Día 03 · Tema 1 — Ejercicios: regularización y *learning rate schedules*

Cómo se usa esta hoja. Hay dos tipos de ejercicio, y sólo dos.

El ejercicio del día. Uno por hoja. Es el obligatorio y es el que entra en el examen, con el cálculo incluido. Basta con hacer ese para seguir el curso con solvencia.

Los ejercicios de ampliación. El resto. Están para enseñar de qué es capaz el método, no para exigir: **no se piden calculados en el examen**. De ellos sólo puede preguntarse la idea, a nivel conceptual.

EL EJERCICIO DEL DÍA *obligatorio · entra en el examen*

1. Dropout con escalado invertido

Consideremos un vector de activaciones $\mathbf{a} = (3{,}0,\ -1{,}0,\ 2{,}5,\ 0{,}8)$ en una capa intermedia de una red neuronal durante entrenamiento. Se aplica *inverted dropout* con probabilidad de desactivación $p = 0{,}3$. La máscara binaria muestreada es $\mathbf{m} = (1, 0, 1, 1)$, donde $m_i \sim \text{Bernoulli}(1-p)$.

La fórmula del *inverted dropout* es

$$\tilde{a}_i \;=\; \frac{1}{1-p}\, m_i\, a_i\,.$$

a) Calcula el factor de escalado $\dfrac{1}{1-p}$.

b) Aplica la fórmula para obtener el vector de salida $\tilde{\mathbf{a}}$. Indica el valor numérico de cada componente (cuatro decimales).

c) Explica por qué el factor $\frac{1}{1-p}$ garantiza que $\mathbb{E}[\tilde{a}_i] = a_i$.

AMPLIACIÓN *opcional · no se pide calculado en el examen*

2. Batch Normalization: cálculo paso a paso

Sea un minibatch de cuatro activaciones (una por ejemplo) que llegan a una capa de *Batch Normalization*: $\mathbf{a} = (2{,}0,\ 4{,}0,\ 6{,}0,\ 8{,}0)$. Los parámetros aprendibles son $\gamma = 1{,}5$ y $\beta = 0{,}5$. Se usa $\varepsilon = 10^{-5}$.

a) Calcula la media μ_B y la varianza σ_B^2 del minibatch.

b) Normaliza cada elemento: $\hat{a}_i = \dfrac{a_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}}$.

c) Aplica la transformación afín: $y_i = \gamma\, \hat{a}_i + \beta$.

d) Si hiciéramos $\gamma = \sqrt{\sigma_B^2 + \varepsilon}$ y $\beta = \mu_B$, ¿qué obtendríamos como y_i?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

3. Weight decay: pérdida total y gradiente de regularización

Una red tiene solo tres pesos $\boldsymbol{\theta} = (0{,}5,\ -1{,}2,\ 0{,}8)$ y la pérdida de datos actual es $L_{\text{data}} = 2{,}3$. Se aplica *weight decay* con $\lambda = 0{,}01$, es decir,

$$L_{\text{total}} = L_{\text{data}} + \lambda\, \|\boldsymbol{\theta}\|^2\,.$$

a) Calcula $\|\boldsymbol{\theta}\|^2$ y después L_{total}.

b) Calcula el gradiente del término de regularización respecto a cada peso: $\dfrac{\partial(\lambda\|\boldsymbol{\theta}\|^2)}{\partial \theta_i}$.

c) Supongamos que el gradiente de L_{data} respecto a θ_1 es $-0{,}15$. ¿Cuál es el gradiente total $\dfrac{\partial L_{\text{total}}}{\partial \theta_1}$? Interpreta el signo y la magnitud del término de regularización.

