Lecture 04a
Artificial Intelligence

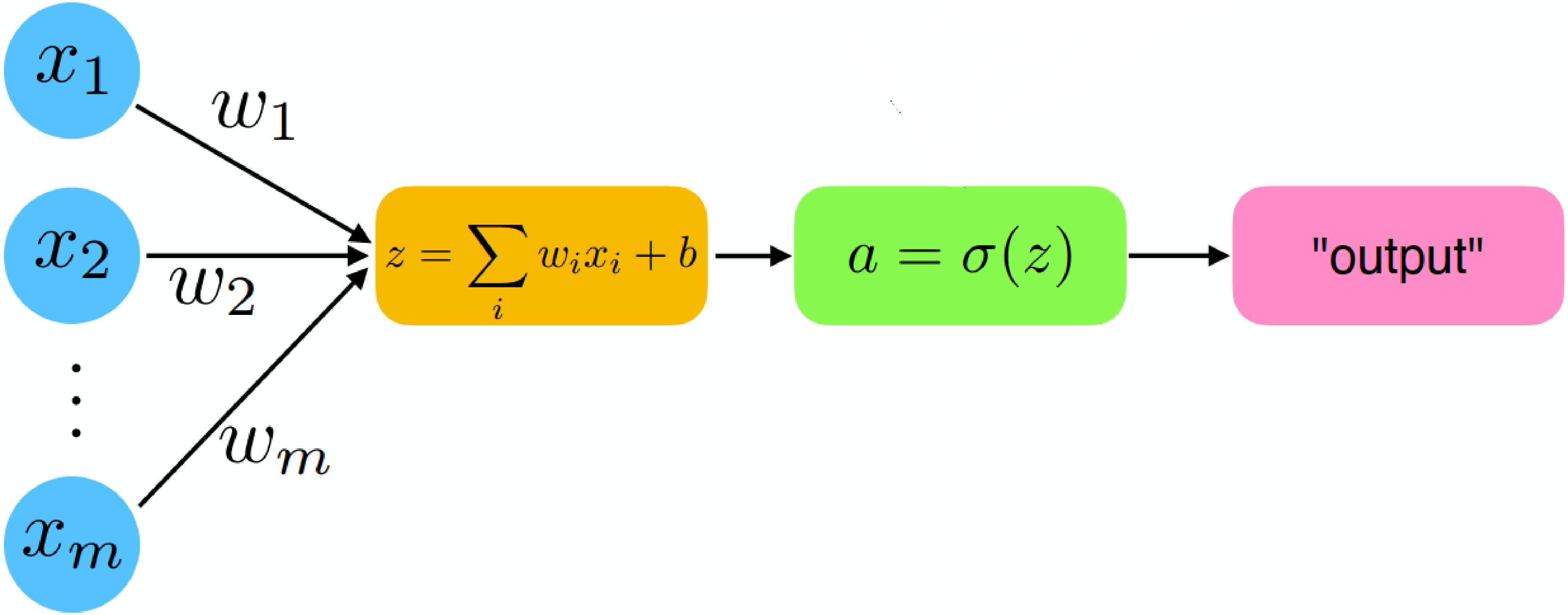
Recapitulando

- Mejoramos los problemas de convergencia del perceptrón vía Adaline
- Conceptualizamos gradiente descendente a través de los grafos computacionales

Agenda

- Regresión logística como una red neuronal
- *Negative Log-Likelihood Loss*
- Regla de aprendizaje de regresión logística
- *Logits* y *Cross-Entropy*
- Ejemplo: Regresión logística
- Generalización a múltiples clases: Regresión *softmax*
- *OneHot Encoding* y *Cross-Entropy* para varias clases
- Regla de aprendizaje de regresión *softmax*
- Ejemplo: Regresión *softmax*

