

Introducción al Deep Learning

Día 2: Introducción a las Redes Neuronales

Manuel Germán y David de la Rosa
Universidad de Jaén



Universidad
de Jaén



`(mgerman, drrosa)@ujaen.es`

Expectativas

Tras esta sesión, sabremos:

- Qué es un perceptrón multicapa (MLP).
- Cómo procesa una ANN sus entradas.
- Cómo se optimizan los parámetros de una ANN.
- Cómo evaluar el rendimiento de una ANN.
- ¿Por qué es esto posible?
 - *Teorema de la aproximación universal*
- Cómo implementar todo lo anterior usando *Pytorch* y *Pytorch Lightning*.

Cuestiones previas

1. ¿Qué es el *Deep Learning*?
2. ¿Qué es una neurona?
3. ¿Qué es una **función de activación**?
4. ¿Por qué es necesario **preprocesar** los datos?
5. ¿Qué es una red neuronal?

Deep learning

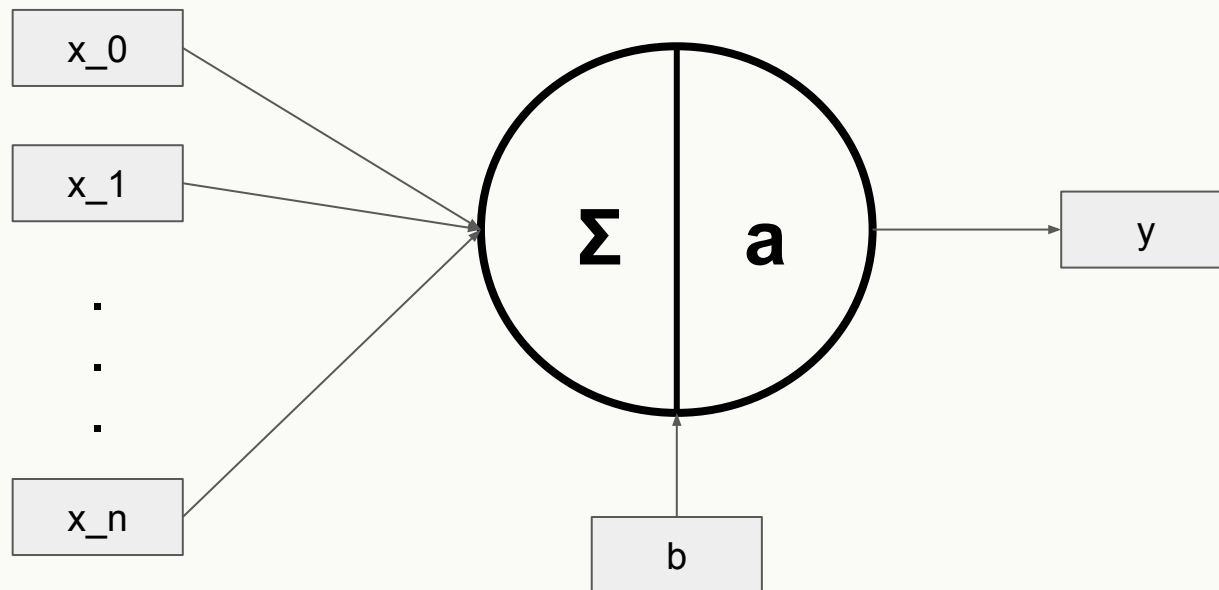
Yann LeCun^{1,2}, Yoshua Bengio³ & Geoffrey Hinton^{4,5}

LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015). <https://doi.org/10.1038/nature14539>

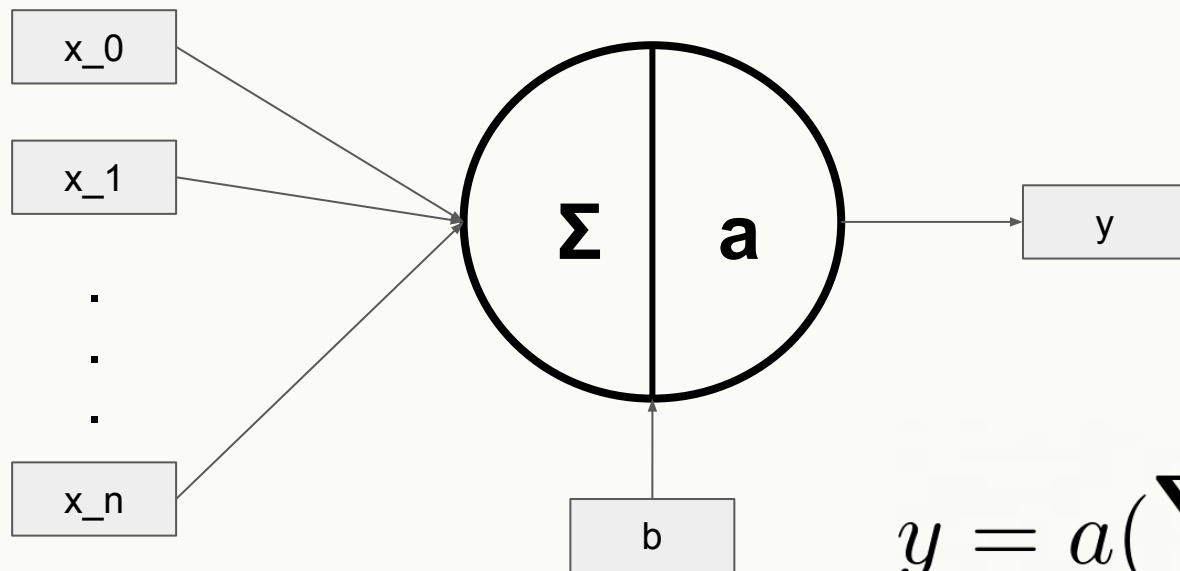
1

El perceptrón multicapa

La neurona, el elemento fundamental de la red

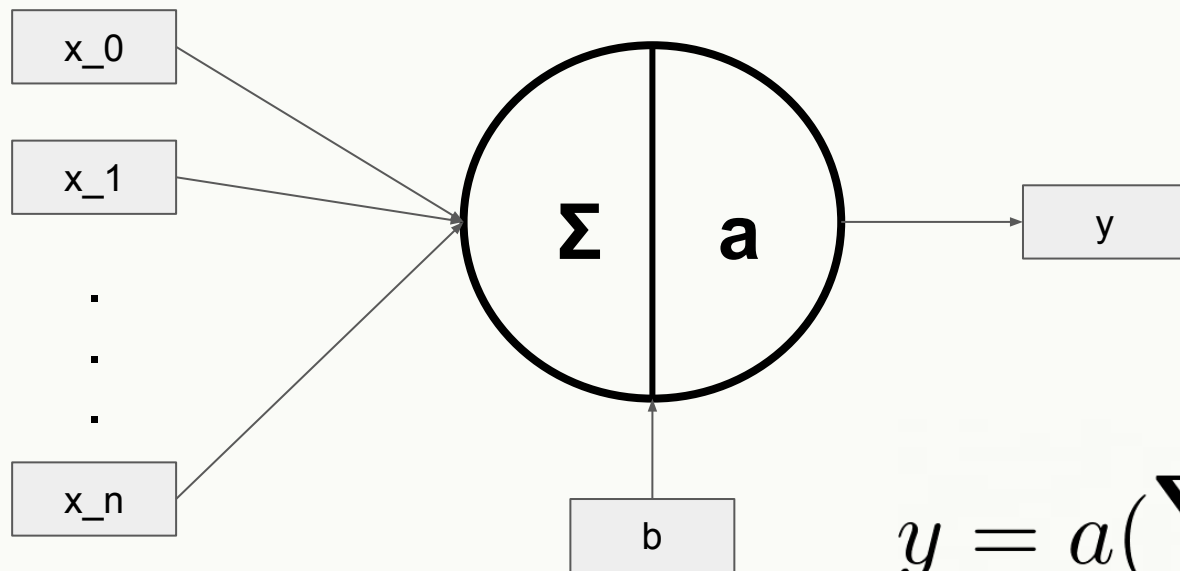


La neurona, el elemento fundamental de la red



$$y = a\left(\sum_{i=0}^n w_i \cdot x_i + b\right)$$

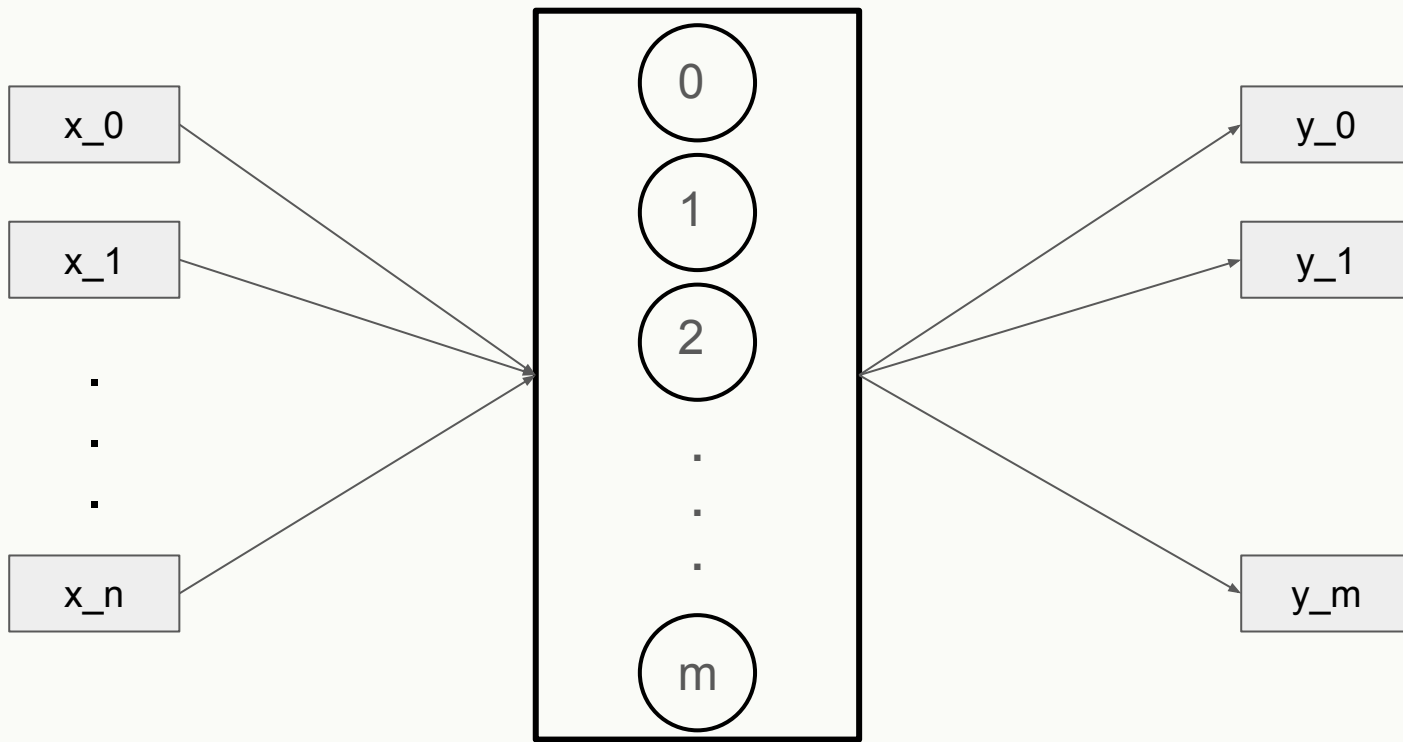
La neurona, el elemento fundamental de la red



