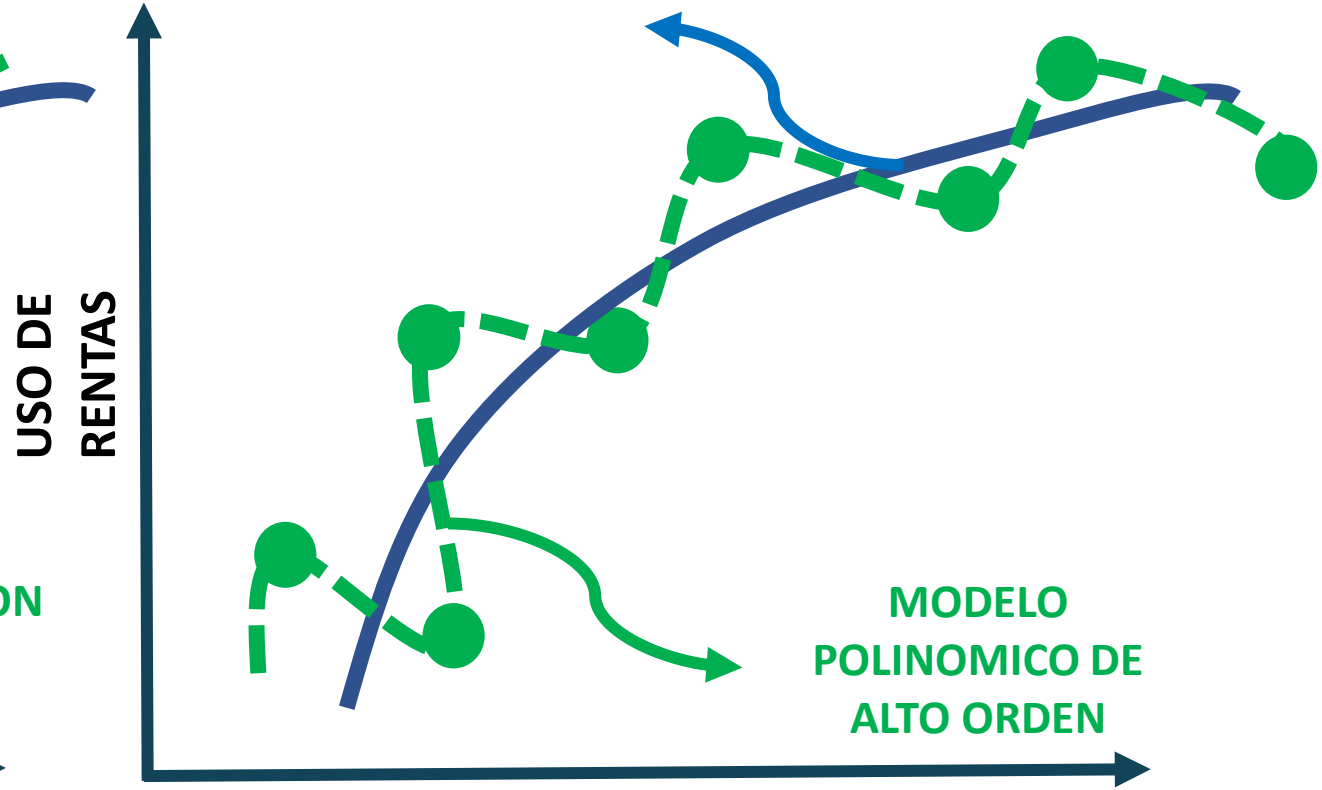
- El modelo de polinomios de alto orden es capaz de tener un sesgo muy pequeño y puede encajar perfectamente en el conjunto de datos de entrenamiento.
- El modelo de polinomio de alto orden es muy flexible



