



Redes neuronales

Redes multi-clase

Softmax

Ejemplo multi-clase: MNIST

Dataset con dígitos manuscritos – 10 clases



Objetivo



???



0 1 2 3 4 5 6 7 8 9

Objetivo




???



0 1 2 3 4 5 6 7 8 9

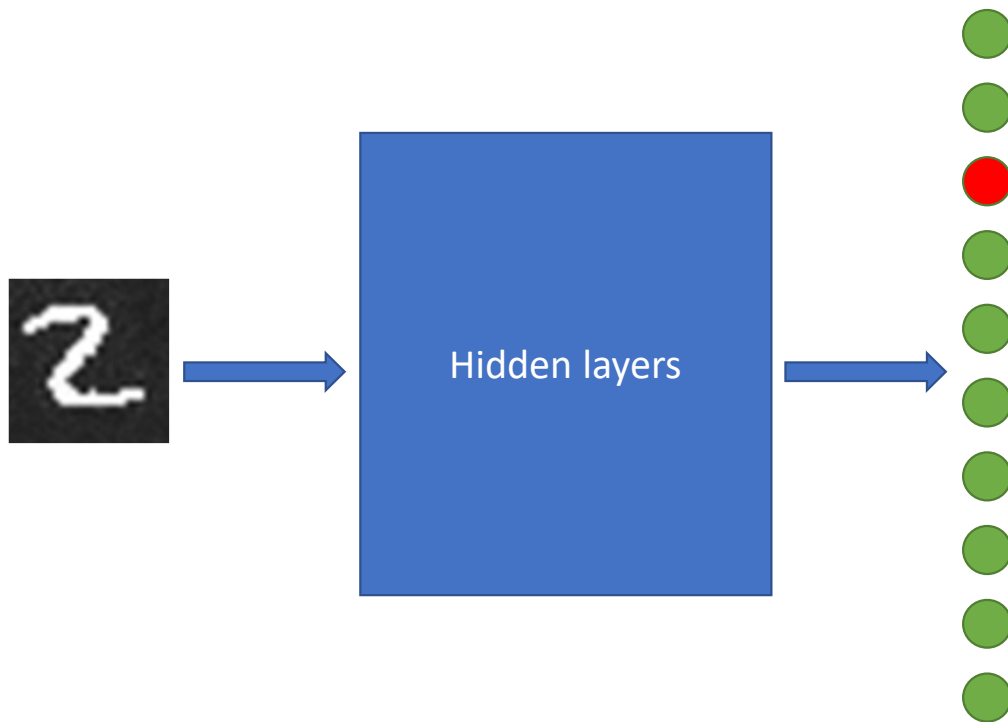
Nuestra red debería computar una probabilidad de que la entrada pertenezca a una clase.

La salida ideal debería ser el one-hot encoding para la clase correcta

$[0, 0, 1, 0, 0, 0, 0, 0, 0, 0]$  Target de entrenamiento!

Salida multi-clase

- Necesitamos una red neuronal con n neuronas de salida



Podemos usar neuronas sigmoide en la capa de salida, sin embargo esto no garantiza que la salida sea una función de densidad de probabilidades(pdf). Podemos normalizar la salida para obtener una pdf.

Función de activación Softmax

Recordemos la computación de la función lineal en la última capa de una NN (antes de la función de activación)

$$z_L = X_{L-1} \cdot W^L + b^L \quad L \text{ es el número de capas}$$

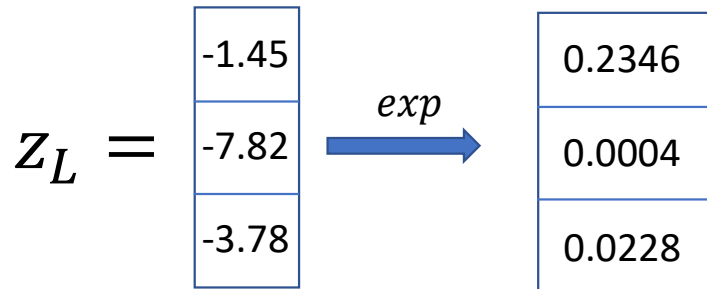
z_L es un vector real de dimensión n . Por ejemplo, supongamos que tenemos una NN con 3 neuronas en la salida:

$$z_L = \begin{array}{|c|} \hline -1.45 \\ \hline -7.82 \\ \hline -3.78 \\ \hline \end{array}$$

Cuál es la interpretación de estos valores?