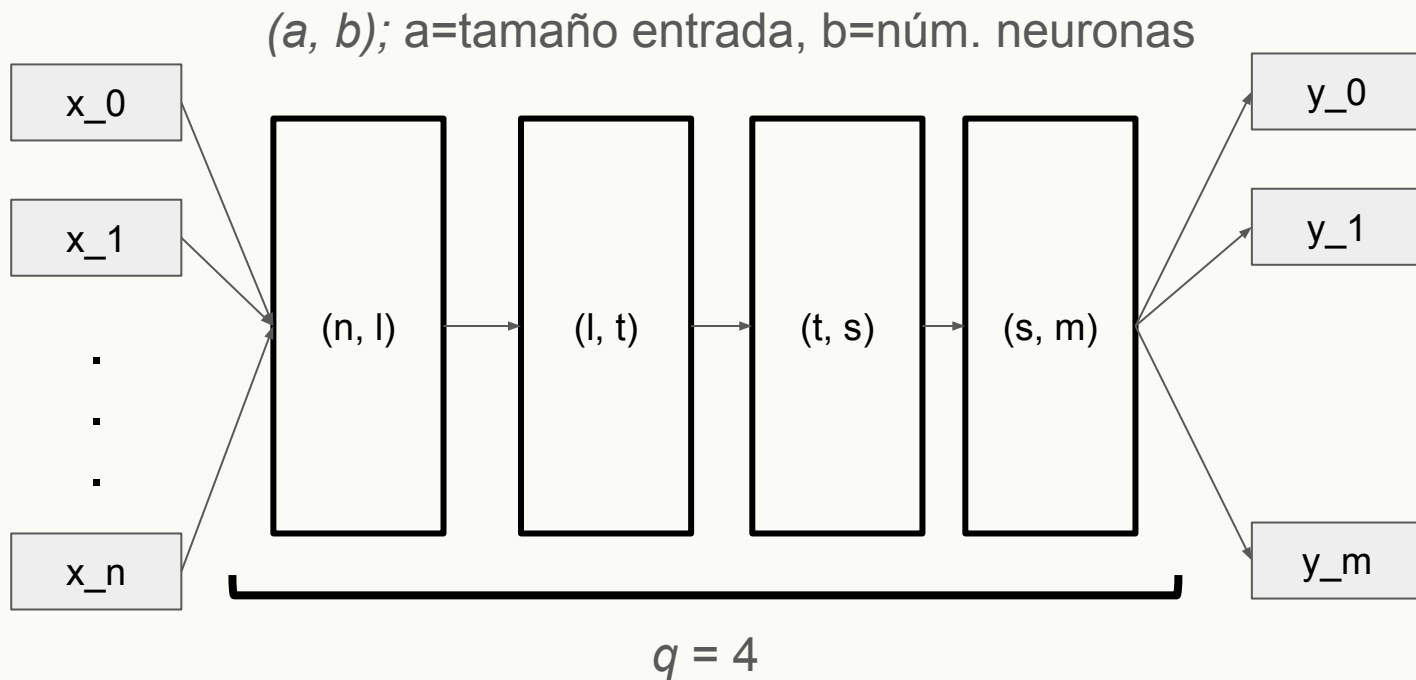
¿Cuál es su propósito?