Sesgo y Varianza: Modelo #1 vs Modelo #2 Durante Entrenamiento

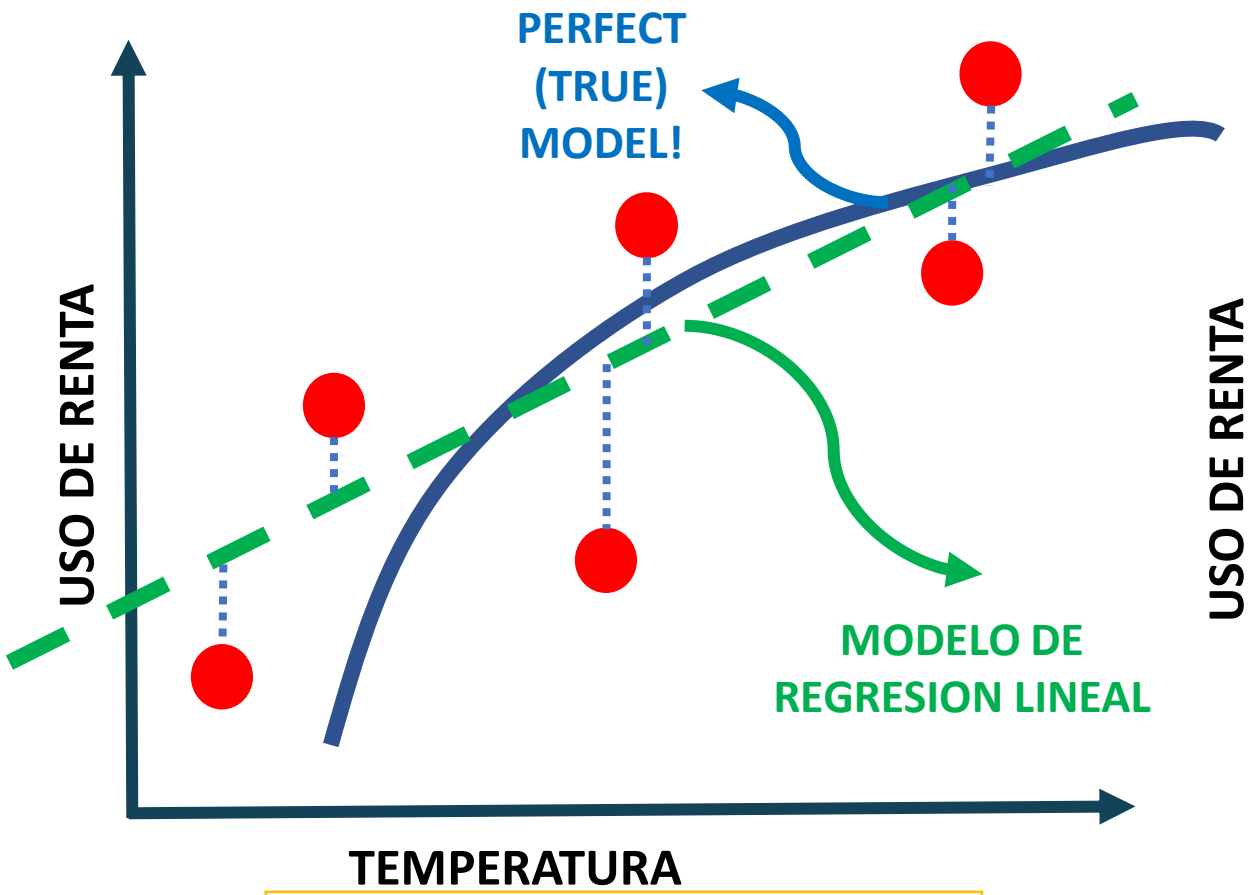


SUMA DE CUADRADOS (LARGA)