AMPLIACIÓN *opcional · no se pide calculado en el examen*

4. Cosine annealing

Usamos *cosine annealing* con $\eta_{\text{máx}} = 0{,}1$, $\eta_{\text{mín}} = 0{,}001$ y periodo total $T = 100$ iteraciones:

$$\eta_t = \eta_{\text{mín}} + \tfrac{1}{2}\left(\eta_{\text{máx}} - \eta_{\text{mín}}\right)\left(1 + \cos\!\left(\tfrac{\pi\, t}{T}\right)\right).$$

a) Calcula η_t en $t = 0,\ 25,\ 50,\ 75,\ 100$. Usa que $\cos 0 = 1$, $\cos(\pi/4) \approx 0{,}7071$, $\cos(\pi/2) = 0$, $\cos(3\pi/4) \approx -0{,}7071$, $\cos\pi = -1$.

b) ¿En qué punto del ciclo la tasa de aprendizaje alcanza exactamente la media $(\eta_{\text{máx}} + \eta_{\text{mín}})/2$? Justifica analíticamente.

AMPLIACIÓN *opcional · no se pide calculado en el examen*

5. Step decay

El *step decay schedule* reduce la tasa de aprendizaje multiplicándola por un factor γ cada T épocas:

$$\eta_t = \eta_0 \, \gamma^{\lfloor t/T \rfloor} .$$

Con $\eta_0 = 0{,}1$, $\gamma = 0{,}1$ y $T = 10$:

a) Calcula η_t en las épocas $t = 0,\ 10,\ 20,\ 30,\ 40$.
b) ¿A partir de qué época la tasa de aprendizaje es inferior a 10^{-4}?
c) Compara este *schedule* con *cosine annealing*: ¿cuál produce transiciones más suaves? ¿Cuál es más fácil de ajustar en la práctica?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

6. Warmup lineal + cosine decay

Un *schedule* habitual en transformers combina una fase de *warmup* lineal con un *cosine decay*:

$$\eta_t = \begin{cases} \eta_{\text{máx}} \dfrac{t}{T_w}, & t < T_w, \\ \eta_{\text{mín}} + \dfrac{1}{2}\left(\eta_{\text{máx}} - \eta_{\text{mín}}\right)\left(1 + \cos\left(\dfrac{\pi\,(t - T_w)}{T - T_w}\right)\right), & t \geq T_w, \end{cases}$$

con $T_w = 1000$, $T = 10\,000$, $\eta_{\text{máx}} = 10^{-3}$ y $\eta_{\text{mín}} = 10^{-6}$.

a) Calcula η_t en $t = 500,\ 1000,\ 5500,\ 10\,000$.
b) Verifica que el *schedule* es continuo en $t = T_w$.
c) ¿Por qué se usa *warmup* al inicio del entrenamiento de transformers?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

7. Early stopping con paciencia

Durante el entrenamiento se registran las siguientes pérdidas de validación a lo largo de diez épocas:

Época	1	2	3	4	5	6	7	8	9	10
L_{val}	1,50	1,35	1,28	1,30	1,25	1,27	1,29	1,31	1,26	1,28

Se aplica *early stopping* con paciencia $P = 3$: el entrenamiento se detiene si la pérdida de validación no mejora el mejor valor registrado durante P épocas consecutivas.

a) Simula el mecanismo época a época: indica en cada una si hay mejora, el valor del contador de paciencia y el mejor L_{val} registrado hasta ese momento.
b) ¿En qué época se detiene el entrenamiento? ¿Cuál es el modelo que se conserva?
c) Si la paciencia fuera $P = 5$, ¿se detendría el entrenamiento dentro de estas 10 épocas?

AMPLIACIÓN *opcional · no se pide calculado en el examen*

8. Gradient clipping por norma

Durante una iteración de entrenamiento, el gradiente calculado es $\mathbf{g} = (3{,}0,\ -4{,}0)$. Se aplica *gradient clipping by norm* con umbral $c = 2{,}5$: si $\|\mathbf{g}\| > c$, se reescala el gradiente como

$$\mathbf{g}_{\text{clip}} = \mathbf{g}\,\frac{c}{\|\mathbf{g}\|} .$$

a) Calcula $\|\mathbf{g}\|$ y determina si es necesario recortar.
b) Calcula $\mathbf{g}_{\text{clip}}$.
c) Verifica que $\|\mathbf{g}_{\text{clip}}\| = c$.
d) ¿Se conserva la dirección del gradiente original? Justifica.