$$y = a\left(\sum_{i=0}^n w_i \cdot x_i + b\right)$$

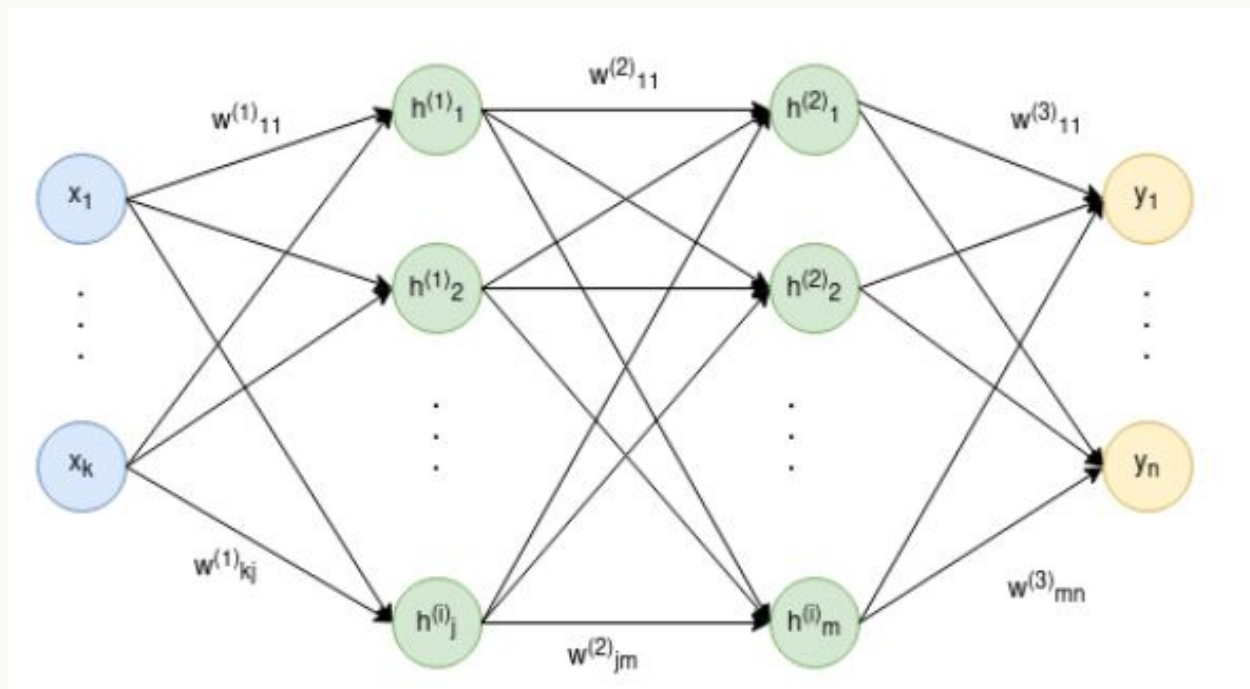
Una capa es un conjunto de m neuronas



Una red neuronal es un conjunto de q capas apiladas



Una red neuronal es un conjunto de q capas apiladas

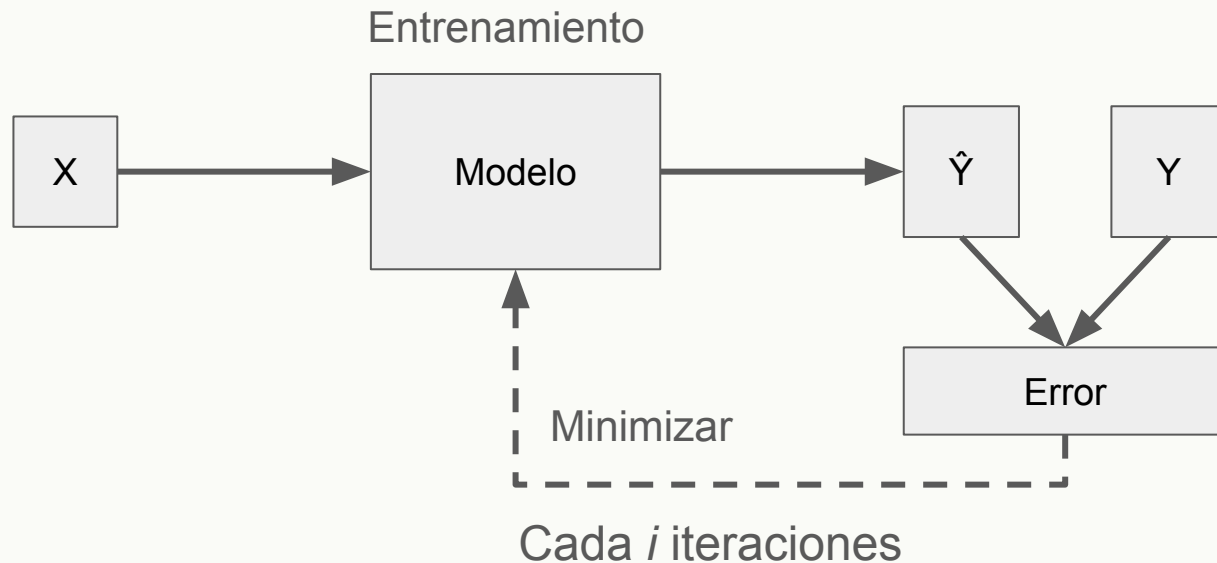


2

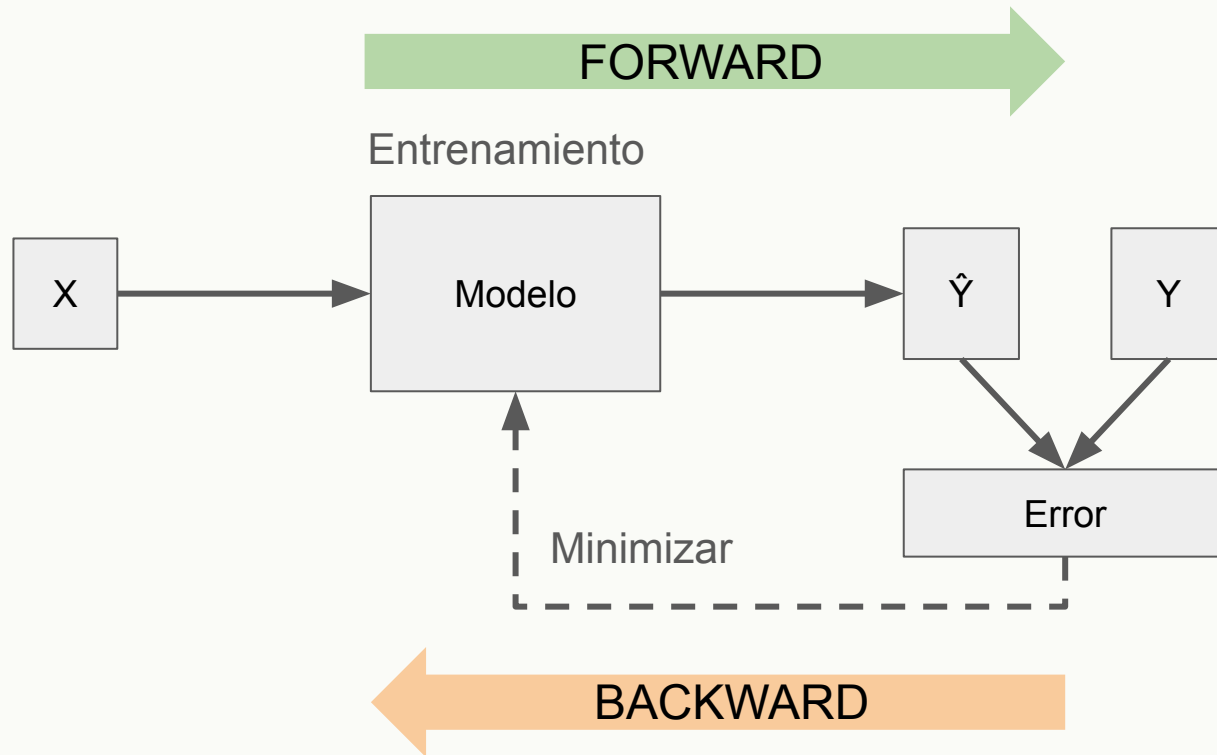
Funcionamiento

Esquema general

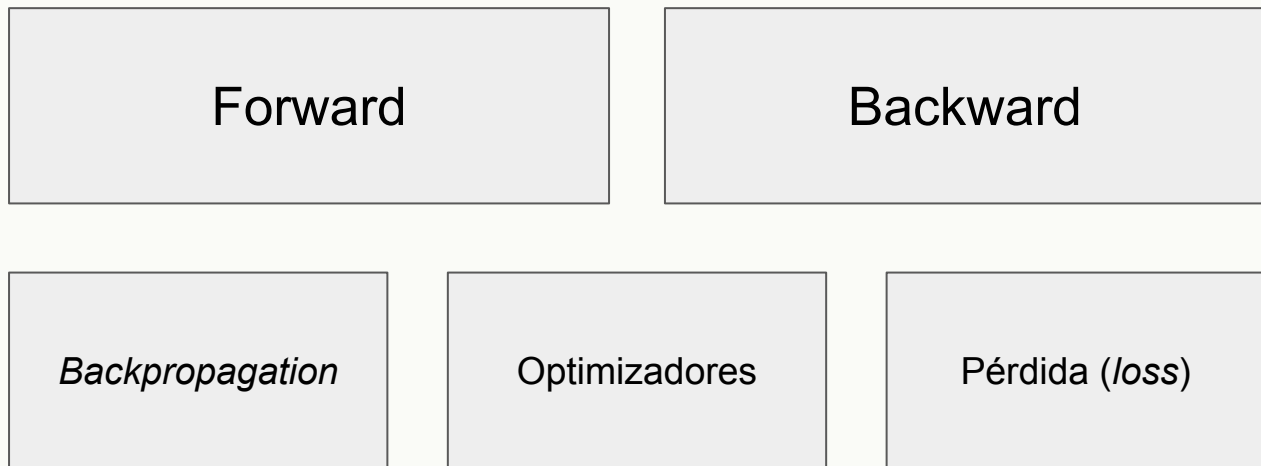
Durante n épocas



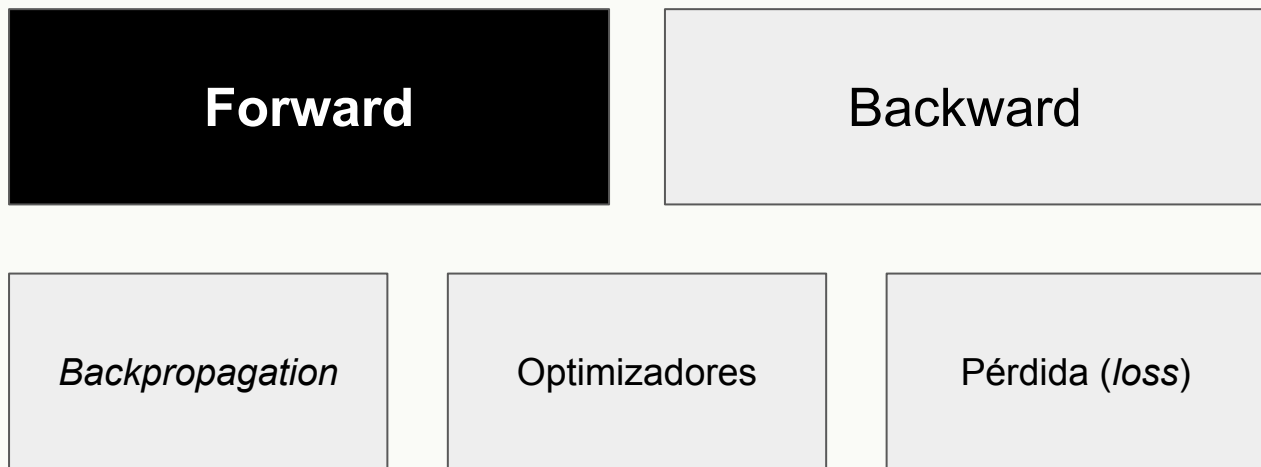
Esquema general



Conceptos clave

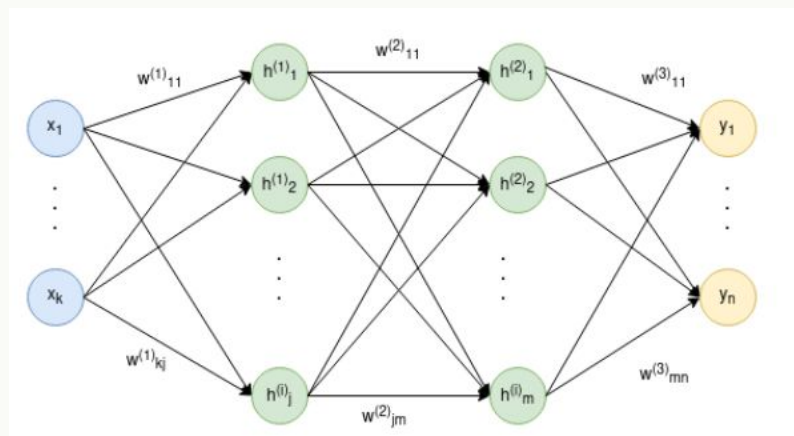


Conceptos clave



Paso hacia adelante (*Forward*)

Obtener la salida del modelo tras procesar los datos de entrada. (¡Fácil!)



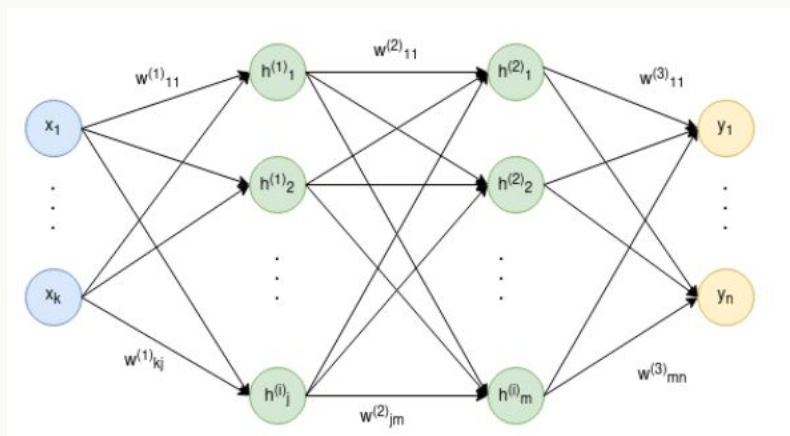
entrada = X

Para desde $i=1$ hasta $q-1$:

 entrada = procesar_capa(q , entrada)
devolver entrada

Paso hacia adelante (*Forward*)

Obtener la salida del modelo tras procesar los datos de entrada. (¡Fácil!)



entrada = X

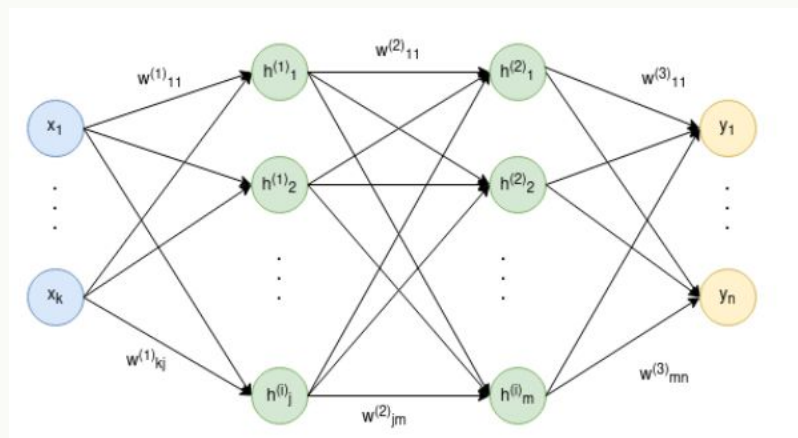
Para desde $i=1$ hasta $q-1$:

 entrada = procesar_capa(q , entrada)
devolver entrada

¿Cómo se implementaría eficientemente la función procesar_capa?