Regresión logística para problemas bi-clase $y^{[i]} \in \{0, 1\}$



Para el sigmoide se tiene

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

Regresión logística para problemas bi-clase $y^{[i]} \in \{0, 1\}$

- En ADALINE, la función de activación era una función identidad

$$\sigma(z) = z$$

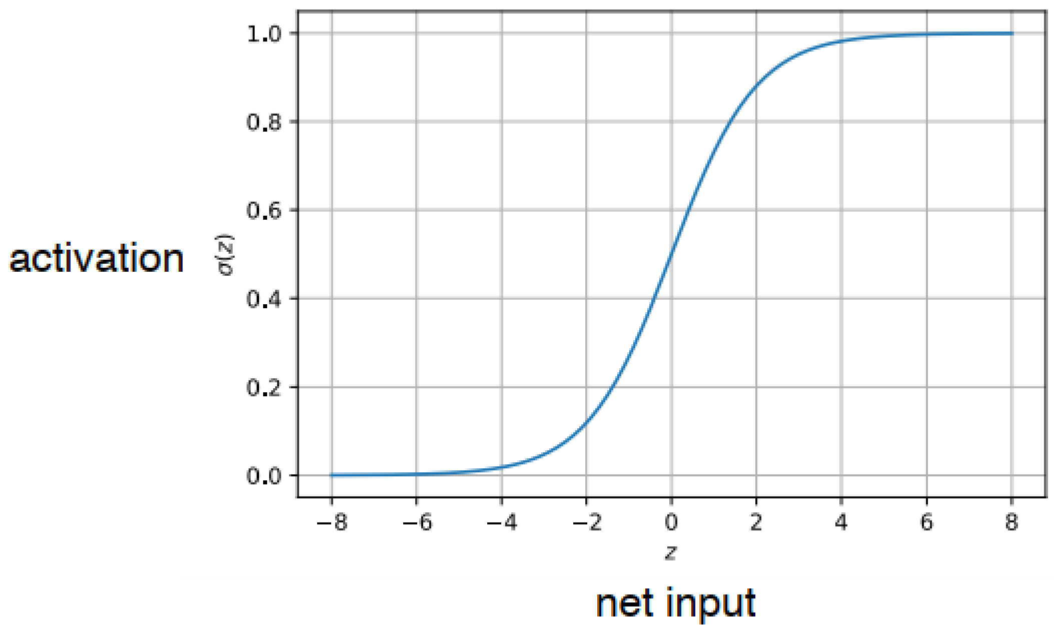
- Se utilizaba MSE como costo

$$MSE = \frac{1}{n}\sum_{i}(a^{[i]} - y^{[i]})^2$$

- Para regresión logística, se utilizará una función de costo diferente

Sigmoide

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



Regresión logística

Dada la salida

$$h(\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b)$$

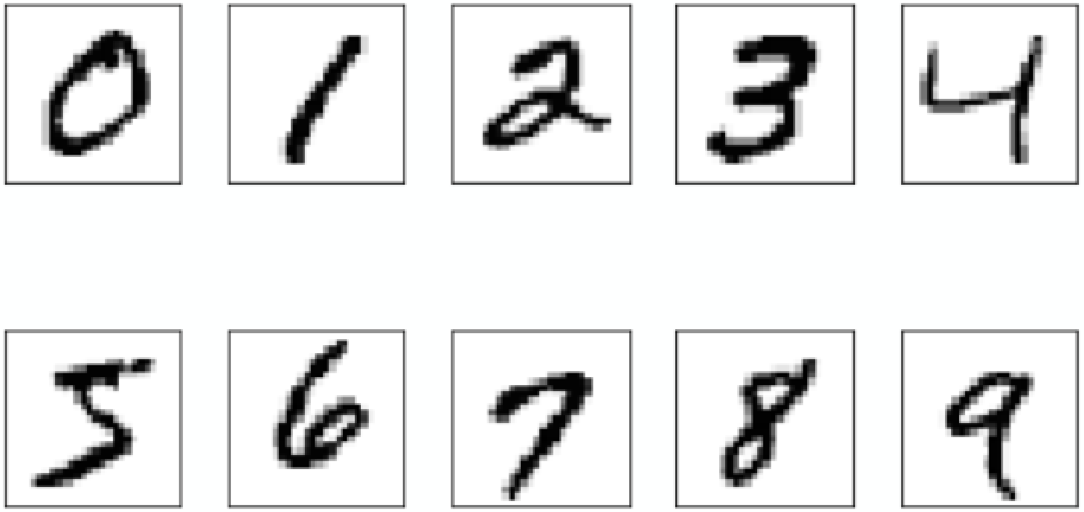
Donde $h(\mathbf{x})$ o hipótesis es la salida de la función de activación

$$h(\mathbf{x}) = a$$

Podemos calcular la probabilidad posterior o *a posteriori* como

$$P(y|\mathbf{x}) = \begin{cases} h(\mathbf{x}) & y = 1 \\ 1 - h(\mathbf{x}) & y = 0 \end{cases}$$

Generalización a múltiples clases: Regresión *softmax*



Generalización a múltiples clases: Regresión *softmax*

Base de datos balanceada:

- 10 clases
- 60.000 dígitos por clase
- Dimensión de las imágenes: $1 \times 28 \times 28$ (CHW)