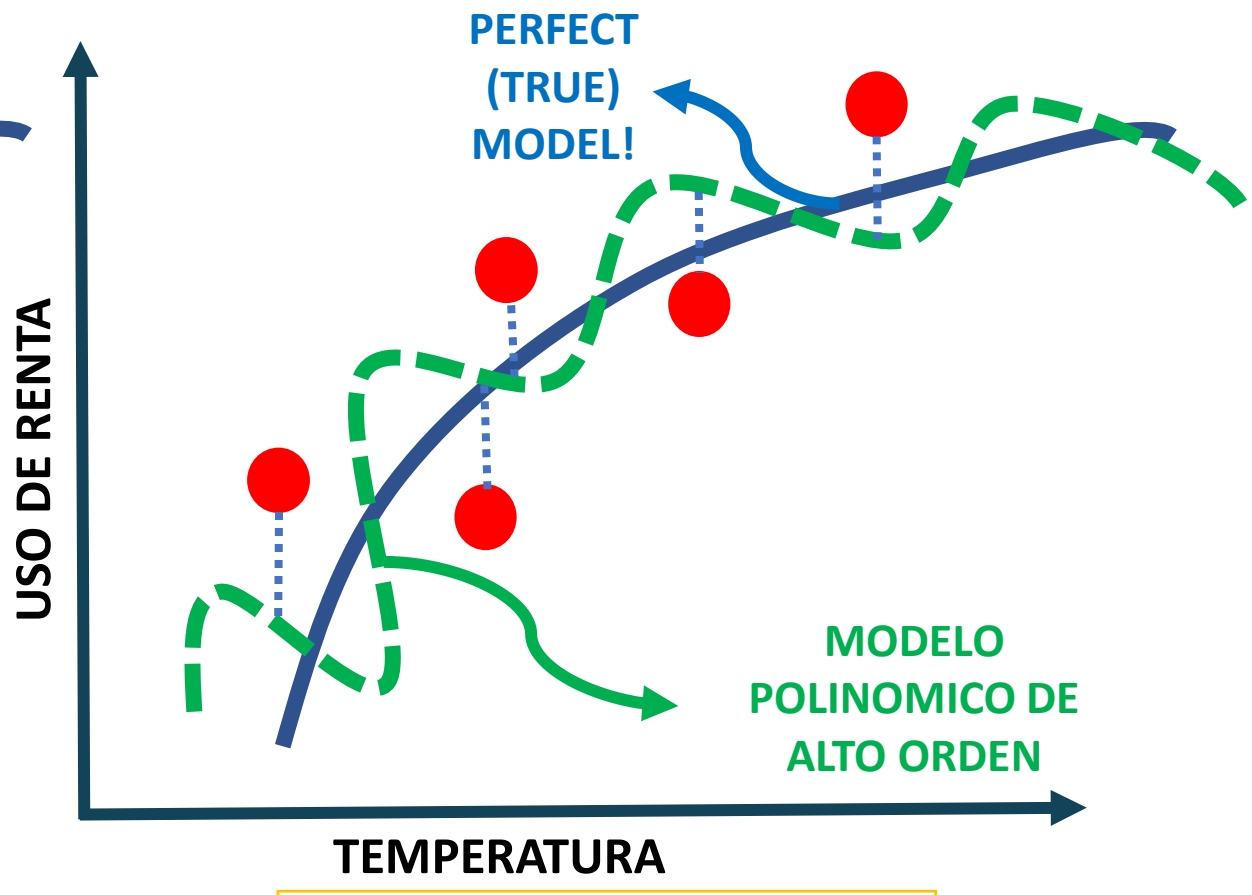


SUMA DE CUADRADO
(PEQUEÑO)

Sesgo y Varianza: Modelo #1 vs Modelo #2 Durante Prueba



SUMA DE CUADRADOS (PEQUEÑO)