Paso hacia adelante (*Forward*)

Obtener la salida del modelo tras procesar los datos de entrada. (¡Fácil!)



multiplicación de matrices

entrada = X

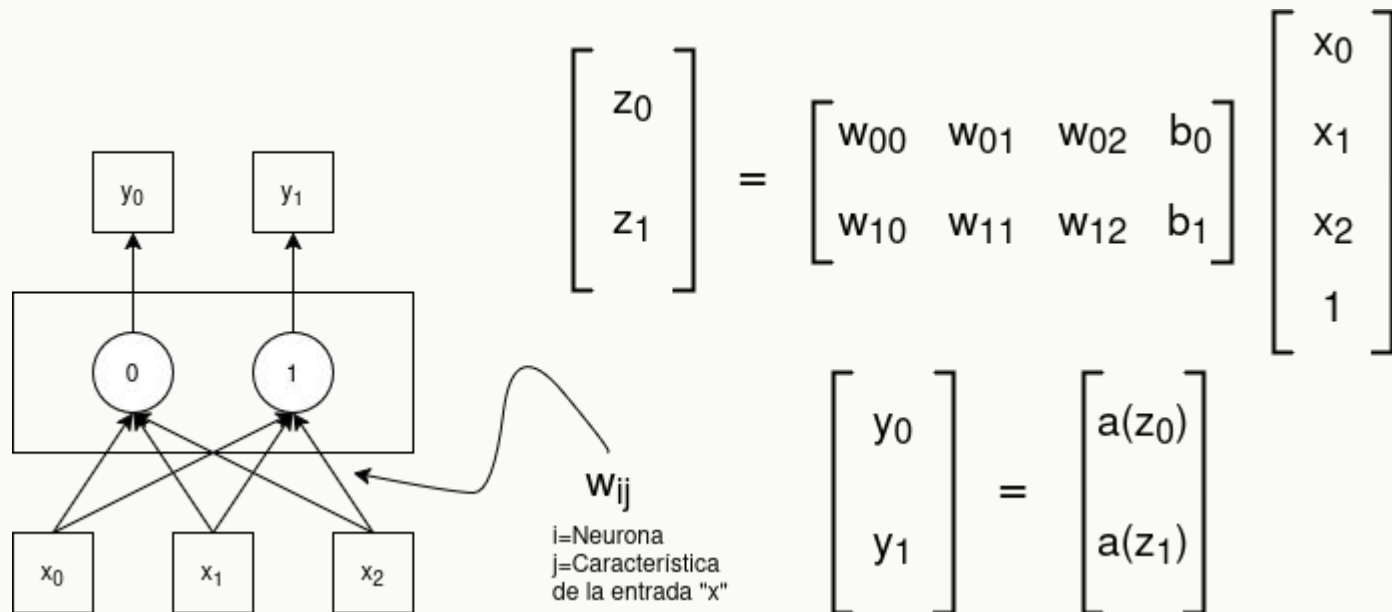
Para desde $i=1$ hasta $q-1$:

 entrada = procesar_capa(q , entrada)
devolver entrada

¿Cómo se implementaría eficientemente la función procesar_capa?

Paso hacia adelante (*Forward*)

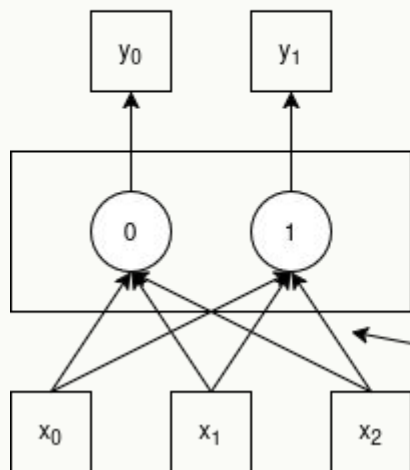
Una capa lineal se puede expresar como una matriz de pesos.



Paso hacia adelante (*Forward*)

Una capa lineal se puede expresar como una matriz de pesos.

$$Y = a(WX)$$



$$\begin{bmatrix} z_0 \\ z_1 \end{bmatrix}$$

=

$$\begin{bmatrix} w_{00} & w_{01} & w_{02} & b_0 \\ w_{10} & w_{11} & w_{12} & b_1 \end{bmatrix}$$

$$\begin{bmatrix} y_0 \\ y_1 \end{bmatrix}$$

$$\begin{bmatrix} y_0 \\ y_1 \end{bmatrix}$$

=

$$\begin{bmatrix} a(z_0) \\ a(z_1) \end{bmatrix}$$

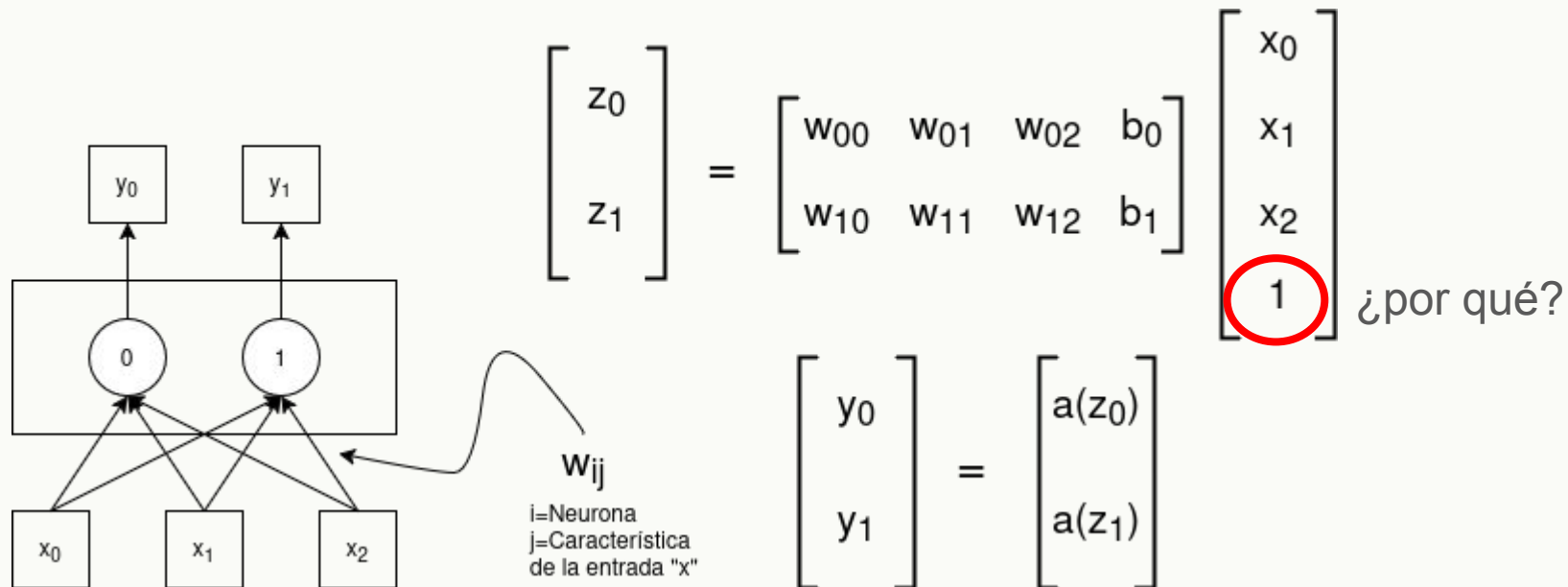
$$X$$

$$\begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ 1 \end{bmatrix}$$

W_{ij}
 i =Neurona
 j =Característica
 de la entrada "x"

Paso hacia adelante (*Forward*)

Una capa lineal se puede expresar como una matriz de pesos.



Paso hacia adelante (*Forward*)

La expresión matricial de una red permite procesar varias instancias en paralelo. Solamente es necesario añadir columnas a la matriz X con las características de las nuevas instancias.

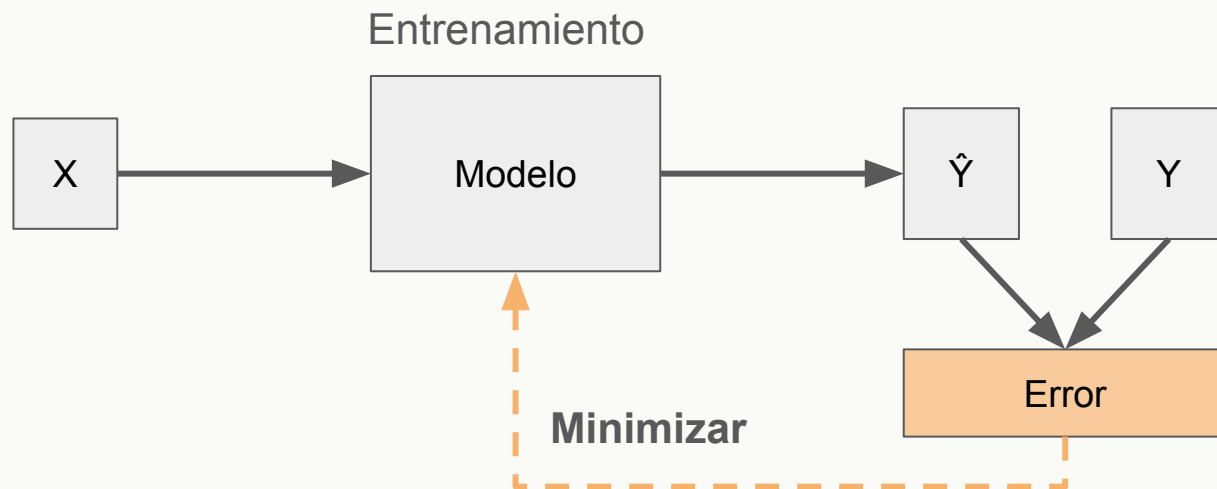
- A esto se le denomina *mini-batching*.
 - La mayoría de modelos actuales emplean lotes en su etapa de entrenamiento.
 - *Optimizar* el modelo para cada instancia puede ser costoso.
 - Al *optimizar* el modelo por lotes, se reducen los tiempos de entrenamiento.
 - Sin embargo, esta *optimización* está más sujeta a ruido y puede ocasionar problemas en función del tamaño del lote.

Conceptos clave



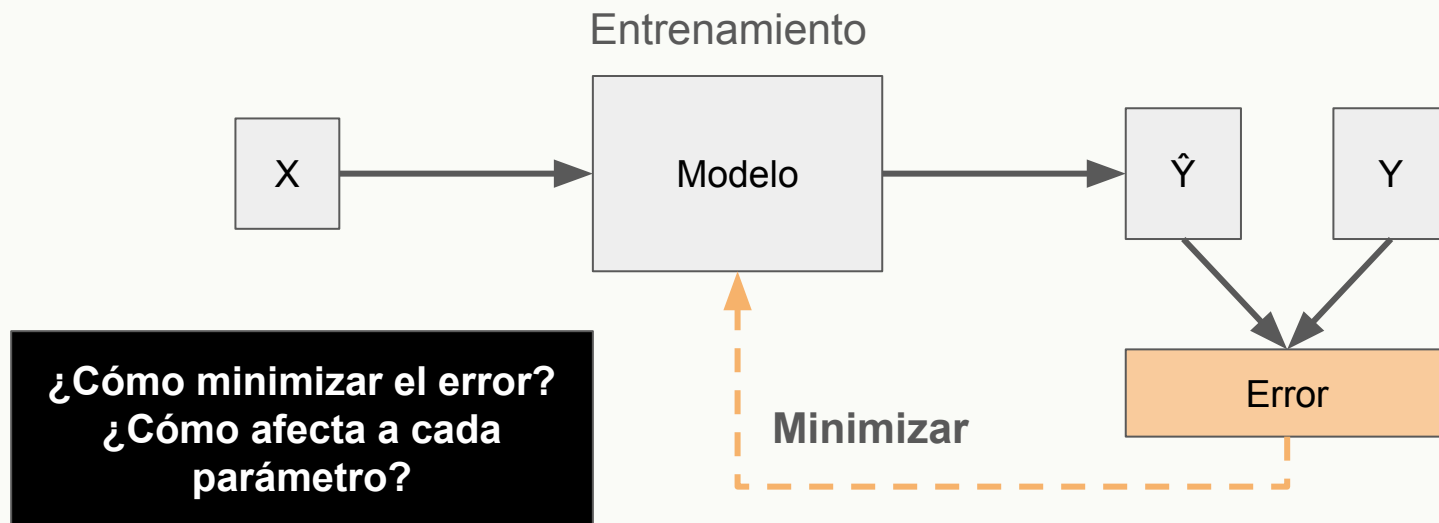
Paso hacia atrás (*Backward*)

Tenemos nuestra salida, ahora tenemos que **comprobar cuánto nos hemos equivocado** y **corregir dicho error**.