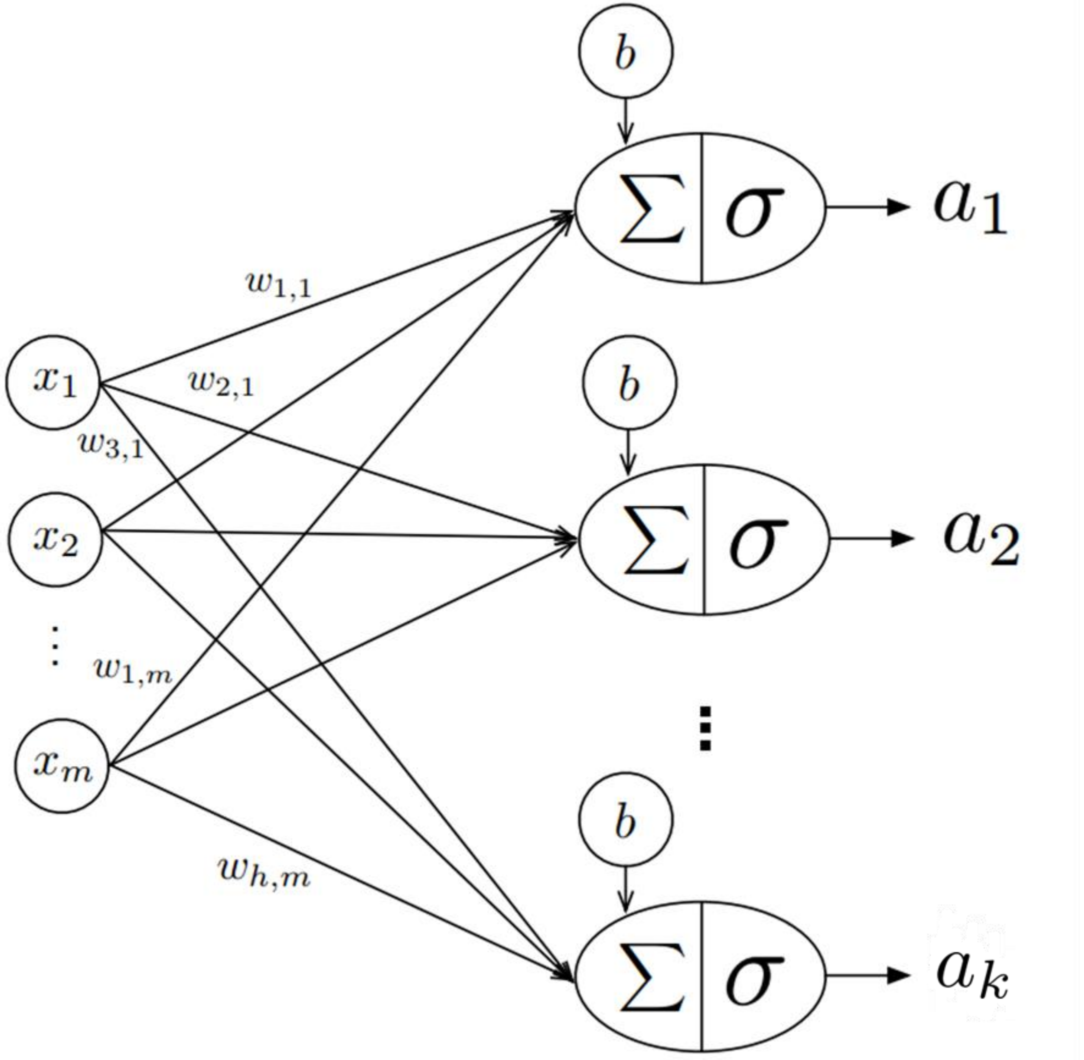
La forma tradicional de abordar este problema de clasificación es convertir cada imagen en un vector de tamaño 784×1.

Si se quiere hacer entrenamiento en *batches*, cada *batch* tendría dimensión $N_b \times 784$

Generalización a múltiples clases: Regresión *softmax*

En regresión *softmax*, la capa de salida de la red tiene varios nodos, uno por clase.

Las activaciones se pueden ver como proba-bilidad de perte-nencia a cada clase (no mutuamente ex-cluyente)



Generalización a múltiples clases: Regresión *softmax*

Matemáticamente se tiene que:

$\mathbf{a} = \sigma(\mathbf{W}\mathbf{x} + \mathbf{b}) \in \Re^{K \times 1}$, con

$\mathbf{W} \in \Re^{m \times K}$

$\mathbf{b} \in \Re^{K \times 1}$

$\mathbf{x} \in \Re^{m \times 1}$

Cada a_k será:

$a_k = \sigma(\mathbf{w}_k \cdot \mathbf{x} + b_k)$, con

$\mathbf{w}_k \in \Re^{1 \times m}$ k-th fila de $\mathbf{W}$

En caso de batches $\mathbf{X} \in \Re^{N_b \times m}$

$\mathbf{A} = \sigma(\mathbf{X}\mathbf{W}^\top + \mathbf{b}) \in \Re^{N_b \times K}$

