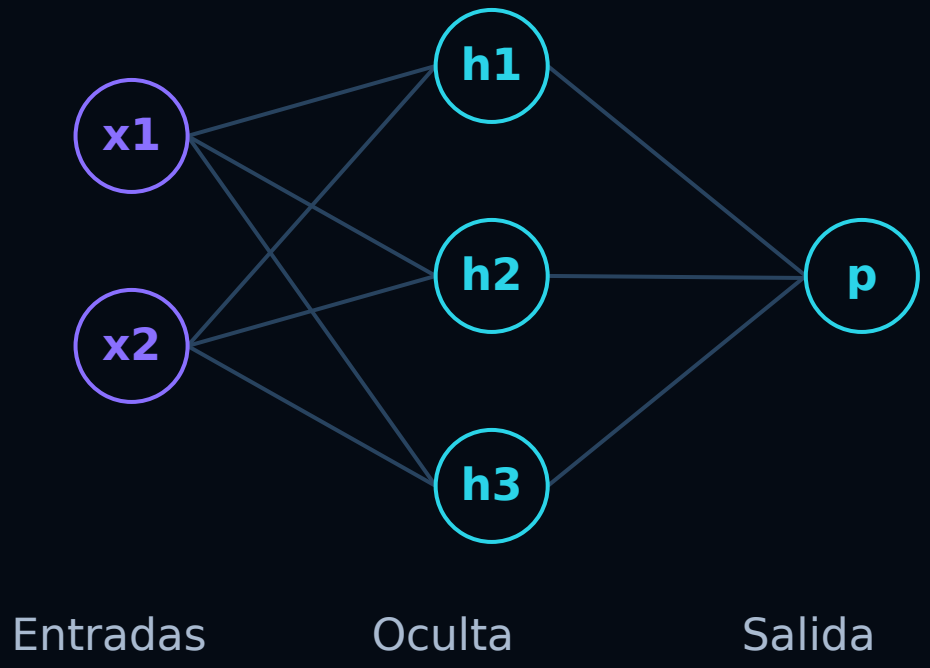
Redes neuronales

Fundamentos, aprendizaje y decisiones de negocio

Ciencias de Datos II · UCA

Sergio Gevatschnaider · 5 de noviembre de 2026



La decisión de retención

Unidad: cliente al cierre de un mes

Objetivo: abandono en los próximos 30 días

Acción: ofrecer una intervención con presupuesto limitado

Datos → modelo → política

Unidad: cliente al cierre de un mes

Objetivo: abandono en los próximos 30 días

¿Una cancelación ya confirmada puede ser una variable predictora?

Resultados de aprendizaje

Calcular una predicción y un gradiente

Explicar capas, activaciones y parámetros

Defender una política con evidencia de validación

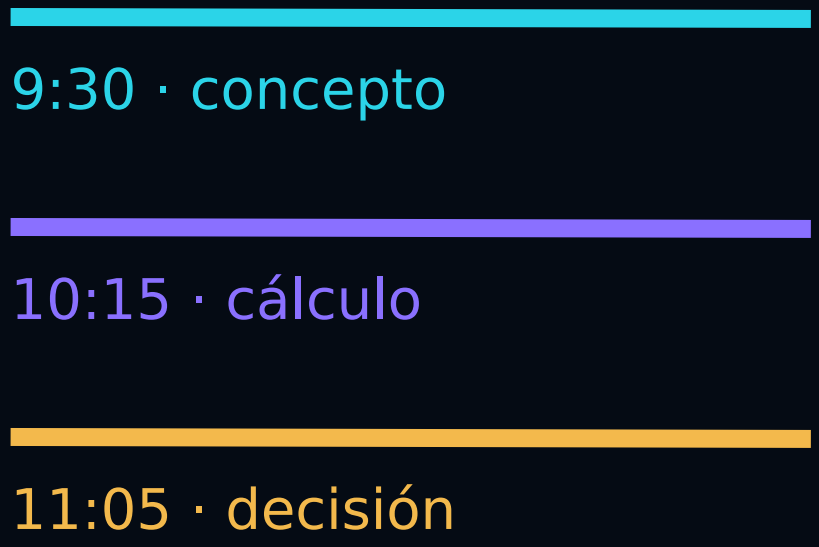
¿Qué evidencia distinguiría comprensión de repetición de fórmulas?

Ruta de clase: 9:30 a 12:00

9:30–10:15 · decisión, neurona, XOR y activaciones

10:15–11:05 · forward, pérdidas y backpropagation

11:05–12:00 · generalización, churn y defensa



¿Qué pregunta conecta los tres bloques?

Cuándo conviene una solución simple

Pocos ejemplos o etiquetas poco confiables

Exigencias estrictas de interpretación o latencia

Baselines tabulares que ya cumplen el objetivo

Comparación con evidencia

Pocos ejemplos o etiquetas poco confiables

Exigencias estrictas de interpretación o latencia

¿Más datos garantizan que una red supere a un árbol?

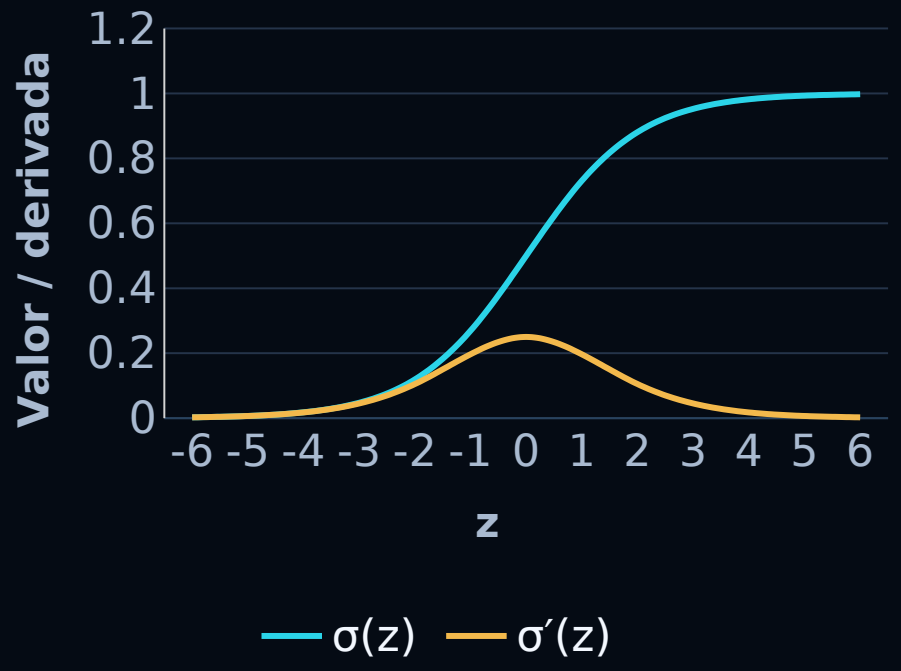
Regresión logística como baseline

$$z = w^T x + b$$
$$p = \sigma(z) = 1 / (1 + \exp(-z))$$

Un score afín se transforma en una probabilidad modelada

La frontera con umbral 0,5 es un hiperplano

La calibración debe comprobarse fuera de entrenamiento



¿Qué agrega una capa oculta a esta función?