Paso hacia atrás (*Backward*)

Tenemos nuestra salida, ahora tenemos que **comprobar cuánto nos hemos equivocado** y **corregir dicho error**.

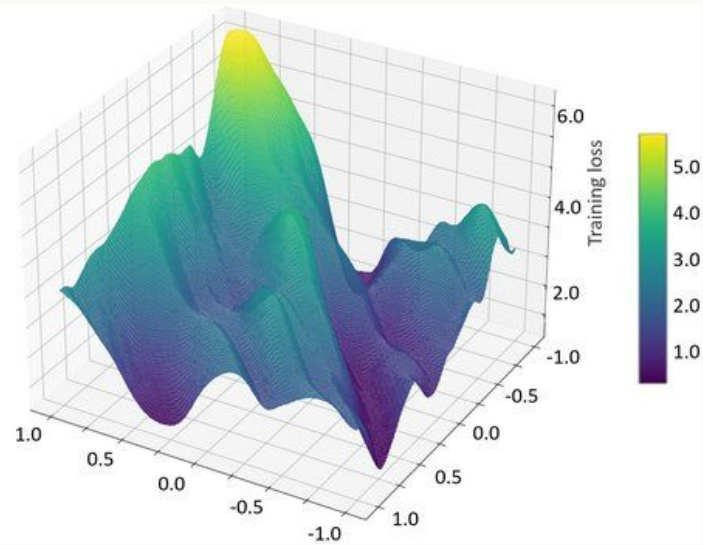
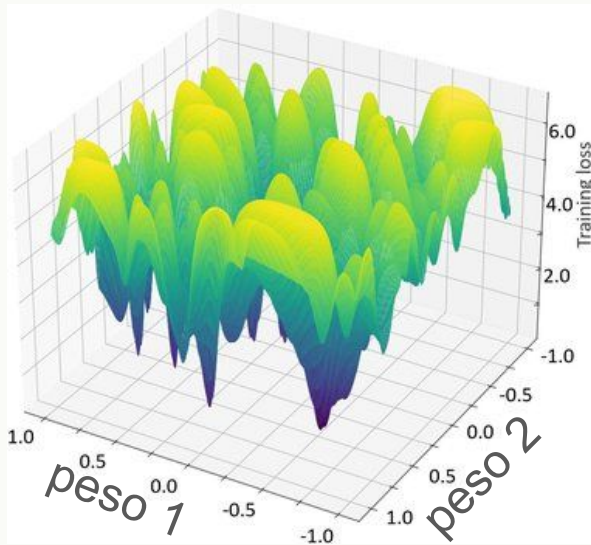


¿Cómo minimizar el error?

Problema de optimización en un espacio de búsqueda de alta dimensionalidad

Buscar el valor óptimo de los parámetros del modelo (W) tal que $L(W;x)$ sea mínimo.

Una simplificación



¿Cómo minimizar el error?

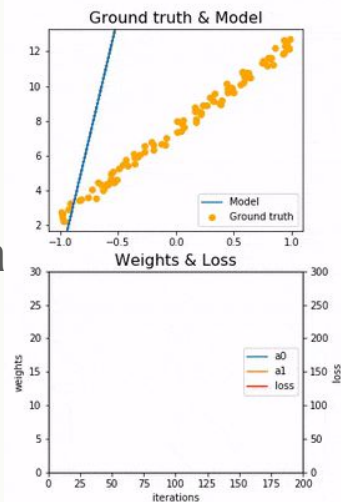
Gradiente descendiente

$$w_{k+1} = w_k - \alpha \nabla \mathcal{L}(w_k)$$

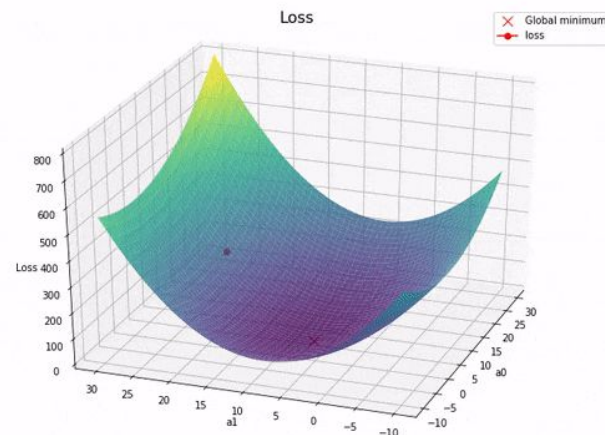
Esto hay que hacerlo para cada parámetro de la red...

- Derivar es costoso
- ¿Cómo calculo el gradiente para cada parámetro?

Buscar el valor óptimo de los parámetros del modelo (W) tal que $L(W)$ sea mínimo.



Stochastic Gradient Descent
epoch number: = 1



¿Cómo minimizar el error?

Gradiente descendiente

$$w_{k+1} = w_k - \alpha \nabla \mathcal{L}(w_k)$$

Buscar el valor óptimo de los parámetros del modelo (W) tal que $L(W)$ sea mínimo.

Esto hay que hacerlo para cada parámetro de la red...

- Derivar es costoso
- **¿Cómo calculo el gradiente para cada parámetro?**

Regla de la cadena y
Backpropagation
(Más info en el cuaderno)

Otros optimizadores

| Optimizer | State Memory [bytes] | # of Tunable Parameters | Strengths | Weaknesses |
|-------------------|----------------------|-------------------------|---|--|
| SGD | 0 | 1 | Often best generalization (after extensive training) | Prone to saddle points or local minima Sensitive to initialization and choice of the learning rate α |
| SGD with Momentum | $4n$ | 2 | Accelerates in directions of steady descent Overcomes weaknesses of simple SGD | Sensitive to initialization of the learning rate α and momentum β |
| AdaGrad | $\sim 4n$ | 1 | Works well on data with sparse features Automatically decays learning rate | Generalizes worse, converges to sharp minima Gradients may vanish due to aggressive scaling |
| RMSprop | $\sim 4n$ | 3 | Works well on data with sparse features Built in Momentum | Generalizes worse, converges to sharp minima |
| Adam | $\sim 8n$ | 3 | Works well on data with sparse features Good default settings Automatically decays learning rate α | Generalizes worse, converges to sharp minima Requires a lot of memory for the state |
| AdamW | $\sim 8n$ | 3 | Improves on Adam in terms of generalization Broader basin of optimal hyperparameters | Requires a lot of memory for the state |
| LARS | $\sim 4n$ | 3 | Works well on large batches (up to 32k) Counteracts vanishing and exploding gradients Built in Momentum | Computing norm of gradient for each layer can be inefficient |

<https://www.lightly.ai/blog/which-optimizer-should-i-use-for-my-machine-learning-project>

Funciones de pérdida

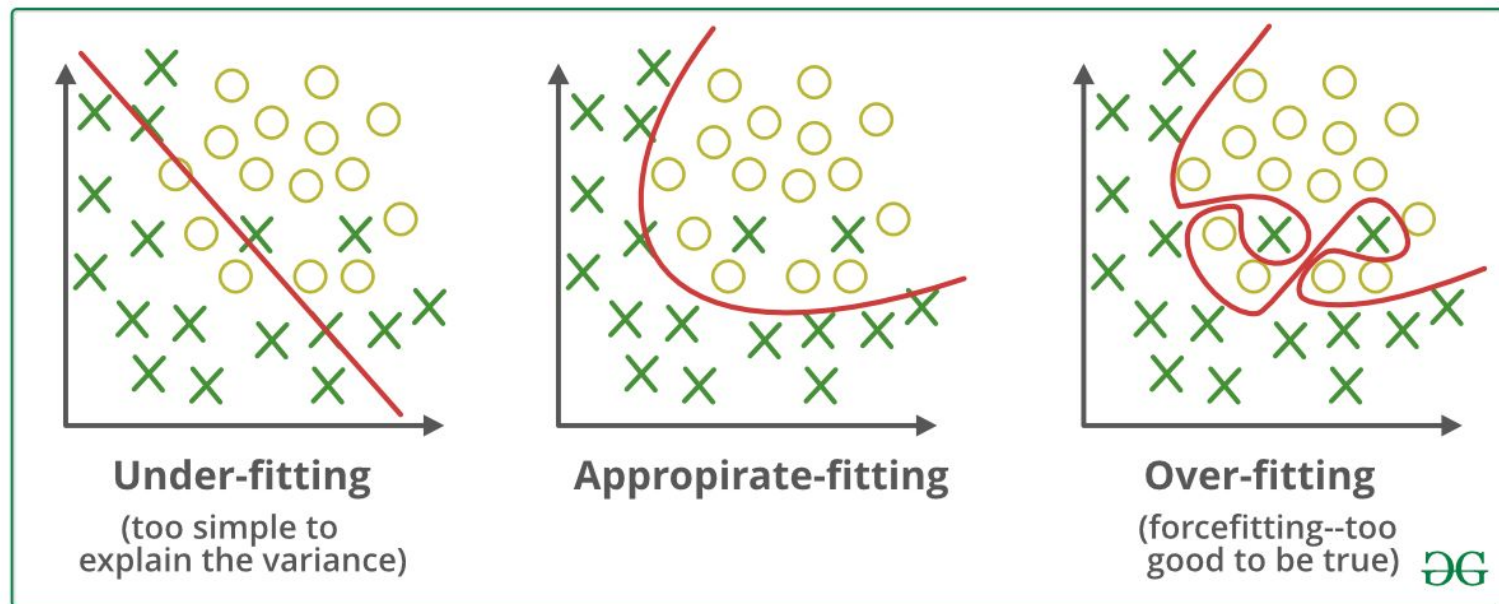
Depende del tipo de tarea en la que estemos trabajando.