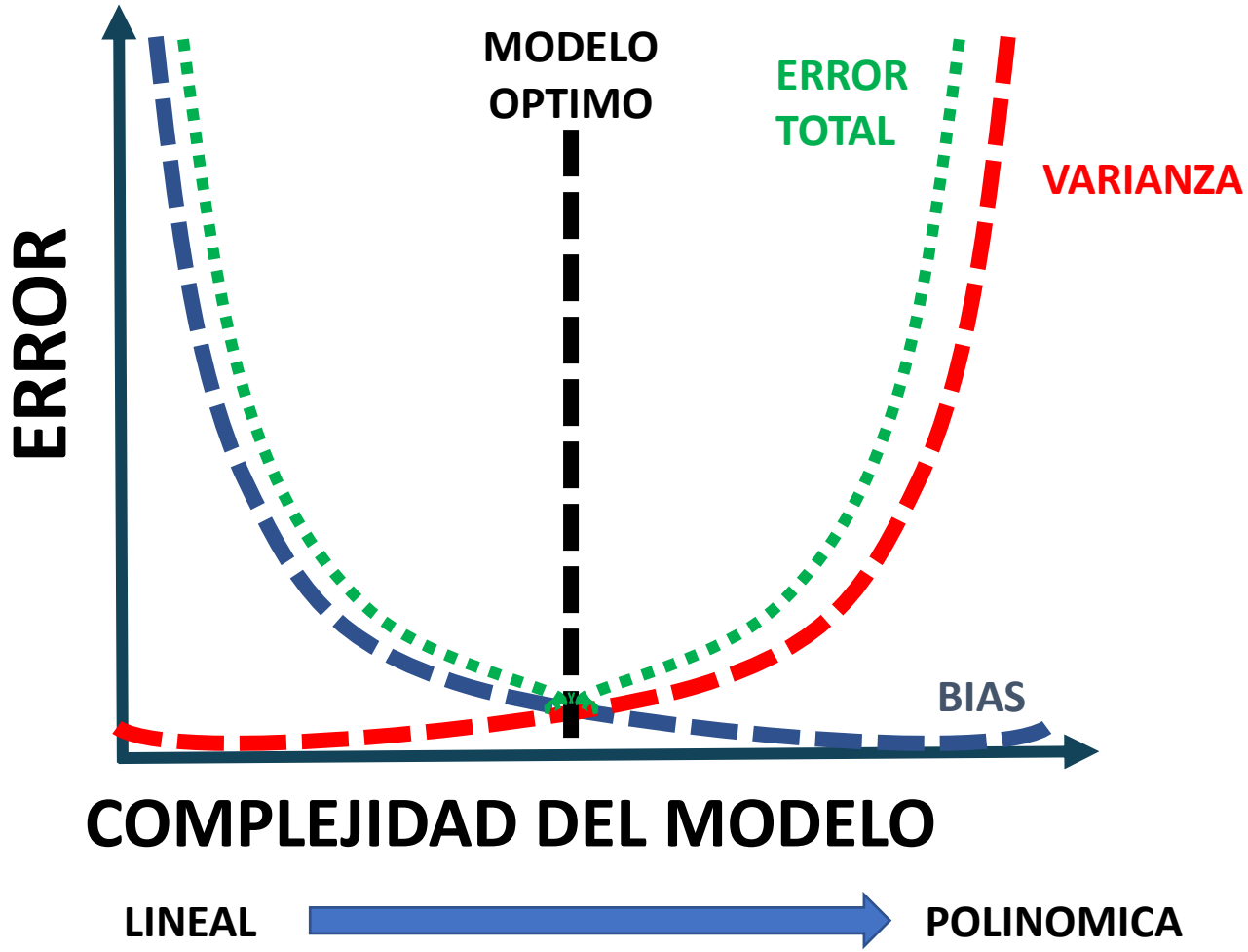


SUMA DE CUADRADOS (LARGO)

El modelo polinómico funciona mal en el conjunto de datos de prueba y por lo tanto tiene una gran variación

Complejidad vs Error del Modelo

- La regularización funciona reduciendo la variación a costa de añadir algún sesgo al modelo.
- Se produce una compensación entre la varianza y el sesgo



Complejidad vs Error del Modelo

MODELO #1 (REGRESION LINEAL) (SIMPLE)	MODELO #2 (POLINOMICO DE ALTO ORDEN) (COMPLEJO)
El modelo tiene un **alto sesgo** porque es muy rígido (no flexible) y no puede encajar bien en el conjunto de datos de entrenamiento.	El modelo tiene un **pequeño sesgo** porque es flexible y puede encajar muy bien en el conjunto de datos de entrenamiento.
Tiene una **pequeña variación (variabilidad)** porque puede ajustarse a los datos de capacitación y a los datos de las pruebas con un nivel similar (el modelo es capaz de generalizar mejor) y evita el sobreajuste	Tiene una **gran variación (variabilidad)** porque el modelo se ajusta en exceso al conjunto de datos de entrenamiento y tiene un mal desempeño en el conjunto de datos de prueba
El rendimiento es consistente entre el conjunto de datos de entrenamiento y el conjunto de datos de prueba	El rendimiento varía enormemente entre el conjunto de datos de entrenamiento y el conjunto de datos de prueba (alta variabilidad)
Buena Generalizacion	Sobre Ajustado (Overfitting)