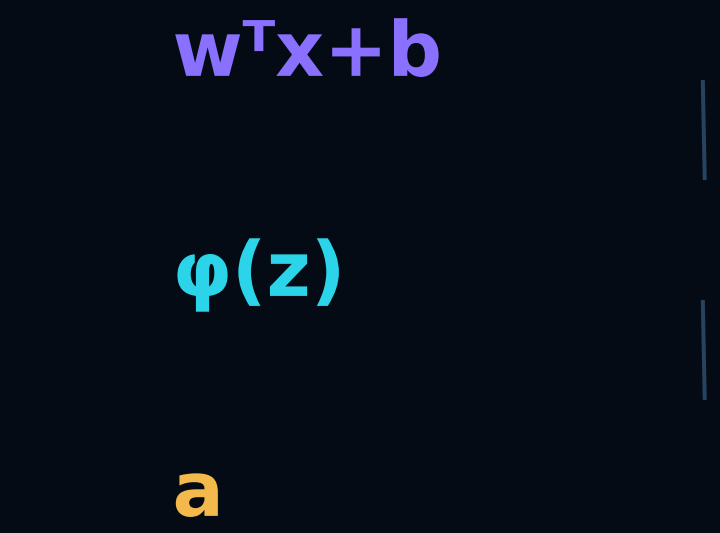
Entradas, pesos, sesgo y activación

$$z = w_1x_1 + w_2x_2 + b$$
$$a = \varphi(z)$$

x contiene atributos medibles

w y b son parámetros aprendidos

φ transforma la suma ponderada



¿Cambiar la unidad de x cambia la interpretación de w?

Inferencia: un cálculo verificable

$z = 0{,}8 - 0{,}2 - 0{,}1 = 0{,}5$
$p = \sigma(0{,}5) = 0{,}622459$

$x_1 = 0{,}8$ y $x_2 = 0{,}4$

$w_1 = 1$, $w_2 = -0{,}5$ y $b = -0{,}1$

Con umbral 0,5, la clase predicha es 1

0,622459

Valores didácticos verificables

¿Qué ocurre si b baja a −0,6?

Una frontera lineal

$$\hat{y} = 1[z \geq 0]$$
$$w \leftarrow w + \eta(y - \hat{y})x$$

El signo de $w^T x + b$ separa dos semiespacios

La regla del perceptrón corrige errores de clasificación

Convergencia: datos separables y condiciones de aprendizaje apropiadas



Frontera: $w^T x + b = 0$

¿Qué cambia el sesgo?

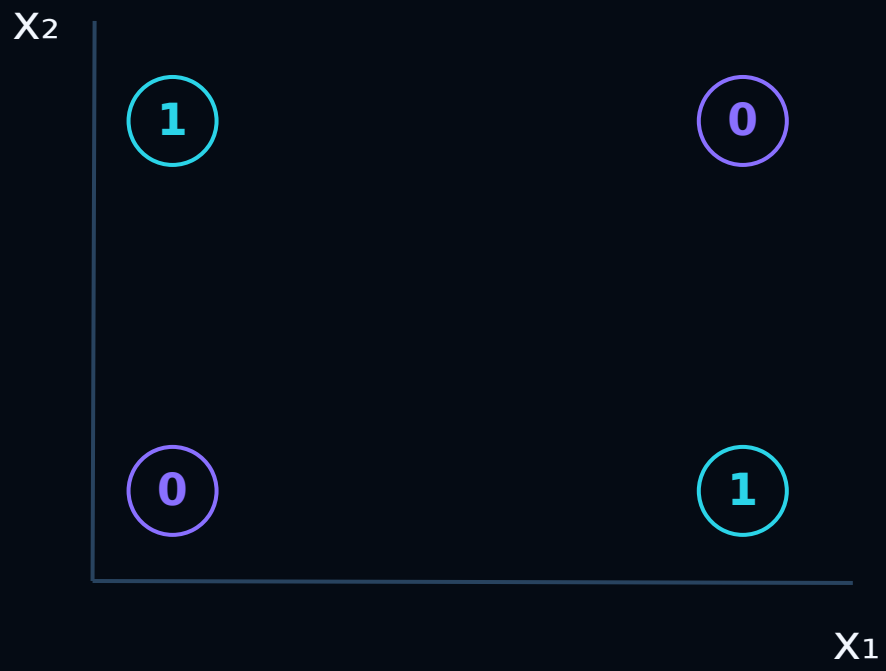
XOR exige una representación distinta

XOR(0,0)=0 XOR(0,1)=1
XOR(1,0)=1 XOR(1,1)=0

(0,0) y (1,1) pertenecen a la clase 0

(1,0) y (0,1) pertenecen a la clase 1

Ninguna recta separa las clases en las entradas originales



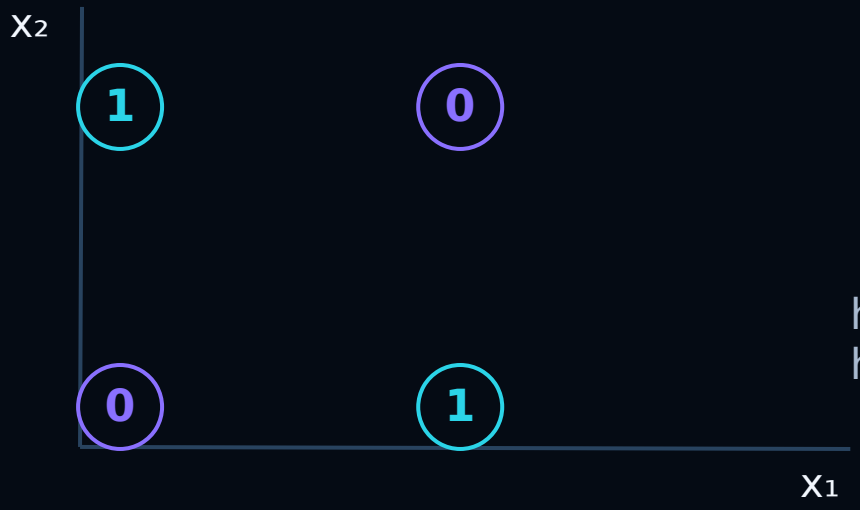
Clase 0: violeta / Clase 1: cian

¿Basta con entrenar más épocas un perceptrón?

XOR con dos unidades ReLU

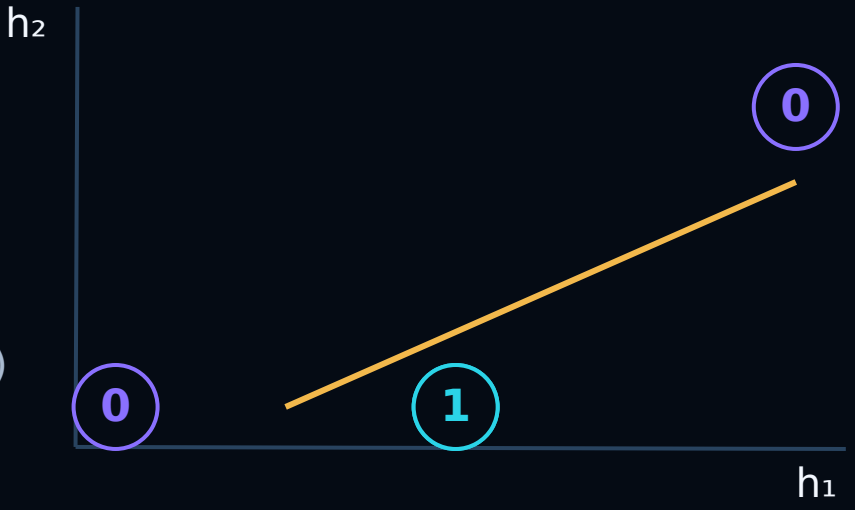
$h_1 = \mathrm{ReLU}(x_1+x_2)$

Entradas originales



$h_1 = \mathrm{ReLU}(x_1+x_2)$
$h_2 = \mathrm{ReLU}(x_1+x_2-1)$

Representación con dos ReLU



Clase 1: cian · clase 0: violeta · frontera: $h_1-2h_2 = 0{,}5$

¿Cuánto vale la salida para (1,1)?