<https://docs.pytorch.org/docs/stable/nn.html#loss-functions>

3

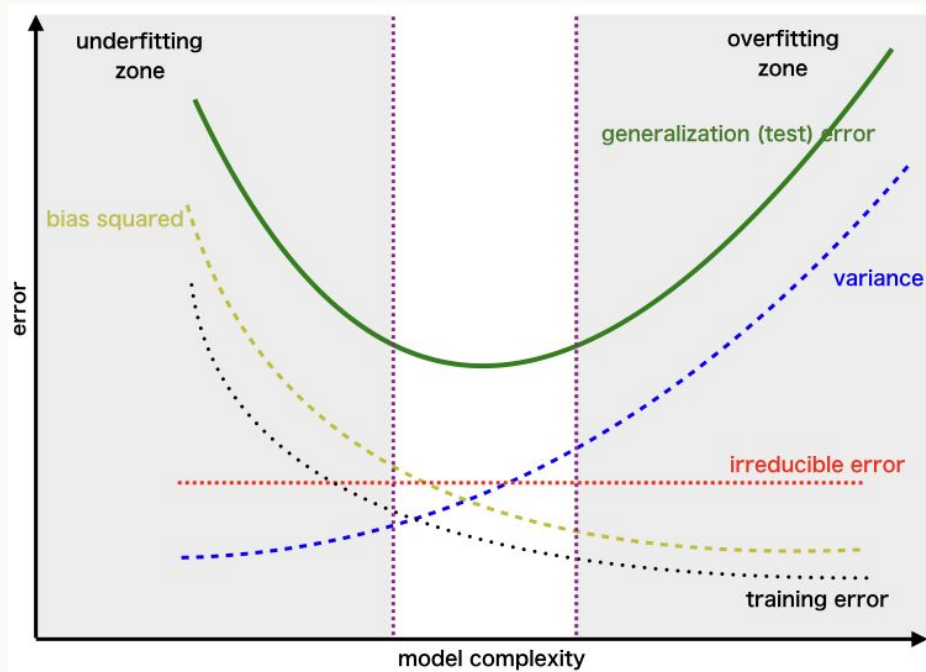
Evaluación

¿Nuestro modelo es correcto?



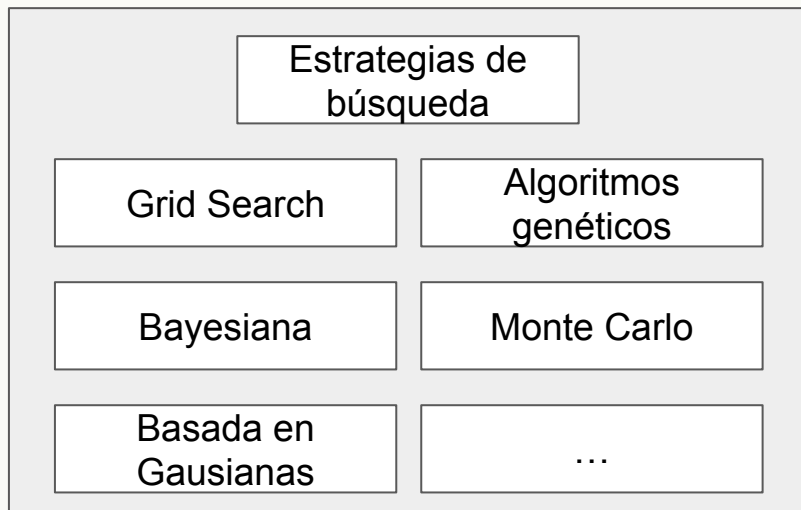
¿Nuestro modelo es correcto?

Bias-Variance trade-off
Balance sesgo-varianza



Optimización de hiper-parámetros

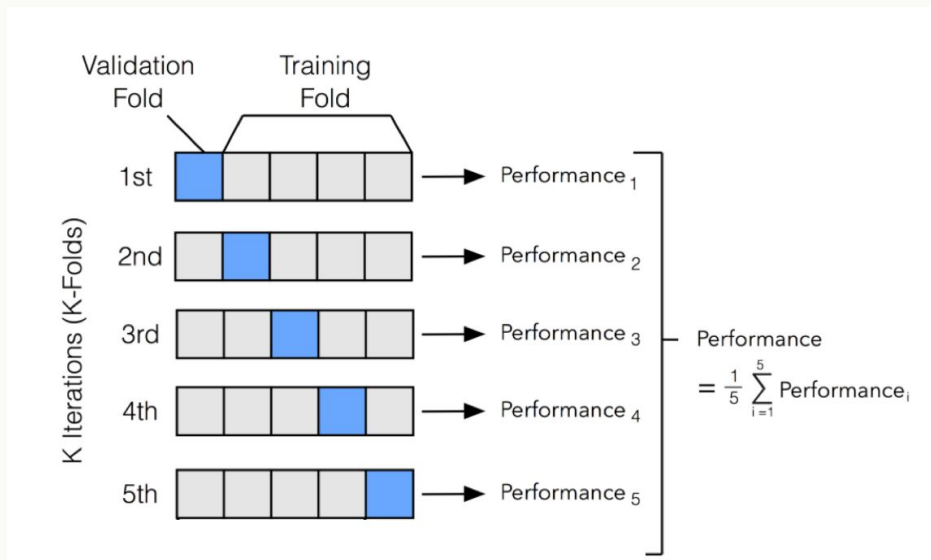
Un **hiper-parámetro** es un parámetro que controla de forma directa la estructura, funciones y rendimiento de los modelos. Es necesario encontrar su mejor configuración para obtener un buen rendimiento en cada problema.



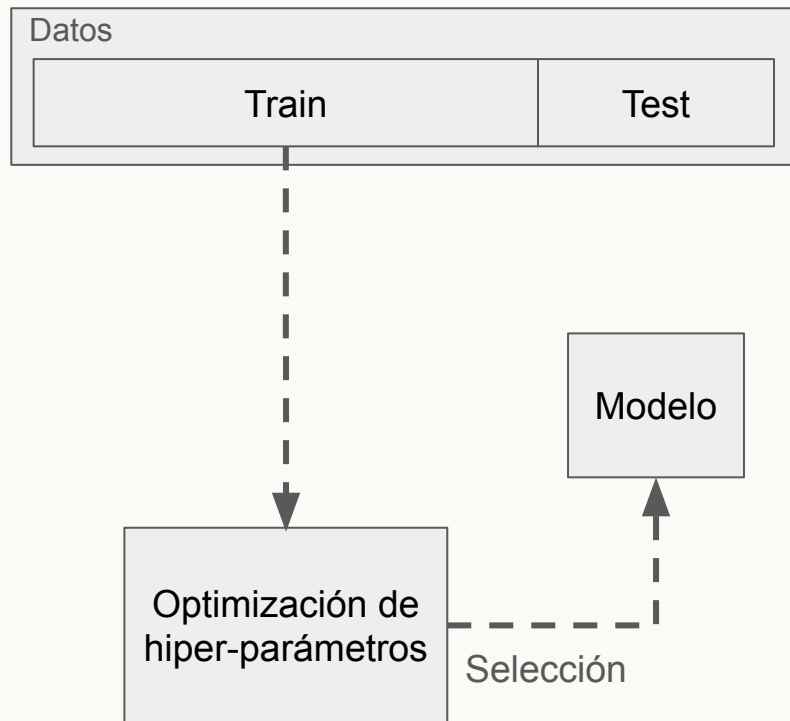
https://optuna.readthedocs.io/en/stable/tutorial/10_key_features/03_efficient_optimization_algorithms.html

Optimización de hiper-parámetros

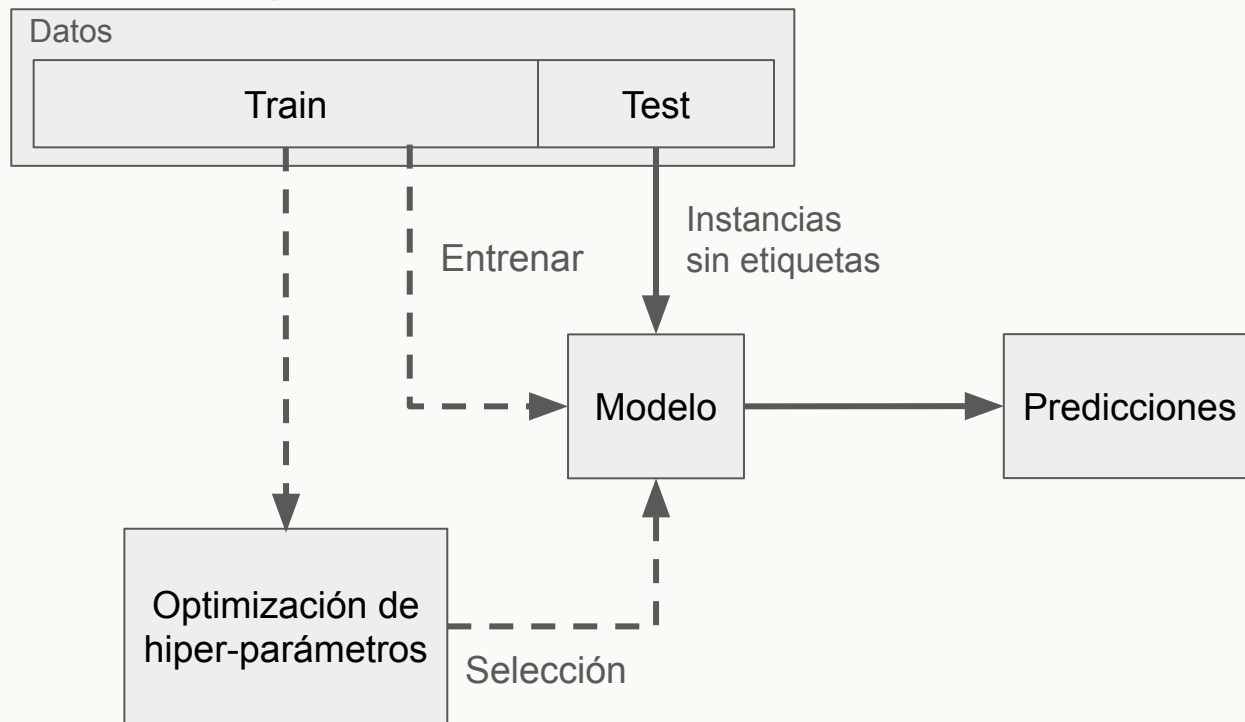
La **validación cruzada** permite evaluar el rendimiento del modelo sobre **todo** el conjunto de entrenamiento. Es normalmente utilizada, junto a una estrategia de búsqueda, para encontrar el mejor conjunto de hiper-parámetros.