Complejidad vs Error del Modelo

- La varianza mide la diferencia en los ajustes entre el conjunto de datos de entrenamiento y el conjunto de datos de prueba
- Si el modelo se generaliza mejor, el modelo tiene una pequeña variación, lo que significa que el rendimiento del modelo es consistente entre los conjuntos de datos de entrenamiento y prueba
- Si el modelo se ajusta al conjunto de datos de entrenamiento, el modelo tiene una gran variación

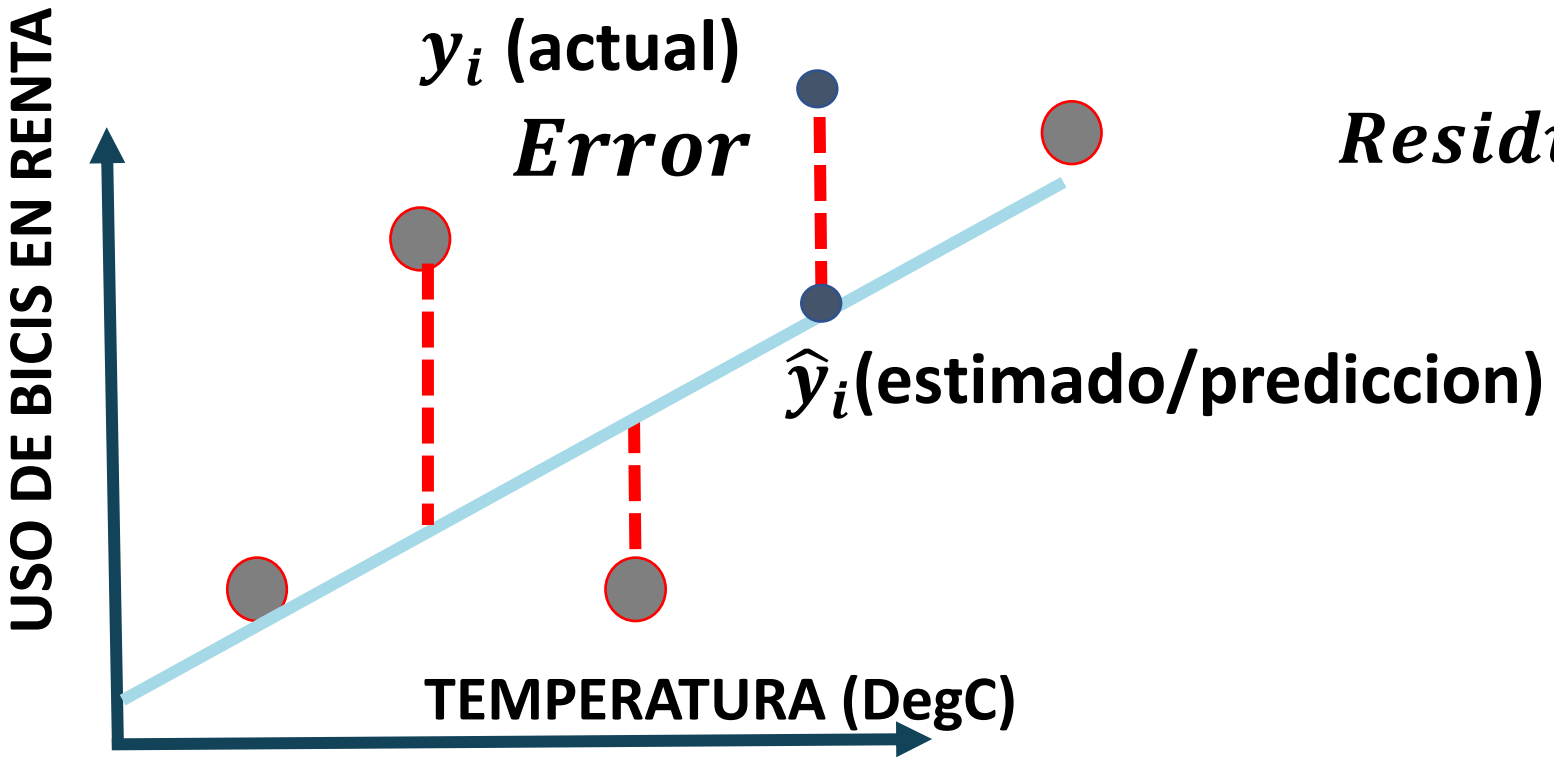
EL MODELO DE REGRESIÓN PERFECTA TENDRÁ UN PEQUEÑO SESGO Y UNA PEQUEÑA VARIABILIDAD!