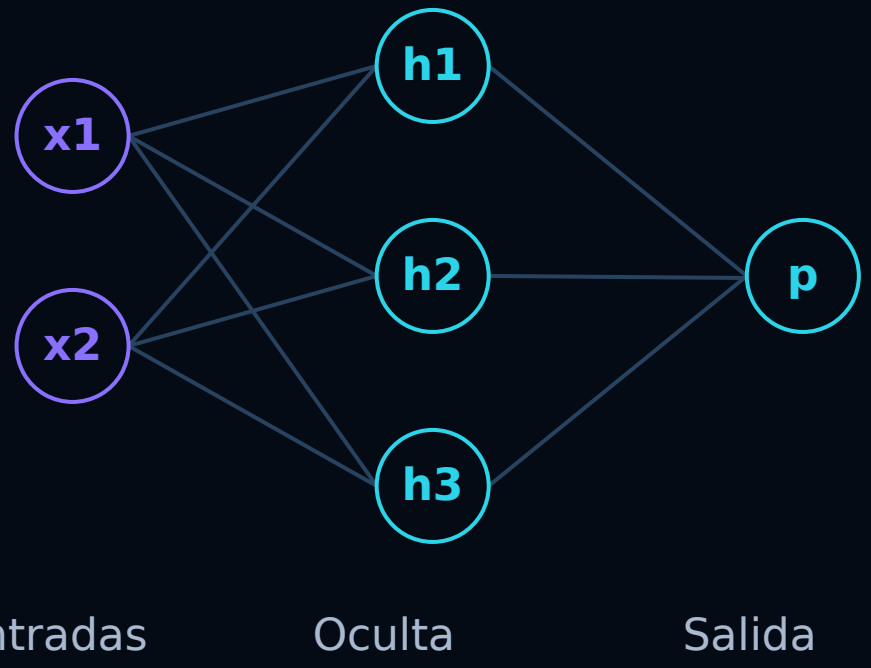
Dimensiones y parámetros de un MLP

Parámetros = 3·2 + 3 + 1·3 + 1 = 13

Dos entradas, tres unidades ocultas y una salida

W_1: 3×2, b_1: 3, W_2: 1×3, b_2: 1

Cada conexión y cada sesgo agrega un parámetro



¿Cuántos parámetros tiene una capa d→m con sesgo?

La no linealidad cambia la capacidad

$$W_2(W_1x+b_1)+b_2 = Ax+c$$

Dos capas afines sin activación colapsan en una

La profundidad por sí sola no introduce no linealidad

Las activaciones permiten representaciones por regiones

Comparación con evidencia

Dos capas afines sin activación colapsan en una

La profundidad por sí sola no introduce no linealidad

¿Qué ocurre si todas las capas ocultas usan identidad?

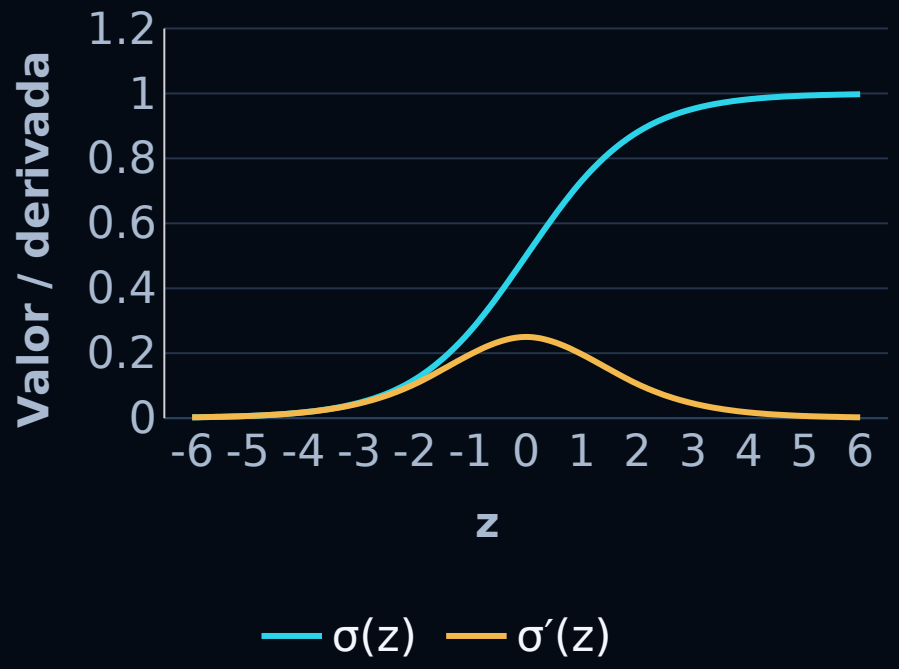
Sigmoide: salida acotada y saturación

$\sigma(z) = 1/(1+e^{-z})$
$\sigma'(z) = \sigma(z)(1-\sigma(z))$

Rango: (0,1)

Derivada máxima: 0,25 en z=0

En los extremos la sensibilidad se aproxima a cero



¿Cuál es la derivada en z=0?

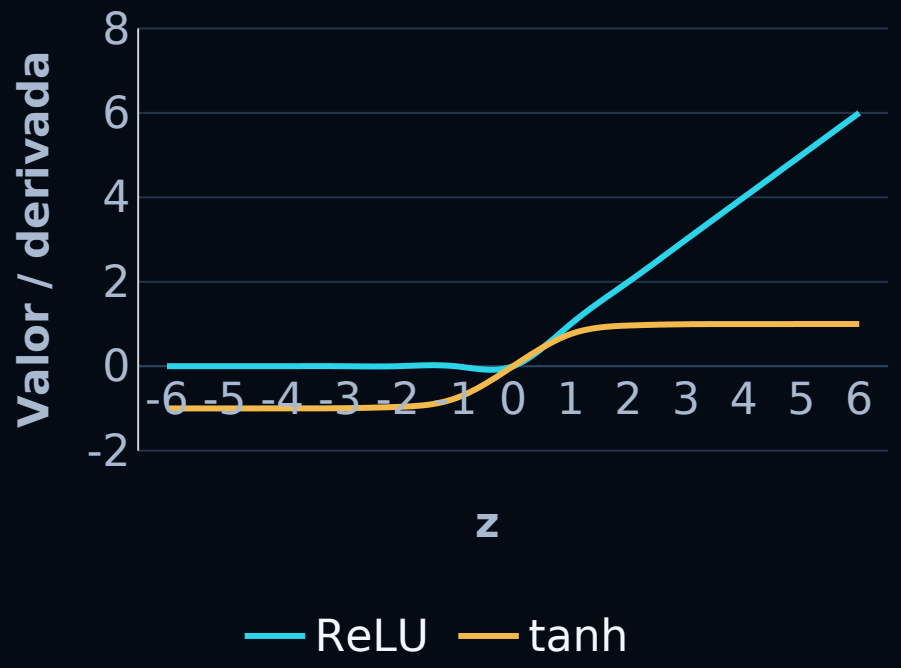
ReLU, tanh y Leaky ReLU

$\tanh'(z) = 1-\tanh^2(z)$
$\mathrm{ReLU}(z)=\max(0,z),\ \ \varphi'(z)=1[z>0]$
$\mathrm{Leaky}(z)=\max(\alpha z,z),\ 0<\alpha<1$

tanh se centra en cero, pero también satura

ReLU mantiene pendiente 1 en el lado positivo

Leaky ReLU conserva una pendiente α en el lado negativo



¿Una activación con salida grande tiene siempre derivada grande?