$b \in \Re^{K,}$ (Tensor 1D)

Generalización a múltiples clases: Regresión *softmax*

Sin embargo, para que se cumpla que a_k es una probabilidad de pertenencia a cada clase:

$\sum_{k=1}^{K} a_k = 1$

Para esto, se utiliza *softmax*

$p(y = k|z_k^{[i]}) = \sigma_{\text{softmax}}(z_k^{[i]}) = \frac{e^{z_k^{[i]}}}{\sum_{k=1}^{K} e^{z_k^{[i]}}}$

K : Número de clases

$k = 1, \ldots, K$

Softmax es solo una función exponencial que normaliza las activaciones para que la suma de 1

One-hot encoding

class labels
0
1
3
2

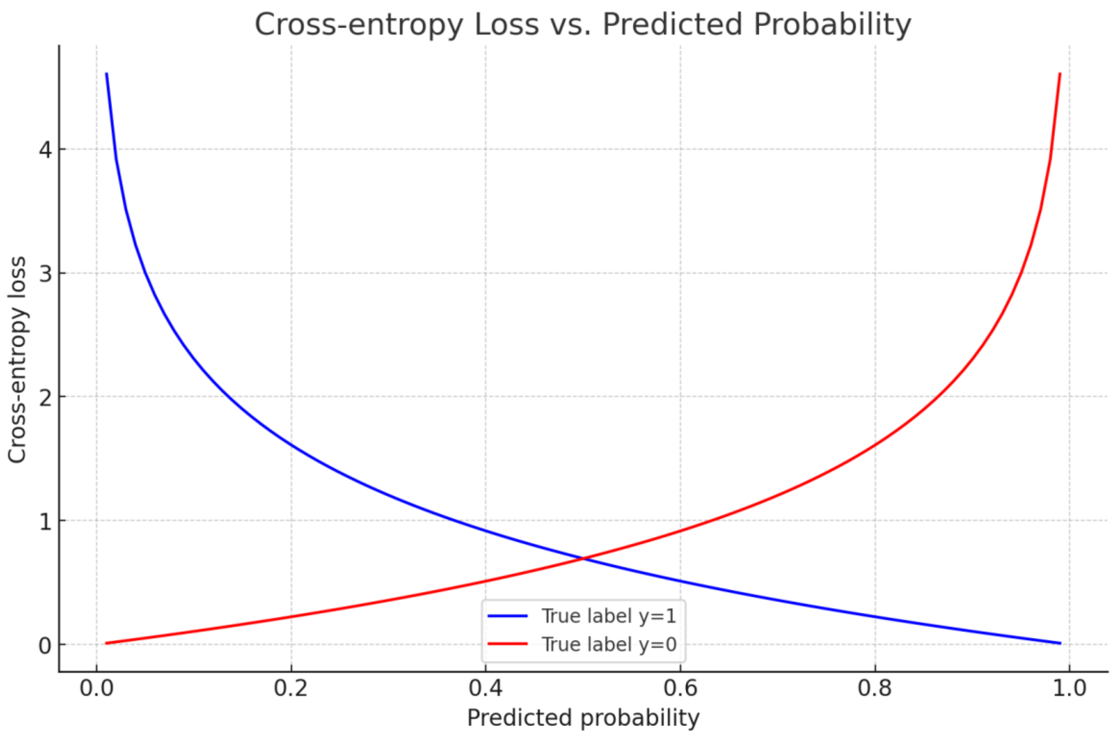
class_0	class_1	class_2	class_3
1	0	0	0
0	1	0	0
0	0	0	1
0	0	1	0

Función de costo

Cross-entropía multicategórica para h clases:

$$\mathcal{L} = -\sum_{i=1}^{n}\sum_{k=1}^{K} y_k^{[i]} \log(a_k^{[i]})$$

Asume etiquetas en *one-hot encoding*



Ejemplo

$$\mathbf{Y}_{\text{onehot}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \quad \mathbf{A}_{\text{softmax outputs}} = \begin{bmatrix} 0.3792 & 0.3104 & 0.3104 \\ 0.3072 & 0.4147 & 0.2780 \\ 0.4263 & 0.2248 & 0.3490 \\ 0.2668 & 0.2978 & 0.4354 \end{bmatrix}$$

$$\begin{aligned} \mathcal{L}^{[1]} &= [(-1)\cdot\log(0.3792)] \\ &+ [(-0)\cdot\log(0.3104)] \\ &+ [(-0)\cdot\log(0.3104)] \\ &= 0.969692... \end{aligned}$$

$$\begin{aligned} \mathcal{L}^{[2]} &= [(-0)\cdot\log(0.3072)] \\ &+ [(-1)\cdot\log(0.4147)] \\ &+ [(-0)\cdot\log(0.2780)] \\ &= 0.880200... \end{aligned}$$