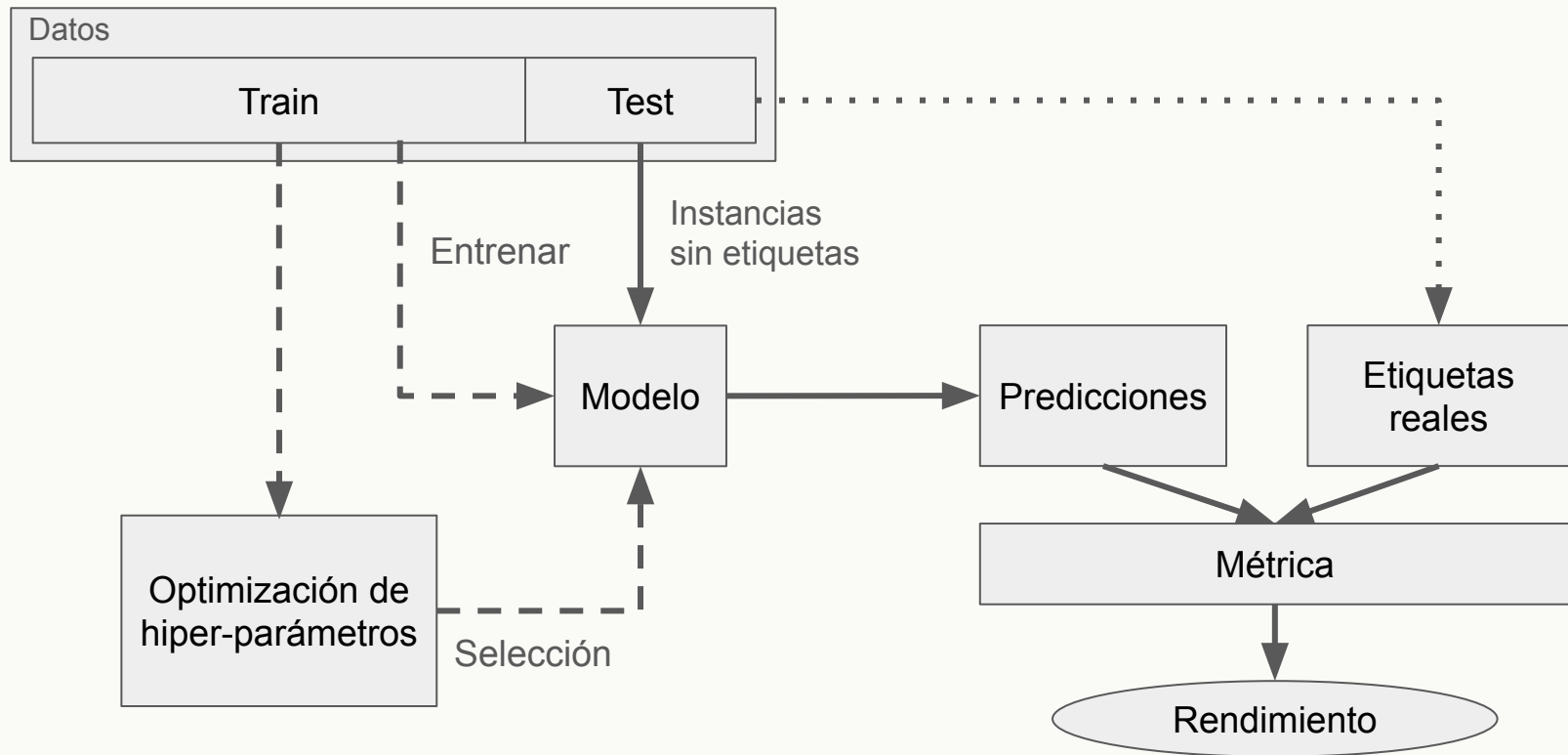
Proceso general



Proceso general



Proceso general

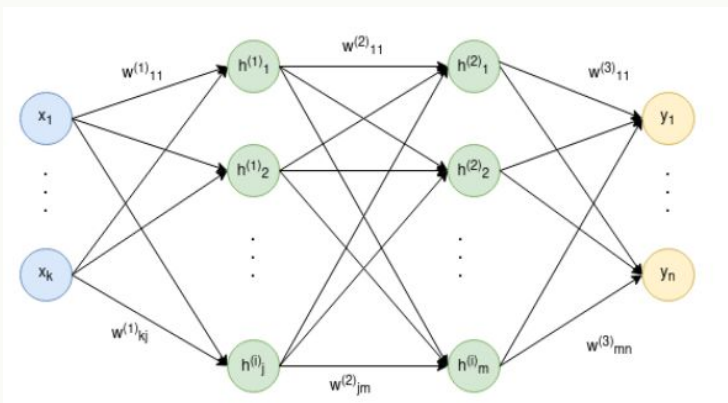
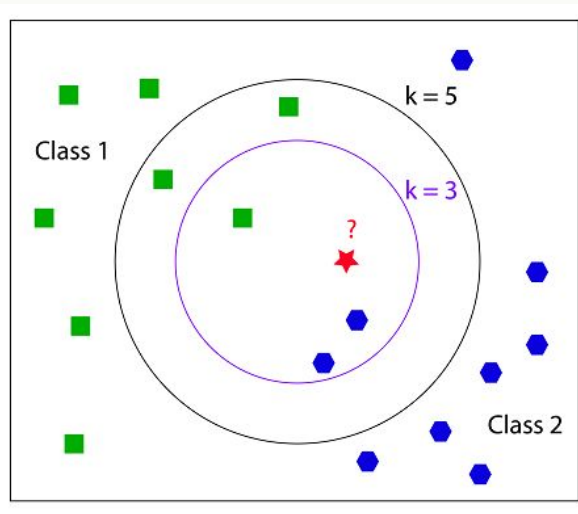


4

Otros aspectos
a considerar

Interpretabilidad

La **interpretabilidad** se refiere al **grado de claridad** que ofrece un sistema para entender sus decisiones. Está relacionada con la **opacidad** del modelo.



IA Explicable

Entre todas las definiciones existentes, la más adecuada creemos que es:

“Dada una **audiencia**, una IA Explicable es aquella que produce detalles o razones para que su funcionamiento sea fácil de entender” (*Arrieta et al., 2020*)

A diferencia de otras, se tiene en cuenta el **usuario del sistema**, alineándose con marcos sociotecnológicos que diferencian la XAI en dos grandes conjuntos:

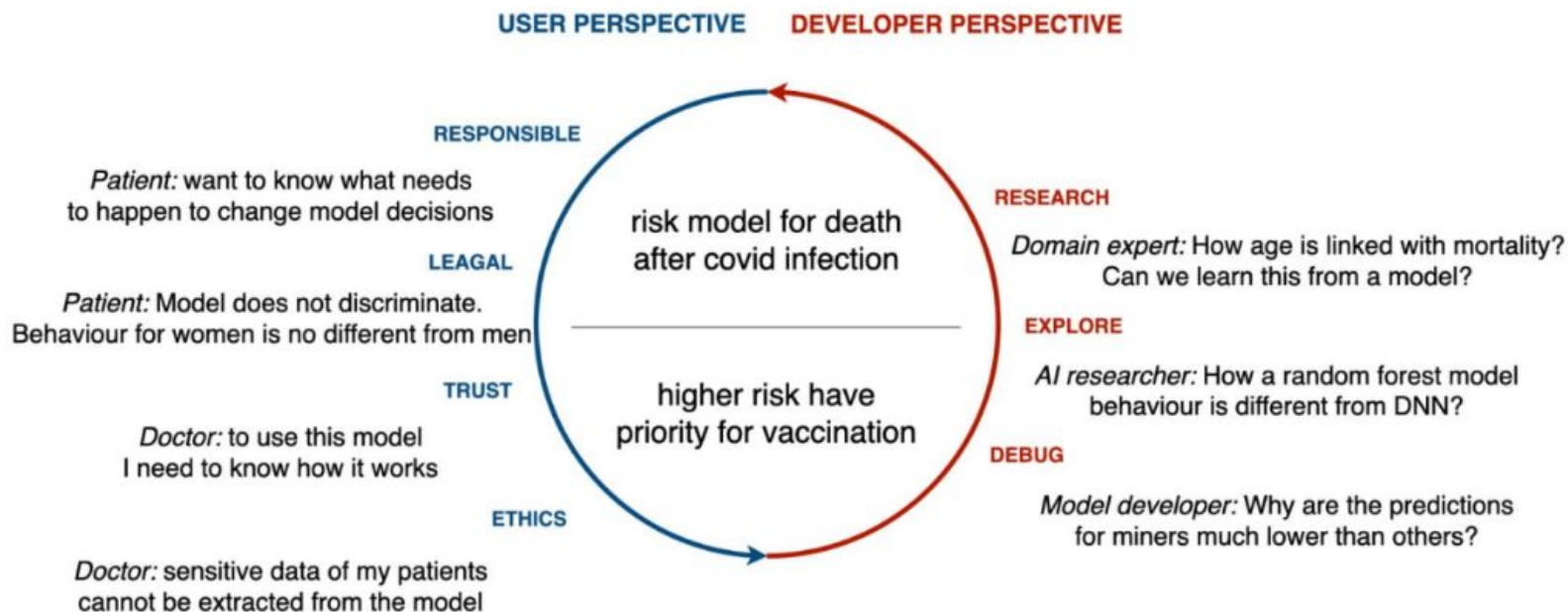
- **RED** (Research, Explore, Debug)
- **BLUE** (responsiBle, Legality, trUst, Ethics)

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, 82-115.

Herrera, F. (2025). Making Sense of the Unsensible: Reflection, Survey, and Challenges for XAI in Large Language Models Toward Human-Centered AI. arXiv preprint arXiv:2505.20305.

RED y BLUE XAI

Position: Explain to Question not to Justify





Biecek, P., & Samek, W. (2024). Position: Explain to question not to justify. arXiv preprint arXiv:2402.13914.

¿Cómo dotar de explicabilidad a un modelo?