SE REALIZARÁ UN BALANCE ENTRE EL SESGO Y LA VARIANZA PARA LOS RESULTADOS FINALES

Modelo de Evaluación del Rendimiento - Metrica

Metricas de Regresion: ¿Como Evaluar el Rendimiento de un Modelo?

- Después del ajuste del modelo, nos gustaría evaluar el rendimiento del mismo comparando las predicciones del modelo con los datos reales (Verdaderos)



$$Residuales\ (Error) = \hat{y}_i - y_i$$

Proyecto #3: Metricas de Regresion: Error Absoluto Medio (MAE)

- El error absoluto medio (MAE) se obtiene calculando la diferencia absoluta entre las predicciones del modelo y los valores verdaderos (reales)
- El MAE es una medida de la magnitud promedio del error generado por el modelo de regresión
- El error absoluto medio (MAE) se calcula de la siguiente manera:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

El MAE se calcula siguiendo estos pasos:

1. Calcular el residuo de cada punto de datos
2. Calcular el valor absoluto (para deshacerse del signo)
3. Calcular el promedio de todos los residuos

Si el MAE es cero, esto indica que las predicciones del modelo son perfectas.

Metricas de Regresion: Error del Cuadrado Medio (MSE)

- El Error Medio Cuadrado (MSE) es muy similar al Error Medio Absoluto (MAE), pero en lugar de utilizar valores absolutos, se calculan los cuadrados de la diferencia entre las predicciones del modelo y el conjunto de datos de entrenamiento (valores reales).
- Los valores del MSE son generalmente más grandes comparados con el MAE, ya que los residuos se están cuadrando.
- En el caso de los datos atípicos, la MSE será mucho mayor en comparación con el MAE
- En MSE, el error aumenta de forma cuadrática mientras que el error aumenta de forma proporcional en MAE
- En MSE, como el error se está cuadrando, cualquier error de predicción está siendo fuertemente penalizado
- El MSE se calcula de la siguiente manera:

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

- El MSE se calcula siguiendo estos pasos:
  1. Calcular el residuo para cada punto de datos
  2. Calcular el valor cuadrado de los residuos
  3. Calcular el promedio de los resultados del paso #2

Metricas de Regresion: Error Cuadratico Medio de la Raiz (RMSE)

- El Root Mean Square Error (RMSE) representa la desviación estándar de los residuos (es decir, las diferencias entre las predicciones del modelo y los valores reales (datos de entrenamiento)).
- El RMSE puede interpretarse fácilmente en comparación con el MSE porque las unidades del RMSE coinciden con las unidades de la salida.
- La RMSE proporciona una estimación del tamaño de la dispersión de los residuos.
- La RMSE se calcula de la siguiente manera:

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

La RMSE se calcula siguiendo estos pasos:

1. Calcular el residuo para cada punto de datos
2. Calcular el valor cuadrado de los residuos
3. Calcular el promedio de los residuos cuadrados
4. Obtener la raíz cuadrada del resultado