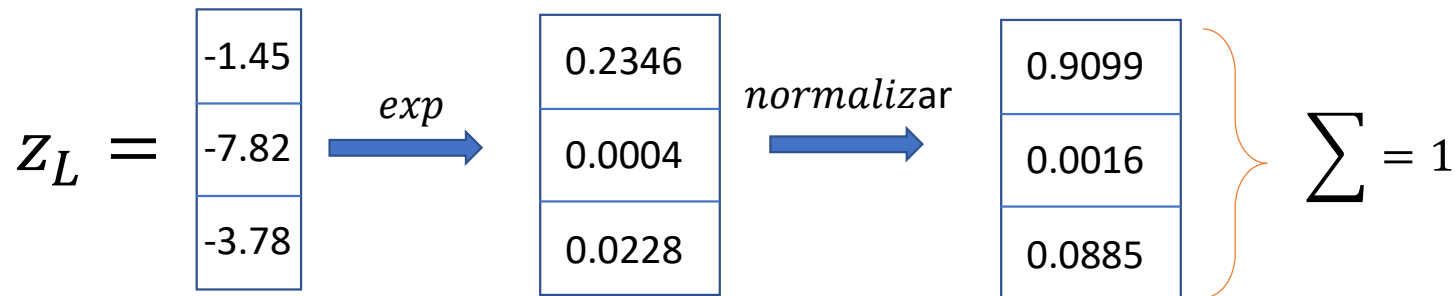
Función de activación Softmax

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_k e^{z_k}}$$



Función de activación Softmax

$$\text{Softmax}(z_i) = \frac{e^{z_i}}{\sum_k e^{z_k}}$$



Softmax transforma cualquier salida en una *función de densidad de probabilidades (pdf)*

Softmax como función de activación de la capa de salida

Resumen

- Softmax como función de activación para producir una pdf
- Clasificador multi-clase

Redes neuronales

Redes multi-clase

Cross-entropy Loss



Cross-entropy Loss


- Nuestra red ahora produce una pdf
- La salida objetivo (one-hot encoding) es también una pdf
- Cómo comparamos dos pdf's: Cross-entropy loss

$$L(y, y') = - \sum_{i=1}^n y_i \log(y'_i)$$

Cross-entropy Loss

- Por ejemplo, para el problema de 3 clases, podemos tener

$$L(y, y') = - \sum_{i=1}^n y_i \log(y'_i)$$



1
0
0

One-hot encoding

0.9099
0.0016
0.0885

Salida Softmax

Cross-entropy Loss

- Por ejemplo, para el problema de 3 clases, tenemos

$$L(y, y') = - \sum_{i=1}^n y_i \log(y'_i)$$

1
0
0

-0.0944
-6.4377
-2.4247

One-hot encoding

Cross-entropy Loss

- Por ejemplo, para el problema de 3 clases, tenemos

$$L(y, y') = - \sum_{i=1}^n y_i \log(y'_i)$$

Por lo tanto,

$$L(y, y') = 0.0944$$

One-hot encoding

1	-0.0944	-0.0944
0	-6.4377	0
0	-2.4247	0

Cross-entropy Loss: Explicación

- Ejemplo: queremos usar una antena para enviar el estado del clima



code

bits

				
Soleado	Nublado	Nublado parcial	Lluvia	Tormenta
000	001	010	011	100
3	3	3	3	3

Cantidad promedio de bits usados = 3

Cross-entropy Loss: Explicación

- Ejemplo: queremos usar una antena para enviar el estado del clima



Probabilidad del clima en una ciudad A

$p(x)$
code
bits

				
Soleado	Nublado	Nublado parcial	Lluvia	Tormenta
0.25	0.45	0.2	0.09	0.01
000	001	010	011	100
3	3	3	3	3

$$\text{Bits promedio} = 3(0.25) + 3(0.45) + 3(0.2) + 3(0.09) + 3(0.01) = 3$$

Cross-entropy Loss: Explanation

- Ejemplo: queremos usar una antena para enviar el estado del clima



Usemos mejores códigos dependiendo de la probabilidad

$p(x)$
code
bits

				
Soleado	Nublado	Nublado parcial	Lluvia	Tormenta
0.25	0.45	0.2	0.09	0.01
1100	0	10	1101	1110
4	1	2	4	4

$$\text{Bits promedio} = 4(0.25) + 1(0.45) + 2(0.2) + 4(0.09) + 4(0.01) = 2.25$$

Cross-entropy Loss: Explicación

- Ejemplo: queremos usar una antenna para enviar el estado del clima

					
	Soleado	Nublado	Nublado parcial	Lluvia	Tormenta
$p(x)$	0.25	0.45	0.2	0.09	0.01

Cómo obtenemos el mínimo número de bits?