$$\begin{aligned} \mathcal{L}^{[3]} &= [(-0)\cdot\log(0.4263)] \\ &+ [(-0)\cdot\log(0.2248)] \\ &+ [(-1)\cdot\log(0.3490)] \\ &= 1.05268... \end{aligned}$$

$$\begin{aligned} \mathcal{L}^{[4]} &= [(-0)\cdot\log(0.2668)] \\ &+ [(-0)\cdot\log(0.2978)] \\ &+ [(-1)\cdot\log(0.4354)] \\ &= 0.831490... \end{aligned}$$

$$\rightarrow \mathcal{L} = \boxed{\frac{1}{n}\sum_{i=1}^{n}}\sum_{j=1}^{h} -y_j^{[i]} \log\left(a_j^{[i]}\right)$$

$$\approx 0.9335$$

$n = 4$

$h = 3$

Derivadas de regresión *softmax* por gradiente descendente

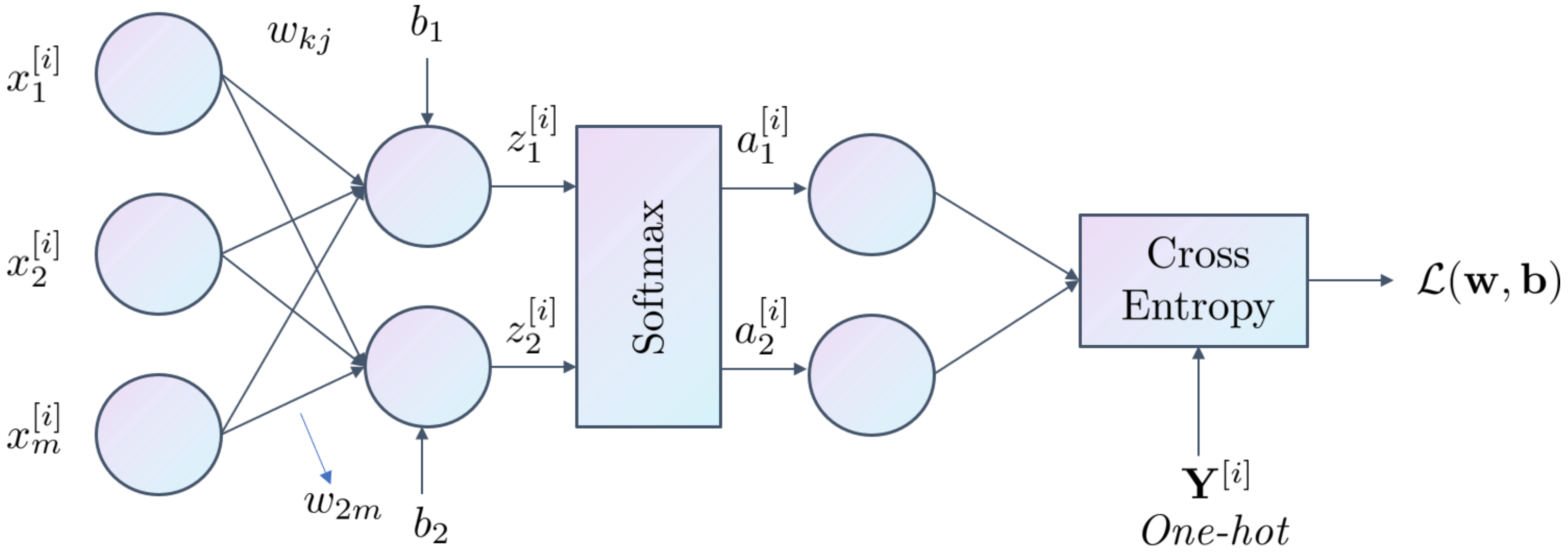
Para la regla de aprendizaje, necesitamos calcular:

$$\frac{\partial \mathcal{L}}{\partial w_{k,j}} = \frac{\partial \mathcal{L}}{\partial \mathbf{a}} \frac{\partial \mathbf{a}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial w_{k,j}}$$

$$\frac{\partial \mathcal{L}}{\partial b_k} = \frac{\partial \mathcal{L}}{\partial \mathbf{a}} \frac{\partial \mathbf{a}}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial b_k}$$