Salida según el objetivo

$$\mathrm{softmax}(z)_k = \exp(z_k - m) \,/\, \Sigma_j \exp(z_j - m)$$
$$m = \max_j z_j$$

Binaria: una sigmoide y BCE

Multiclase exclusiva: softmax y entropía cruzada

Regresión: identidad y una pérdida adecuada al ruido

¿Dos etiquetas que pueden coexistir requieren softmax?

Laboratorio de activaciones

Escenario 1: recorrer z entre −6 y 6

Escenario 2: observar saturación

Escenario 3: cruzar el punto no derivable de ReLU

Antes de simular

1. Predecir
2. Observar
3. Explicar

¿Qué se pierde al observar solo la curva de activación?

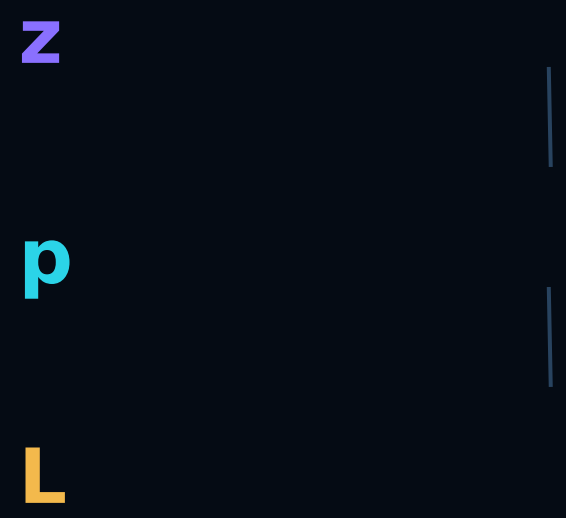
Forward propagation

$$h = \varphi(W_1 x + b_1)$$
$$z_2 = W_2 h + b_2$$
$$p = \sigma(z_2)$$

Las entradas pasan por las capas con pesos fijos

El entrenamiento agrega una pérdida y sus derivadas

La inferencia termina en la predicción



¿El forward cambia los pesos?

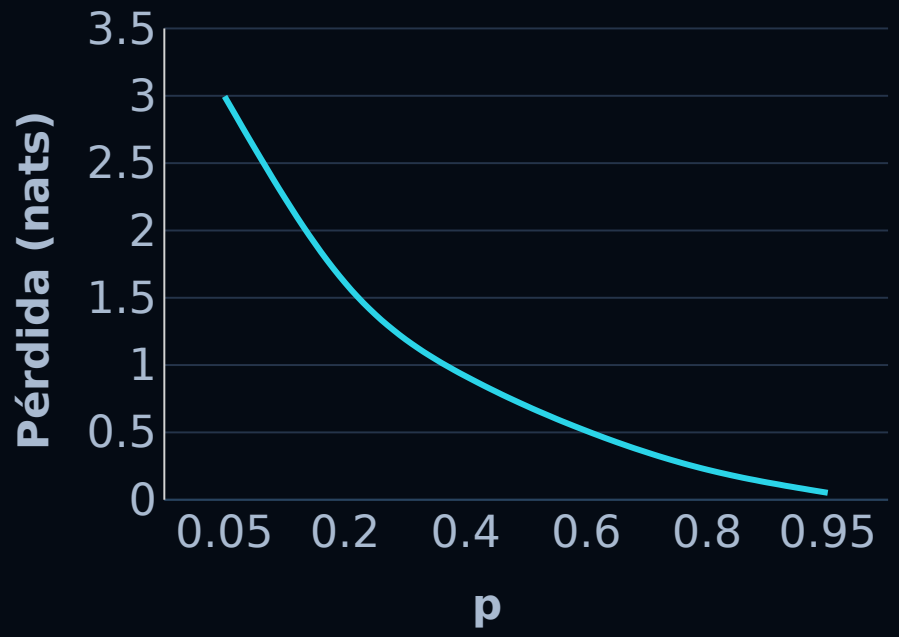
BCE para clasificación binaria

$L = -[y \cdot \ln(p) + (1-y) \cdot \ln(1-p)]$
$y=1$, $p=0{,}622459$: $L=0{,}474077$

Penaliza probabilidades incompatibles con y

Usa logaritmo natural y promedia sobre ejemplos

La implementación con logits mejora estabilidad numérica



¿Qué pérdida tiene un acierto con probabilidad 0,5?

MSE para regresión

MSE = (1/n) $\Sigma_i(\hat{y}_i - y_i)^2$
y=10, ŷ=8: error²=4

Penaliza más los errores grandes

Se expresa en unidades al cuadrado

RMSE devuelve la escala original del objetivo

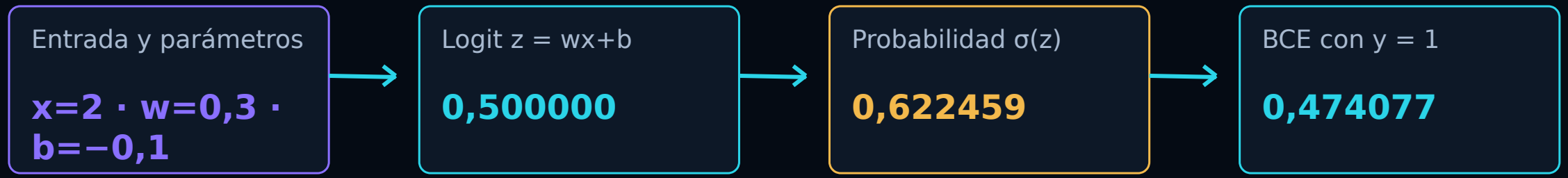
4 unidades²

Valores didácticos verificables

¿Cambiar la unidad monetaria altera MSE?

El grafo de una neurona logística

z=0,5 p=0,622459 L=0,474077

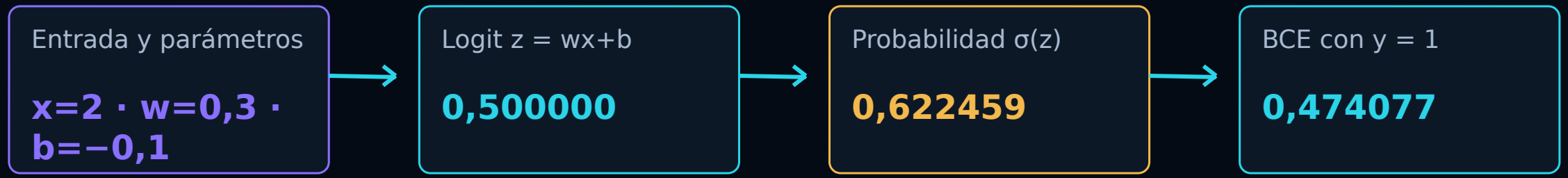


El forward conserva los valores intermedios para calcular derivadas después.

¿Qué valores deben guardarse para la derivación?