Metricas de Regresion: Porcentaje de Error Absoluto Medio (MAPE)

- Los valores de MAE pueden ir de 0 a infinito, lo que dificulta la interpretación del resultado en comparación con los datos de entrenamiento.
- El porcentaje de error absoluto medio (MAPE) es el equivalente al MAE pero proporciona el error en forma de porcentaje y por lo tanto supera las limitaciones del MAE.
- MAPE puede presentar algunas limitaciones si el valor del punto de datos es cero (ya que hay una operación de división implicada)
- El MAPE se calcula de la siguiente manera:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} |(y_i - \widehat{y}_i)/y_i|$$

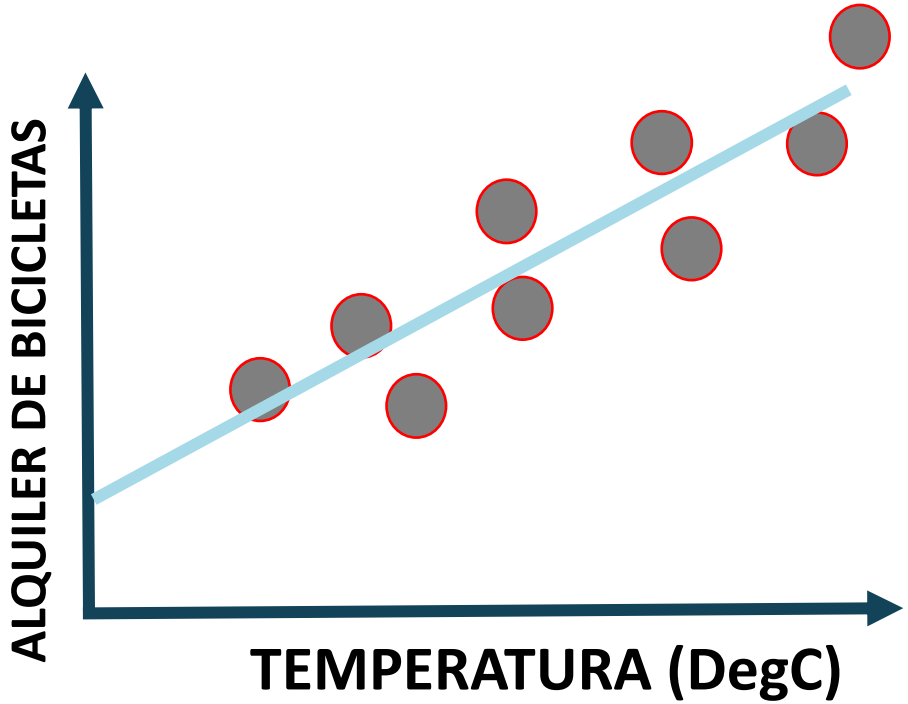
Metricas de Regresion: Porcentaje Medio de Error (MPE)

- El MPE es similar al MAPE pero sin la operación absoluta
- El MPE es útil para proporcionar una idea de cuántos errores positivos en comparación con los negativos
- El MPE se calcula de la siguiente manera:

$$MPE = \frac{100\%}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)/y_i$$

Metricas de Regresion: R Cuadrado (R^2)- Coeficiente de Determinacion

- R-cuadrado o el coeficiente de determinación representa la proporción de varianza (de y) que ha sido explicada por las variables independientes del modelo.
- Si R^2=80, esto significa que el 80% del incremento en el uso de la bicicleta se debe al aumento de la temperatura.