Existen **técnicas** que permiten determinar qué elementos ha considerado el modelo para emitir cierta **decisión**.

| | SÍ | NO |
|---------------------------------------|---|---------------------------------------|
| ¿Se aplica tras obtener el resultado? | Post-hoc (Mapas de saliencia) | Ante-hoc (Modelo interpretable) |
| ¿Depende del modelo? | Model-dependent (Mapas de saliencia) | Model-agnostic (Basadas en reglas) |

Técnicas de explicabilidad

| TABULAR | IMAGE | TEXT | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|--|--|---|---------------|-------|--------------|------|--------------|------|--|--|--|-----|-------|----|-----|------|-----|-----|------|-----|-----|-----|-----|-----|-------|-----|------|-----|-----|-----|-----|----|-----|-----|------|-----|-----|-----|-----|-----|-----|-----|------|-----|-----|------|-----|-----|-----|-----|------|-----|-----|-----|-----|-----|-----|-----|------|
| <p>Rule-Based (RB)</p> <p>A set of premises that the record must satisfy in order to meet the rule's consequence.</p> $r = Education \leq College$ $\rightarrow \leq 50k$ | <p>Saliency Maps (SM)</p> <p>A map which highlight the contribution of each pixel at the prediction.</p>  | <p>Sentence Highlighting (SH)</p> <p>A map which highlight the contribution of each word at the prediction.</p> <p>the movie is not that bad</p> | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| <p>Feature Importance (FI)</p> <p>A vector containing a value for each feature. Each value indicates the importance of the feature for the classification.</p> <table><tr><td>capitalgain</td><td>0.00</td></tr><tr><td>education-num</td><td>14.00</td></tr><tr><td>relationship</td><td>1.00</td></tr><tr><td>hoursperweek</td><td>3.00</td></tr></table> | capitalgain | 0.00 | education-num | 14.00 | relationship | 1.00 | hoursperweek | 3.00 | <p>Concept Attribution (CA)</p> <p>Compute attribution to a target “concept” given by the user. For example, how sensitive is the output (a prediction of zebra) to a concept (the presence of stripes)?</p>  | <p>Attention Based (AB)</p> <p>This type of explanation gives a matrix of scores which reveal how the word in the sentence are related to each other.</p> <table><tr><td></td><td>the</td><td>movie</td><td>is</td><td>not</td><td>that</td><td>bad</td></tr><tr><td>the</td><td>High</td><td>Low</td><td>Low</td><td>Low</td><td>Low</td><td>Low</td></tr><tr><td>movie</td><td>Low</td><td>High</td><td>Low</td><td>Low</td><td>Low</td><td>Low</td></tr><tr><td>is</td><td>Low</td><td>Low</td><td>High</td><td>Low</td><td>Low</td><td>Low</td></tr><tr><td>not</td><td>Low</td><td>Low</td><td>Low</td><td>High</td><td>Low</td><td>Low</td></tr><tr><td>that</td><td>Low</td><td>Low</td><td>Low</td><td>Low</td><td>High</td><td>Low</td></tr><tr><td>bad</td><td>Low</td><td>Low</td><td>Low</td><td>Low</td><td>Low</td><td>High</td></tr></table> | | the | movie | is | not | that | bad | the | High | Low | Low | Low | Low | Low | movie | Low | High | Low | Low | Low | Low | is | Low | Low | High | Low | Low | Low | not | Low | Low | Low | High | Low | Low | that | Low | Low | Low | Low | High | Low | bad | Low | Low | Low | Low | Low | High |
| capitalgain | 0.00 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| education-num | 14.00 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| relationship | 1.00 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| hoursperweek | 3.00 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | the | movie | is | not | that | bad | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| the | High | Low | Low | Low | Low | Low | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| movie | Low | High | Low | Low | Low | Low | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| is | Low | Low | High | Low | Low | Low | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| not | Low | Low | Low | High | Low | Low | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| that | Low | Low | Low | Low | High | Low | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| bad | Low | Low | Low | Low | Low | High | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Técnicas de explicabilidad

Prototypes (PR)

The user is provided with a series of examples that characterize a class of the black box

$p = \text{Age} \in [35, 60], \text{Education} \in [\text{College}, \text{Master}] \rightarrow \geq 50k$

$p =$

 \rightarrow


$p = \text{"... not bad ..."} \rightarrow \text{"positive"}$


Counterfactuals (CF)

The user is provided with a series of examples similar to the input query but with different class prediction

$q = \text{Education} \leq \text{College} \rightarrow \leq 50k$

$c = \text{Education} \geq \text{Master} \rightarrow \geq 50k$

$q =$

 $\rightarrow \text{"3"}$

$c =$

 $\rightarrow \text{"8"}$

$q =$
 The movie is not that bad \rightarrow "positive"

$c =$
 The movie is that bad \rightarrow "negative"

Librerías y marcos populares



“Querida comunidad XAI, ¡tenemos que hablar!”

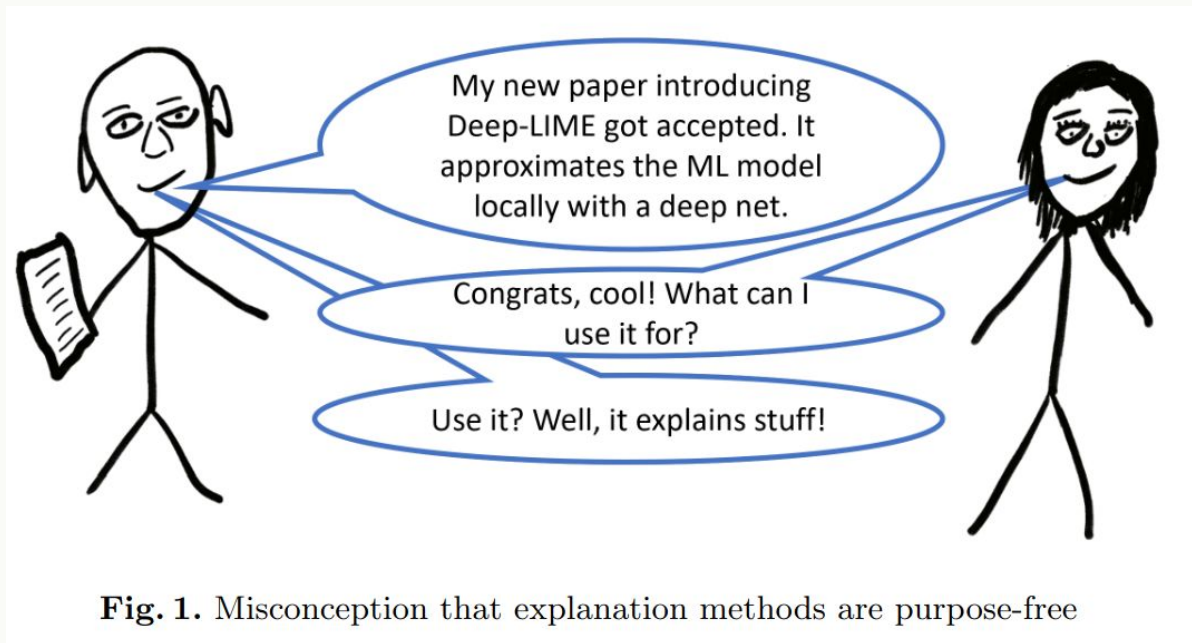


*One technique to explain it all,
One technique to find bugs,
One technique to convince them all
and in the black-box bind them.*

Fig. 2. Misconception that there is one true explanation technique

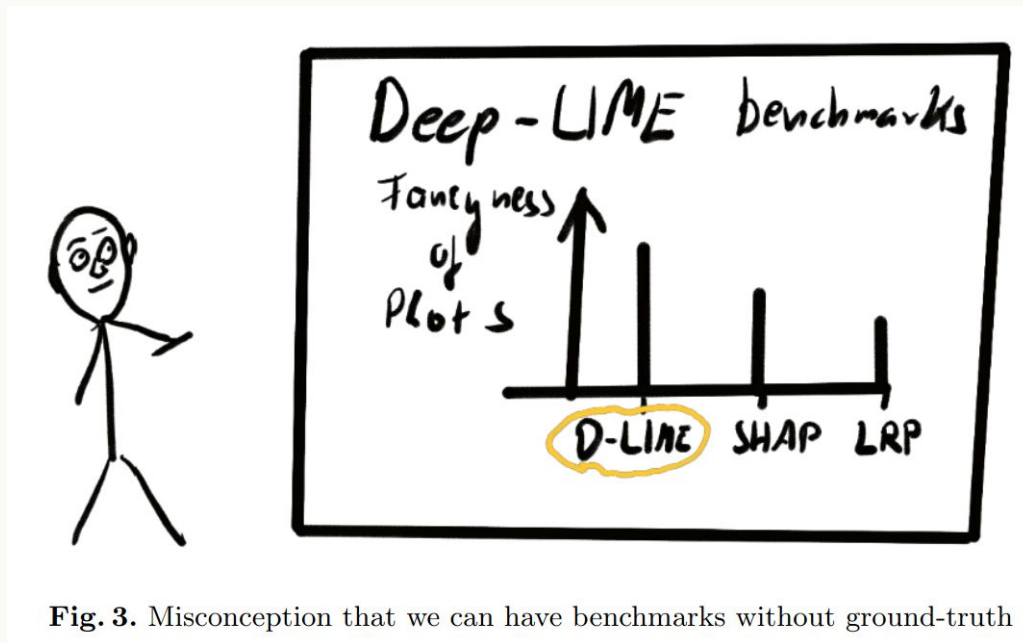
Freiesleben, T., & König, G. (2023, July). Dear XAI community, we need to talk! Fundamental misconceptions in current XAI research. In World Conference on Explainable Artificial Intelligence (pp. 48-65). Cham: Springer Nature Switzerland.

“Querida comunidad XAI, ¡tenemos que hablar!”



Freiesleben, T., & König, G. (2023, July). Dear XAI community, we need to talk! Fundamental misconceptions in current XAI research. In World Conference on Explainable Artificial Intelligence (pp. 48-65). Cham: Springer Nature Switzerland.

“Querida comunidad XAI, ¡tenemos que hablar!”



Freiesleben, T., & König, G. (2023, July). Dear XAI community, we need to talk! Fundamental misconceptions in current XAI research. In World Conference on Explainable Artificial Intelligence (pp. 48-65). Cham: Springer Nature Switzerland.

“Querida comunidad XAI, ¡tenemos que hablar!”



Fig. 4. Misconception that the goal is to give people explanations they find intuitive

Freiesleben, T., & König, G. (2023, July). Dear XAI community, we need to talk! Fundamental misconceptions in current XAI research. In World Conference on Explainable Artificial Intelligence (pp. 48-65). Cham: Springer Nature Switzerland.

Introducción al Deep Learning

Día 2: Introducción a las Redes Neuronales

Manuel Germán y David de la Rosa
Universidad de Jaén



Universidad
de Jaén



`(mgerman, drrosa)@ujaen.es`