Regla de la cadena

$$\partial L/\partial w = (\partial L/\partial p)(\partial p/\partial z)(\partial z/\partial w)$$



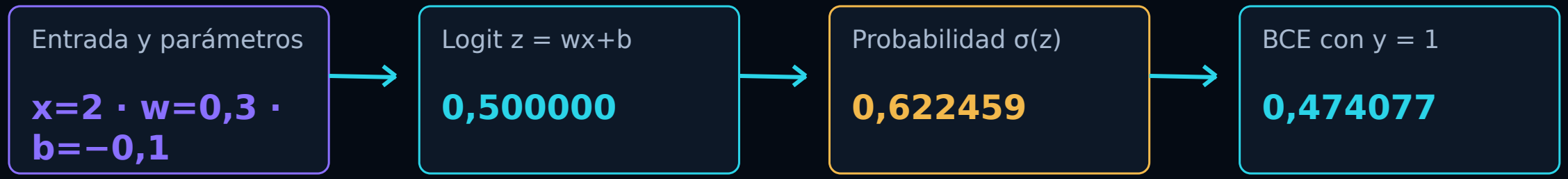
$\partial L/\partial p = -1/p$ $\partial p/\partial z = p(1-p)$ $\partial z/\partial w = x = 2$

Producto: $\partial L/\partial w = (p-y)x = -0{,}755081$

¿Por qué se suman gradientes cuando un parámetro tiene varios caminos?

Backpropagation: resultado numérico

$\partial L/\partial z = p - y$



La señal vuelve desde la pérdida

$\delta = p - y = -0{,}377541$

$\partial L/\partial w = -0{,}755081$ **$\partial L/\partial b = -0{,}377541$**

¿Por qué el gradiente de w duplica al de b?

Una actualización de descenso

$$w_{nuevo} = w - \eta \cdot \partial L / \partial w$$

Entrada y parámetros

x=2 · w=0,3 · b=−0,1

→

Logit z = wx+b

0,500000

→

Probabilidad σ(z)

0,622459

→

BCE con y = 1

0,474077

Actualización con η=0,1

w′ = 0,375508

Actualización del sesgo

b′ = −0,062246

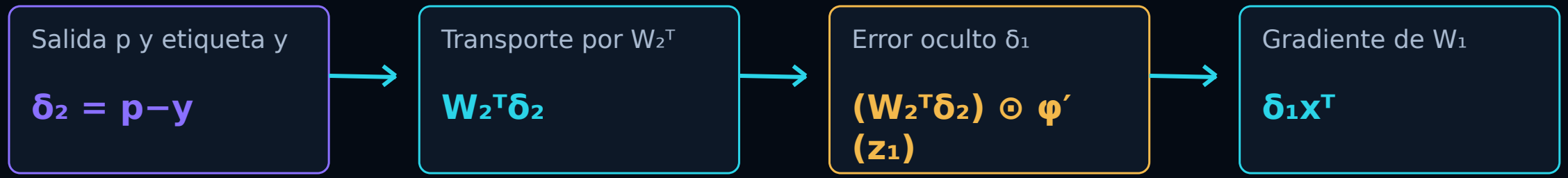
Nuevo forward

BCE′ = 0,406926

¿En qué dirección se movieron ambos parámetros?

Backpropagation en una capa oculta

$$\delta_2 = p - y$$



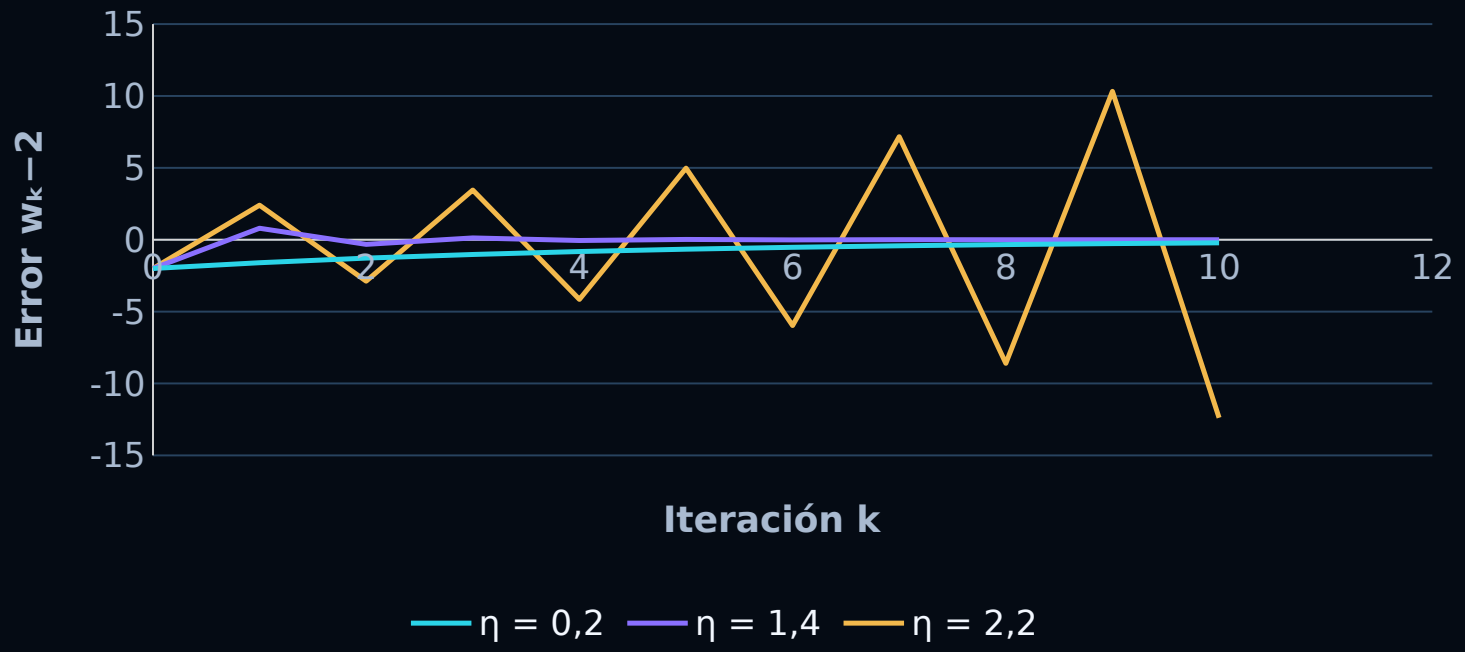
W_2 reparte la señal hacia la capa oculta. φ' modula cuánto atraviesa cada unidad.

Una ReLU inactiva tiene derivada 0 en $z<0$ y bloquea esa contribución local.

¿Qué dimensión tiene el gradiente de W_1?

Learning rate y estabilidad

$L(w)=\frac{1}{2}(w-2)^2$



$L(w) = \frac{1}{2}(w-2)^2$

$0 < \eta < 2$
converge

$\eta > 2$
oscila y se aleja

¿Qué sucede con η=1 en esta parábola?

Mini-batch, iteración y época

n=1.000, batch=100
10 actualizaciones por época

Mini-batch: ejemplos para estimar un gradiente

Iteración: una actualización

Época: recorrido del conjunto de entrenamiento

1 **2** **3** **4** **5**

Un lote por actualización

Todos los lotes: una época

¿Cinco épocas equivalen a cinco actualizaciones?

Inicialización y escalado

He para ReLU: Var(w) ≈ 2/fan_in

Romper simetrías entre unidades

Mantener una escala razonable de activaciones

Ajustar transformaciones solo con entrenamiento

Comparación con evidencia

Romper simetrías entre unidades

Mantener una escala razonable de activaciones

¿Por qué no estimar el escalado con todo el dataset?

Gradientes que se atenúan o explotan

$\partial L/\partial h_0$ contiene productos de $\partial h_k/\partial h_{k-1}$

Productos de Jacobianos pueden cambiar mucho de escala

Saturación limita la señal de aprendizaje

Inicialización, arquitectura y clipping cumplen papeles distintos



¿Clipping resuelve un gradiente casi nulo?