$\mathbf{x}^{[i]} \in \Re^{m\times 1}$

$\mathbf{X} \in \Re^{b\times m}$

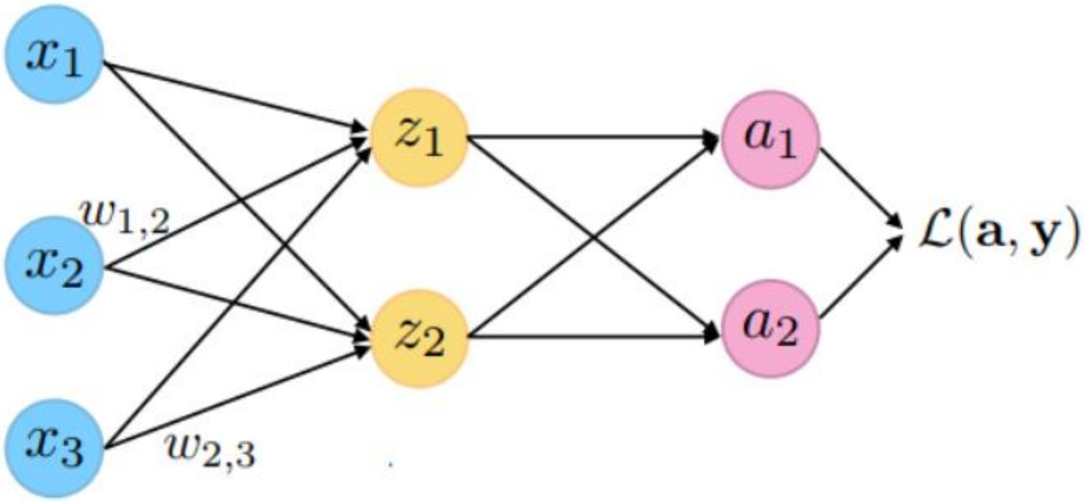


$\mathbf{W} \in \Re^{K\times m}$ $\quad \mathbf{z} \in \Re^{K\times 1}$ $\quad \mathbf{a} \in \Re^{K\times 1}$

$\mathbf{Z} \in \Re^{b\times K}$ $\quad \mathbf{A} \in \Re^{b\times K}$

$k = 1, \ldots, K$: llega
$j = 1, \ldots, m$: sale

Cálculo de las derivadas



$$\frac{\partial L}{\partial w_{1,2}} = \underbrace{\boxed{\frac{\partial L}{\partial a_1}}}_{-\frac{y_1}{a_1}} \underbrace{\boxed{\frac{\partial a_1}{\partial z_1}}}_{a_1(1-a_1)} \underbrace{\boxed{\frac{\partial z_1}{\partial w_{1,2}}}}_{x_2} + \underbrace{\boxed{\frac{\partial L}{\partial a_2}}}_{-\frac{y_2}{a_2}} \underbrace{\boxed{\frac{\partial a_2}{\partial z_1}}}_{-a_2 a_1} \underbrace{\boxed{\frac{\partial z_1}{\partial w_{1,2}}}}_{x_2}$$

$$\frac{\partial L}{\partial a_1} = \frac{\partial}{\partial a_1}\left[\sum_{j=1}^{h} -y_j \log(a_j)\right]$$

$$= \frac{\partial}{\partial a_1}\left[-y_1 \log(a_1)\right]$$

$$= -\frac{y_1}{a_1}$$

$$= \frac{\partial}{\partial w_{1,2}}\left[w_{1,2} \cdot x_2 + b\right]$$

$$= x_2$$

$$\frac{\partial a_1}{\partial z_1} = \frac{\partial}{\partial z_1}\left[\frac{e^{z_1}}{\sum_{j=1}^{h} e^{z_j}}\right]$$

$$= \frac{\left[\sum_{j=1}^{h} e^{z_j}\right]\frac{\partial}{\partial z_1}e^{z_1} - e^{z_1}\frac{\partial}{\partial z_1}\left[\sum_{j=1}^{h} e^{z_j}\right]}{\left[\sum_{j=1}^{h} e^{z_j}\right]^2}$$

$$= \frac{\left[\sum_{j=1}^{h} e^{z_j}\right]e^{z_1} - e^{z_1}e^{z_1}}{\left[\sum_{j=1}^{h} e^{z_j}\right]^2}$$

$$= \frac{e^{z_1}\left(\left[\sum_{j=1}^{h} e^{z_1}\right] - e^{z_1}\right)}{\left[\sum_{j=1}^{h} e^{z_j}\right]^2}$$

$$= \frac{e^{z_1}}{\left[\sum_{j=1}^{h} e^{z_j}\right]} \cdot \frac{\left[\sum_{j=1}^{h} e^{z_j}\right] - e^{z_1}}{\left[\sum_{j=1}^{h} e^{z_j}\right]} \quad = a_1(1-a_1)$$

$$\frac{\partial a_2}{\partial z_1} = \frac{\partial}{\partial z_1}\left[\frac{e^{z_2}}{\sum_{j=1}^{h} e^{z_j}}\right]$$