Metricas de Regresion: R Cuadrado (R^2)- Coeficiente de Determinacion

- R-cuadrado o el coeficiente de determinación representa la proporción de varianza (y) que ha sido explicada por las variables independientes (X) en el modelo.
- Proporciona una indicación de la bondad del ajuste y, por lo tanto, una medida de cuán bien es probable que el modelo prediga las muestras no vistas, a través de la proporción de la varianza explicada.
- La mejor puntuación posible es 1,0
- Un modelo constante que siempre predice el valor esperado de y, sin tener en cuenta las características de entrada, obtendría una puntuación R^2 de 0,0.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2}$$

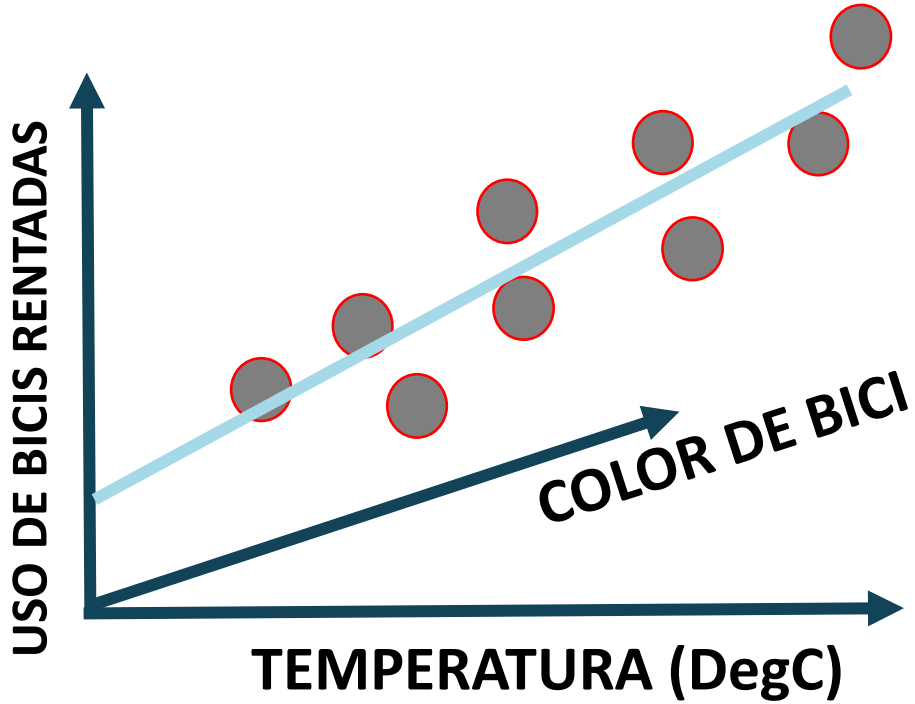
Metricas de Regresion: R Cuadrado (R^$_2$)- Coeficiente de Determinacion

- El cuadrado R representa la proporción de la varianza de la variable dependiente (y) que ha sido explicada por las variables independientes.
- R-cuadrado proporciona una visión de la bondad del ajuste.
- Da una medida de lo bien que las muestras no vistas pueden ser predichas por el modelo, a través de la proporción de la varianza explicada.
- El valor máximo de R^2 es 1
- Un modelo constante que siempre predice el valor esperado de y, sin tener en cuenta las características de entrada, tendrá una puntuación R^2 de 0,0.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2}$$

Metricas de Regresion: R Cuadrado Ajustado (R^2)

- Si R^2=80, esto significa que el 80% del incremento en el uso de la bicicleta de alquiler se debe al aumento de la temperatura.
- Añadamos otra variable independiente "inútil", digamos el color de la bicicleta en el eje Z.
- Ahora R^2 aumenta y se convierte en: R^2=85%



Metricas de Regresion: R Cuadrado Ajustado (R^2)

- Una limitación de R^2 es que aumenta al añadir variables independientes al modelo, lo cual es engañoso ya que algunas variables añadidas podrían ser inútiles con un significado mínimo.
- Ajustado R^2 supera este problema añadiendo una penalización si hacemos un intento de añadir una variable independiente que no mejore el modelo.
- Adjusted R^2 es una versión modificada de la R^2 y tiene en cuenta el número de predictores en el modelo.
- Si se añaden predictores inútiles al modelo, la versión ajustada de R^2 disminuirá
- Si se añaden predictores útiles al modelo, Ajustado R^2 aumentará
- K es el número de variables independientes y n es el número de muestras

$$R^2_{adjusted} = 1 - [\frac{(1 - R^2)(n - 1)}{n - k - 1}]$$