Laboratorio de optimización

BCE: aprender w y b con cuatro observaciones

Cuadrática estable: observar convergencia

Cuadrática inestable: explicar oscilación y divergencia

Antes de simular

1. Predecir
2. Observar
3. Explicar

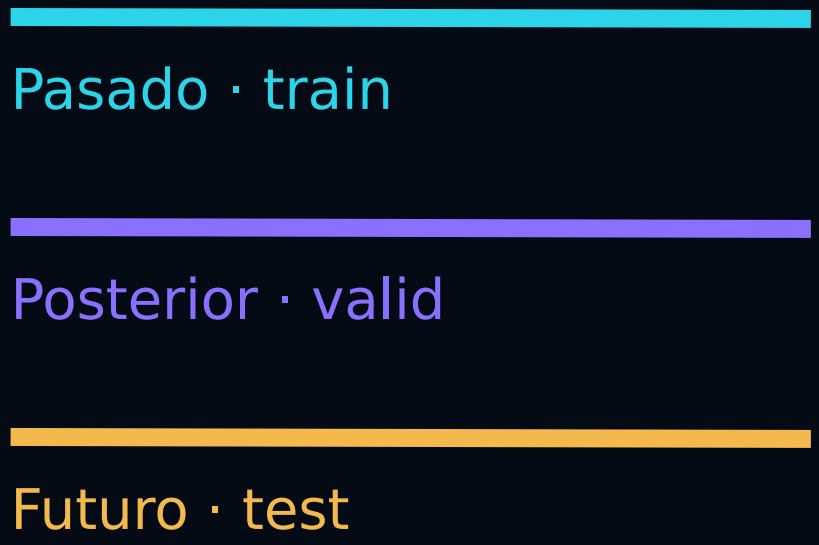
¿Cómo verificarías una derivada implementada?

Train, validación y test

Train ajusta parámetros y transformaciones

Validación selecciona arquitectura y umbral

Test estima desempeño final de la política elegida



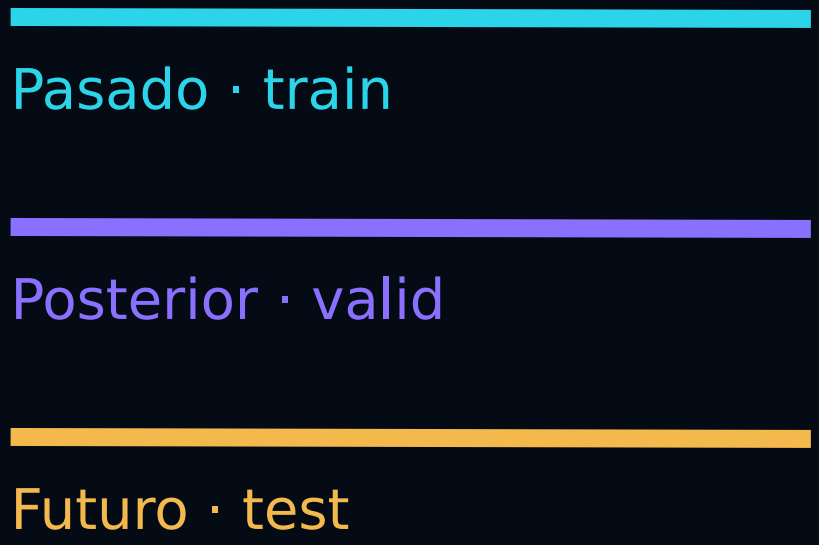
¿Dónde elegimos el umbral de retención?

Fuga de información en churn

Ejemplo: usar el motivo de baja para predecir la baja

Transformaciones y selección de variables dentro del split

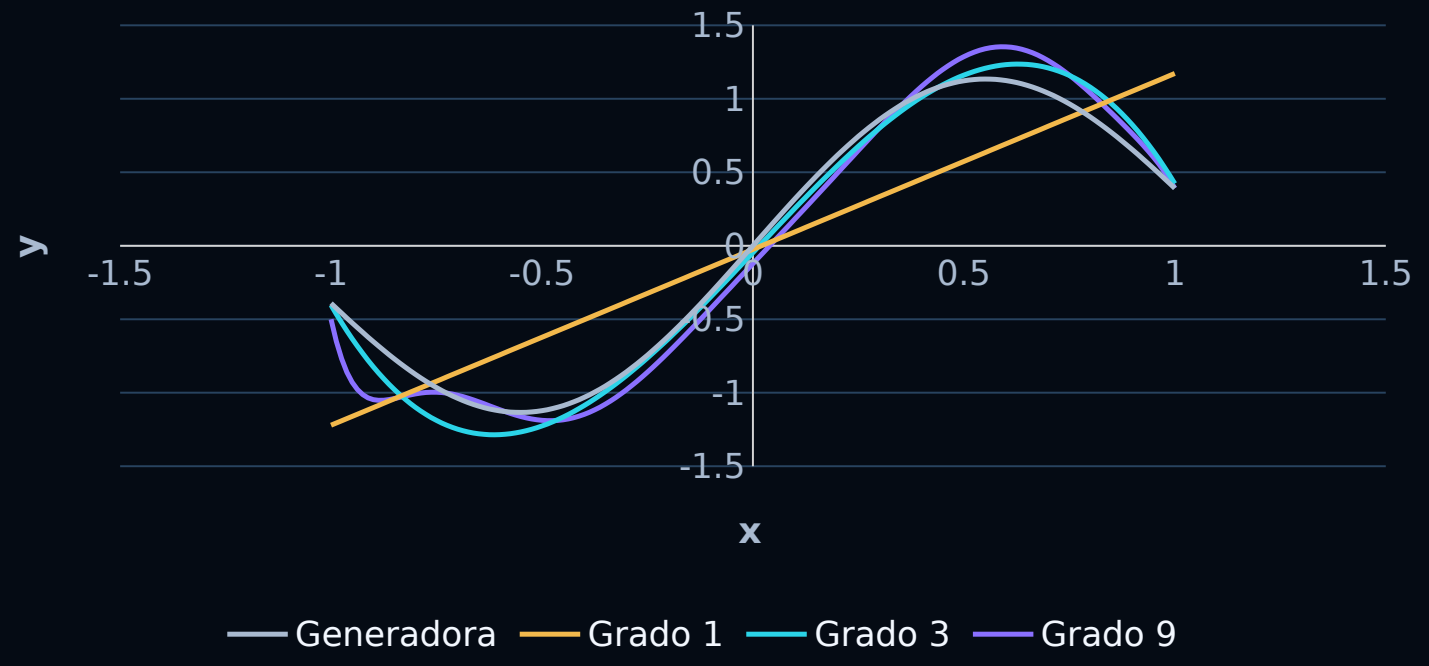
Repeticiones del mismo cliente requieren agrupamiento



¿Es suficiente borrar la columna objetivo?

Capacidad y generalización

$J = MSE_{train} + \lambda\|w\|^2$



MSE en validación

Grado 1	**0.1808**
Grado 3	**0.0338**
Grado 9	**0.0434**

12 puntos train · 60 valid
Semilla 42 · λ=0

¿Una pérdida train más baja implica un mejor modelo?