$$= \frac{\left[\sum_{j=1}^{h} e^{z_j}\right]\frac{\partial}{\partial z_1}e^{z_2} - e^{z_2}\frac{\partial}{\partial z_1}\left[\sum_{j=1}^{h} e^{z_j}\right]}{\left[\sum_{j=1}^{h} e^{z_j}\right]^2}$$

$$= \frac{0 - e^{z_2}e^{z_1}}{\left[\sum_{j=1}^{h} e^{z_j}\right]^2}$$

$$= \frac{-e^{z_2}}{\left[\sum_{j=1}^{h} e^{z_j}\right]} \cdot \frac{e^{z_1}}{\left[\sum_{j=1}^{h} e^{z_j}\right]} \quad = -a_2a_1$$

	Function	Derivative
Sum Rule	$f(x)+g(x)$	$f'(x)+g'(x)$
Difference Rule	$f(x)-g(x)$	$f'(x)-g'(x)$
Product Rule	$f(x)g(x)$	$f'(x)g(x)+f(x)g'(x)$
Quotient Rule	$f(x)/g(x)$	$[g(x)f'(x)-f(x)g'(x)]/[g(x)]^2$
Reciprocal Rule	$1/f(x)$	$-[f'(x)]/[f(x)]^2$
Chain Rule	$f(g(x))$	$f'(g(x))g'(x)$

Forma matricial

$$\nabla_{\mathbf{W}} \mathcal{L} = \mathbf{X}^{\top}(A - Y)$$

¿De dónde viene?

Tomemos los vectores $\mathbf{z}^{[i]} \in \Re^{K \times 1}$ y $\mathbf{y}^{[i]} \in \Re^{K \times 1}$: Probabilidad de pertenencia de la muestra i a cada una de las clases, y etiqueta de la muestra i en *one-hot-encoding*

$l(\mathbf{y}, \mathbf{z}) = -\sum_{j=1}^{K} y_j \log \left(\frac{\exp(z_j)}{\sum_{k=1}^{K} \exp(z_k)} \right)$

$l(\mathbf{y}, \mathbf{z}) = \sum_{j=1}^{K} y_j \log \sum_{k=1}^{K} \exp(z_k) - \sum_{j=1}^{K} y_j z_j$

$l(\mathbf{y}, \mathbf{z}) = \log \sum_{k=1}^{K} \exp(z_k) - \sum_{j=1}^{K} y_j z_j$

Tomando la derivada con respecto a cualquier *logit* z_j

$$\frac{\partial l}{\partial z_j} = \frac{\exp(z_j)}{\sum_{k=1}^{K}\exp(z_k)} - y_j$$

$$\frac{\partial l}{\partial z_j} = a_j - y_j$$

De forma vectorial:

$$\frac{\partial l}{\partial \mathbf{z}} = \mathbf{a}^{[i]} - \mathbf{y}^{[i]}$$

De forma matricial:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{Z}} = \mathbf{A} - \mathbf{Y}$$

La derivada de una función lineal de la forma:

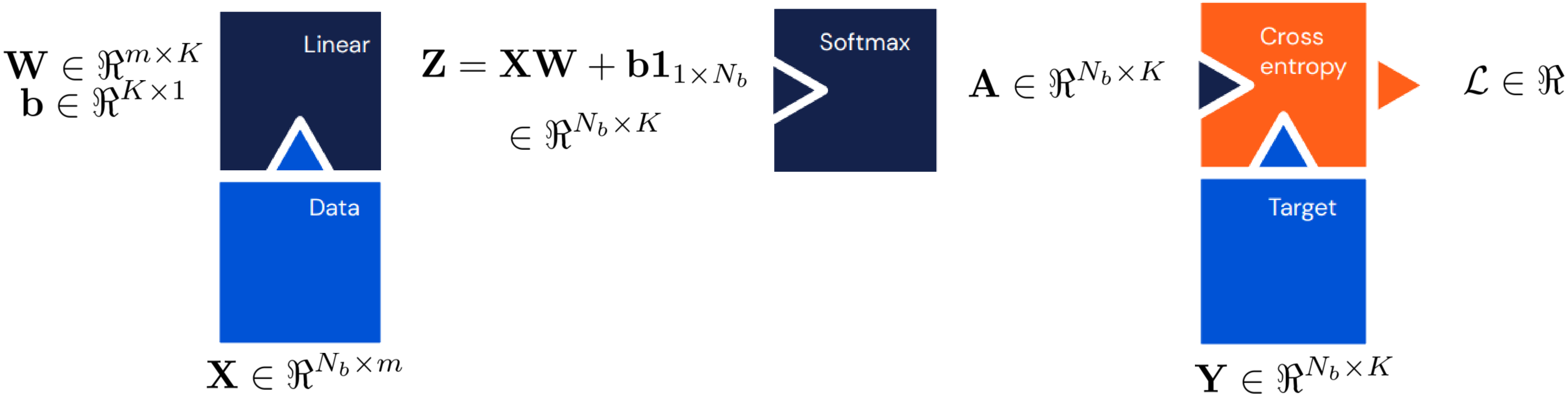
$\mathbf{Z} = \mathbf{X}\mathbf{W}^\top + \mathbf{b}$ con respecto a los parámetros $\mathbf{W}, \mathbf{b}$:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \mathbf{X}^\top \frac{\partial \mathcal{L}}{\partial \mathbf{Z}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}} = \frac{\partial \mathcal{L}}{\partial \mathbf{Z}}^\top \mathbf{1}_n$$