L2 y weight decay

$J = L_{datos} + (\lambda/2)\|W\|^2$



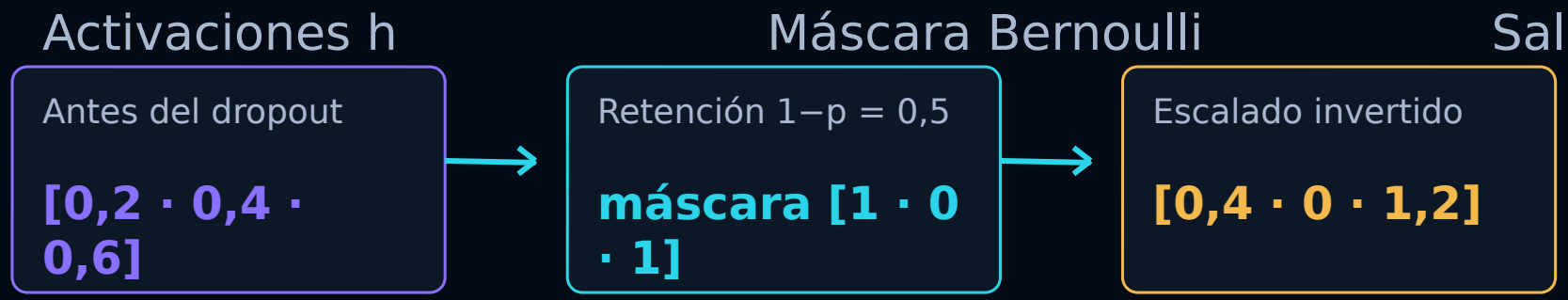
$$J = \tfrac{1}{2}(w-2)^2 + \tfrac{1}{2}\lambda w^2$$

$\lambda=0 \quad w^*=2$
$\lambda=1 \quad w^*=1$
$\lambda=3 \quad w^*=0{,}5$

¿Qué ocurre al aumentar demasiado λ?

Dropout

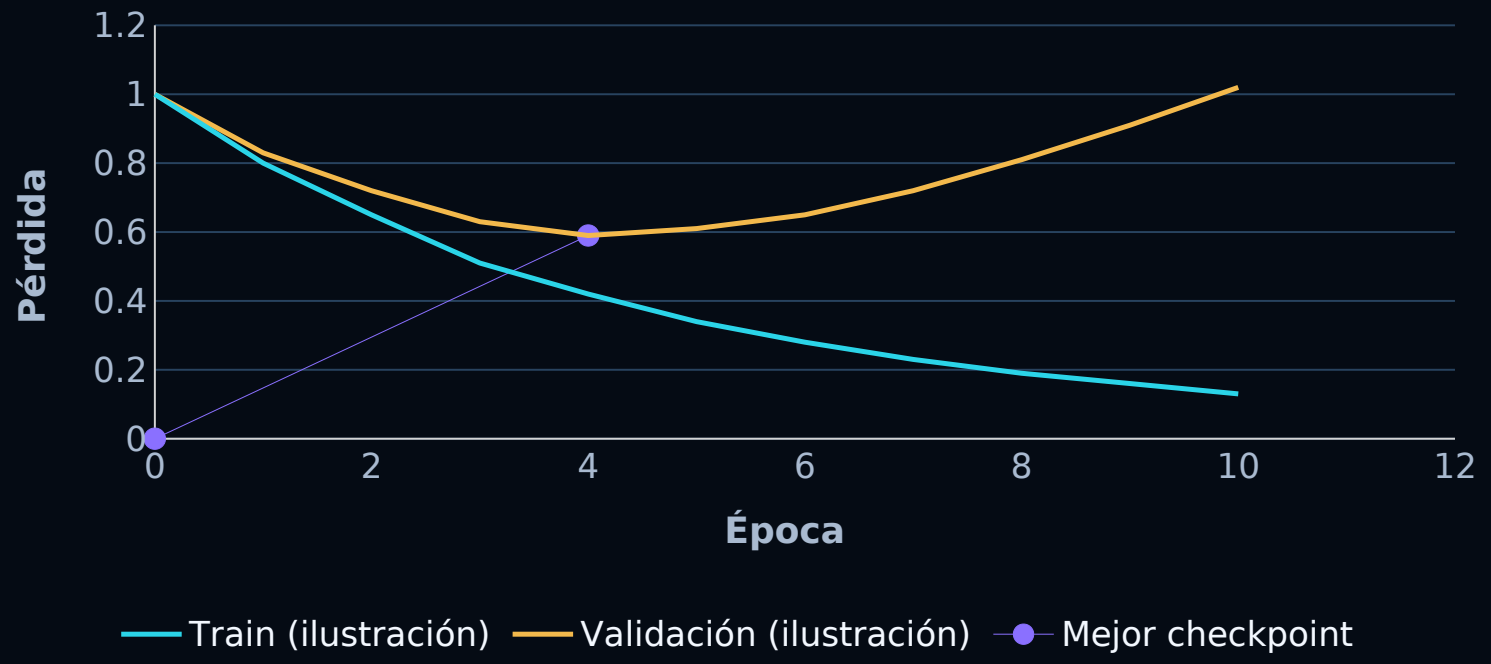
$$\tilde{h} = m \odot h / (1-p_{drop})$$



Cada activación conservada se divide por 1−p. En inferencia se usa la red completa.

¿Dropout=0,5 implica eliminar la mitad de los parámetros al exportar?

Early stopping y checkpoints



Se conserva la época 4

El mínimo de validación selecciona el checkpoint.

Test se consulta al final.

¿Qué pesos evaluamos al final?

Desbalance: 95% de accuracy sin detectar bajas

Accuracy = 95/100 = 95%
Recall = 0/5 = 0%

Ejemplo sintético: 100 clientes, 5 abandonos

Predecir siempre “no abandona” acierta 95

Recall de abandono: 0 de 5

95%

accuracy del clasificador trivial

0 / 5 abandonos detectados

¿Cuántos clientes de riesgo detectó la política?

Matriz de confusión y política

Precision = 3/10 = 0,30

Real / Acción	Contactar	No contactar
Abandona	TP = 3	FN = 2
No abandona	FP = 7	TN = 88

Recall = 3/5 = 60%

Precisión = 3/10 = 30%

100 clientes sintéticos
Ejemplo manual de conteo

¿Cuál es la especificidad?

Precision y recall dependen del umbral

Precision = TP/(TP+FP)
Recall = TP/(TP+FN)

Bajar el umbral suele aumentar contactos y recall

Precision puede subir o bajar según los scores

Capacidad y costo definen qué intercambio conviene

Datos → modelo → política

Bajar el umbral suele aumentar contactos y recall

Precision puede subir o bajar según los scores

¿Qué métrica cuenta cuántos abandonos quedaron fuera?

ROC-AUC y curva precision-recall

TPR = TP/(TP+FN)





Mismos 100 scores sintéticos, semilla 991. Ninguna curva incorpora por sí sola el costo de actuar.

¿Un ROC-AUC mayor garantiza más beneficio a capacidad del 15%?

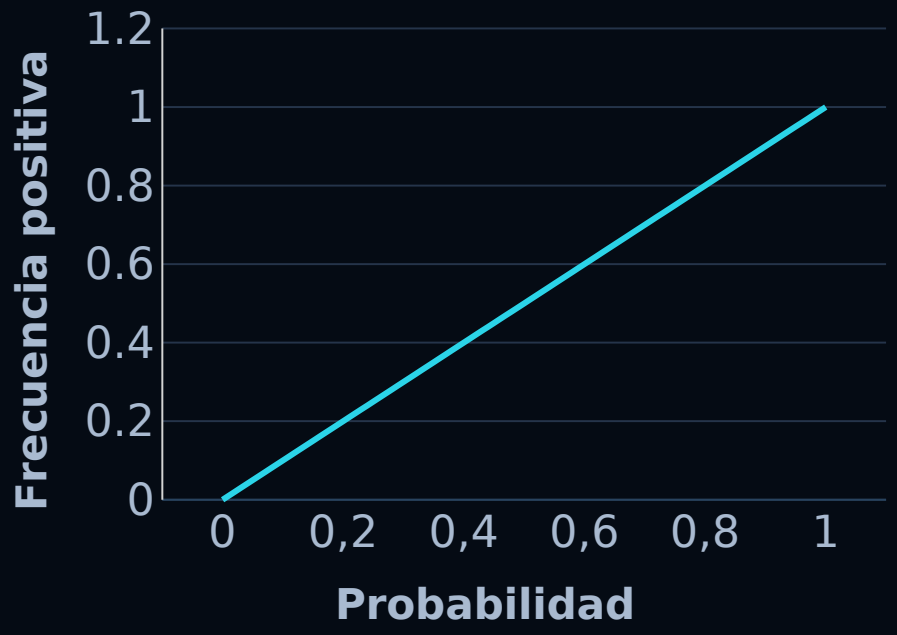
Calibración y ranking

P(Y=1 | p≈0,20) ≈ 0,20

Ranking ordena clientes por score

Calibración relaciona probabilidades con frecuencias

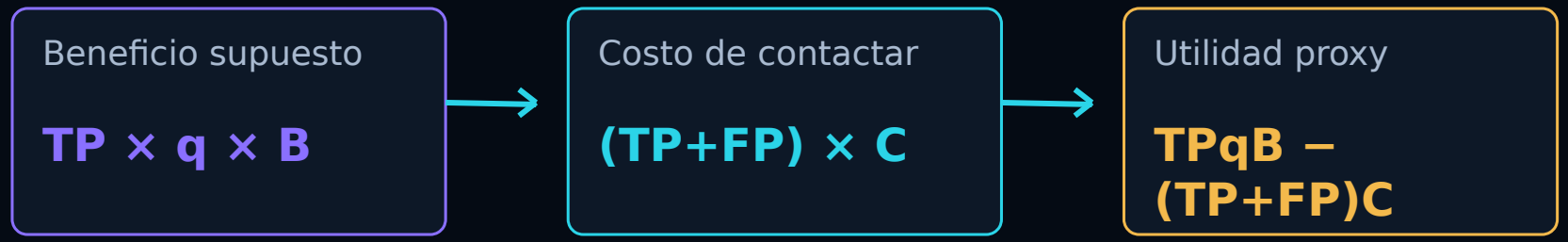
Un buen ordenamiento puede coexistir con mala calibración



¿AUC cambia si aplico una transformación estrictamente creciente al score?

Una utilidad esperada explícita

$U_{proxy} = TP \cdot q \cdot B - (TP+FP) \cdot C$



Ejemplo: TP=3, FP=7, q=0,4, B=100 UM, C=5 UM

Beneficio 120 UM − costo 50 UM = 70 UM

¿Podemos medir q solo con las etiquetas de abandono y los scores?

Umbral con costo y capacidad

Actuar si p·q·B > C
p > C/(q·B) = 0,125

Ejemplo manual: TP=3, FP=7

q=0,4, B=100 UM y C=5 UM por contacto

U_proxy=3·0,4·100−10·5=70 UM

70 UM

Valores didácticos verificables

¿Si el presupuesto solo permite dos contactos, basta el umbral 0,125?

Comparación justa: logística frente a MLP

Mismas filas, objetivo, splits y transformaciones

Selección con validación y presupuesto comparable

Test final con política congelada y costos documentados

Comparación con evidencia

Mismas filas, objetivo, splits y transformaciones

Selección con validación y presupuesto comparable

¿Qué comparación sería injusta?

Un score de riesgo no es un efecto causal

$E[Y(1)-Y(0) \mid X=x]$

Clientes de alto riesgo pueden no responder a una oferta

Medir retención incremental requiere contrafactual

Un experimento aleatorizado permite estimar el efecto

Datos → modelo → política

Clientes de alto riesgo pueden no responder a una oferta

Medir retención incremental requiere contrafactual

¿Podemos atribuir las bajas evitadas al modelo sin grupo de comparación?

Interpretación, equidad y trazabilidad

Inspeccionar errores por segmentos pertinentes

Documentar datos, versiones, semillas y política

Monitorear drift y resultados después de adoptar

Datos → modelo → política

Inspeccionar errores por segmentos pertinentes

Documentar datos, versiones, semillas y política

¿Qué información permitiría reproducir una decisión tomada hoy?

Mapa de arquitecturas

MLP: combinaciones de atributos

CNN: patrones locales en grillas

RNN/LSTM: estado temporal; Transformer: atención

MLP
Atributos

CNN
Estructura local

RNN / LSTM
Secuencia

Transformer
Atención

¿Qué criterio manda al elegir una arquitectura?

Actividad: defensa de una política

Grupos: elegir baseline y MLP para churn mensual

Justificar split, métrica y umbral

Presentar una limitación y una condición de rechazo

Defensa con evidencia

1 Protocolo
2 Resultado
3 Límites

¿Adoptarían una red con mayor AUC y menor utilidad de validación?

Comprobación de comprensión

1. ¿Cuántos parámetros tiene 2→3→1?

2. ¿Por qué una neurona no resuelve XOR?

3. ¿Puede una mejor BCE asegurar más retención?

Defensa con evidencia

1 Protocolo
2 Resultado
3 Límites

¿Qué concepto une los tres problemas?

Material para continuar

Lectura: Deep Learning, capítulos 6, 7 y 8

Videos: intuición, gradiente y backpropagation

Práctica: notebook y laboratorios reproducibles

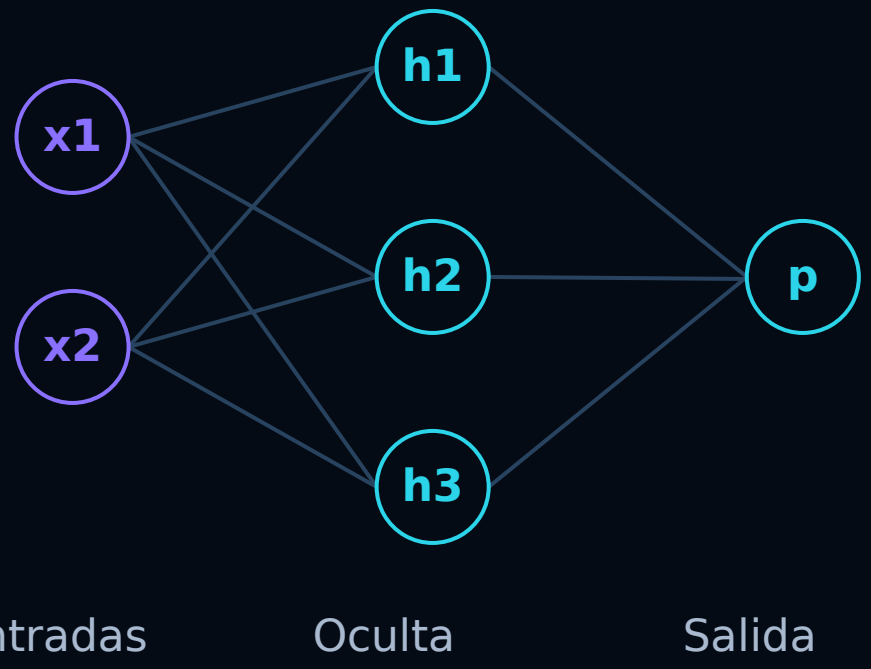
¿Qué experimento harías antes de usar datos reales?

Cierre: evidencia para decidir

Una red compone funciones y aprende parámetros

Backprop calcula gradientes; el optimizador actualiza

La adopción exige utilidad y generalización verificadas



¿Qué dato falta para pasar de riesgo a retención incremental?