`http://cs231n.stanford.edu/handouts/linear-backprop.pdf`

Forward

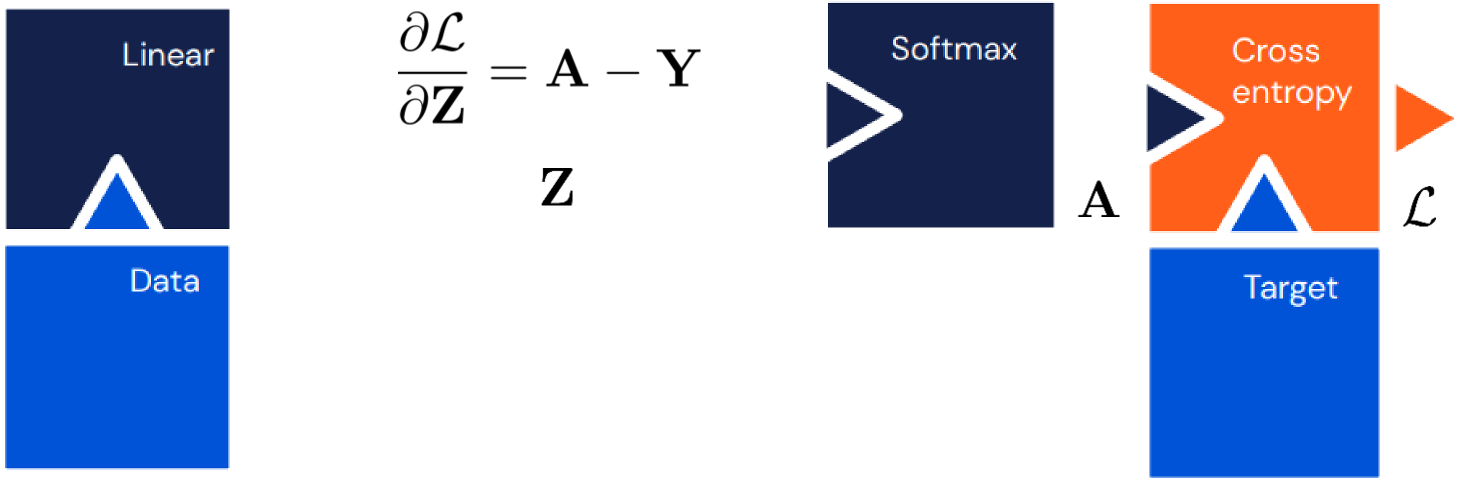

$\mathbf{W} \in \Re^{m \times K}$
$\mathbf{b} \in \Re^{K \times 1}$
Linear
Data
$\mathbf{X} \in \Re^{N_b \times m}$
$\mathbf{Z} = \mathbf{X}\mathbf{W} + \mathbf{b}\mathbf{1}_{1 \times N_b}$
$\in \Re^{N_b \times K}$
Softmax
$\mathbf{A} \in \Re^{N_b \times K}$
Cross entropy
Target
$\mathbf{Y} \in \Re^{N_b \times K}$
$\mathcal{L} \in \Re$

Backward

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}} = \mathbf{X}^\top \frac{\partial \mathcal{L}}{\partial \mathbf{Z}}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}} = \frac{\partial \mathcal{L}}{\partial \mathbf{Z}}^\top \mathbf{1}_{N_b}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{Z}} = \mathbf{A} - \mathbf{Y}$$



$$\mathbf{W} \in \Re^{m \times K}$$
$$\mathbf{b} \in \Re^{1 \times K}$$
$$\mathbf{X} \in \Re^{N_b \times m}$$

$$\mathbf{Z} \in \Re^{N_b \times K}$$
$$\mathbf{A} \in \Re^{N_b \times K}$$
$$\mathbf{Y} \in \Re^{N_b \times K}$$
$$\mathcal{L} \in \Re$$

¿Preguntas?