Cross-entropy Loss: Explicación

- Ejemplo: queremos usar una antena para enviar el estado del clima

					
	Soleado	Nublado	Nublado parcial	Lluvia	Tormenta
$p(x)$	0.25	0.45	0.2	0.09	0.01
$1/p(x)$	4	2.22	5	11.11	100

Cómo obtenemos el mínimo número de bits?

Cross-entropy Loss: Explanation

- Ejemplo: queremos usar una antena para enviar el estado del clima

					
	Soleado	Nublado	Nublado parcial	Lluvia	Tormenta
$p(x)$	0.25	0.45	0.2	0.09	0.01
$1/p(x)$	4	2.22	5	11.11	100
$-\log_2 p(x)$	2	1.15	2.32	3.47	6.64

Cómo obtenemos el mínimo número de bits?

$$\log_2 \left(\frac{1}{p(x)} \right) = -\log_2 p(x)$$

$$-\log_2 p(x)$$

$$\text{Bits promedio} = 2(0.25) + 1.15(0.45) + 2.32(0.2) + 3.47(0.09) + 6.64(0.01) = 1.86$$

Cross-entropy Loss: Explanation

- Ejemplo: queremos usar una antena para enviar el estado del clima

					
	Soleado	Nublado	Nublado parcial	Lluvia	Tormenta
$p(x)$	0.25	0.45	0.2	0.09	0.01
$1/p(x)$	4	2.22	5	11.11	100
$-\log_2 p(x)$	2	1.15	2.32	3.47	6.64

Cómo obtenemos el mínimo número de bits?

$$\log_2 \left(\frac{1}{p(x)} \right) = -\log_2 p(x)$$

$$-\sum p(x) \log p(x) \quad \text{Entropía}$$

$$\text{Bits promedio} = 2(0.25) + 1.15(0.45) + 2.32(0.2) + 3.47(0.09) + 6.64(0.01) = 1.86$$

Cross-entropy Loss: Explanation

- Ejemplo: queremos usar una antena para enviar el estado del clima

					
	Soleado	Nublado	Nublado parcial	Lluvia	Tormenta
$p(x)$	0.25	0.45	0.2	0.09	0.01
$1/p(x)$	4	2.22	5	11.11	100
$-\log_2 p(x)$	2	1.15	2.32	3.47	6.64

Cómo obtenemos el mínimo número de bits?

$$\log_2 \left(\frac{1}{p(x)} \right) = -\log_2 p(x)$$

$$-\sum p(x) \log p(x) \quad \text{Entropía}$$

La entropía es 1.86: cantidad de bits para representar la información que obtienes si sabes el clima de la ciudad A

Y si usamos esto para una antena en ciudad B?

					
	Soleado	Nublado	Nublado parcial	Lluvia	Tormenta
$p(x)$ in A	0.25	0.45	0.2	0.09	0.01
$-\log_2(p_A)$	2	1.15	2.32	3.47	6.64
$p(x)$ in B	0.3	0.2	0.2	0.25	0.05

Cuál es la cantidad de bits para la antena B?

$$-\sum p(x) \log q(x)$$

Cross-entropy

$$\text{Bits promedio} = 0.3(2) + 0.2(1.15) + 0.2(2.32) + 0.25(3.47) + 0.05(6.64) = 2.5$$

Y si usamos esto para una antena en ciudad B?

					
	Soleado	Nublado	Nublado parcial	Lluvia	Tormenta
$p(x)$ in A	0.25	0.45	0.2	0.09	0.01
$-\log_2(p_A)$	2	1.15	2.32	3.47	6.64
$p(x)$ in B	0.3	0.2	0.2	0.25	0.05

Cuál es la cantidad de bits para la antena B?

$$-\sum p(x) \log q(x)$$

Cross-entropy

La cross-entropy es 2.5: cantidad de información que obtienes en promedio usando la información de clima de la ciudad B y la codificación óptima de la ciudad A.

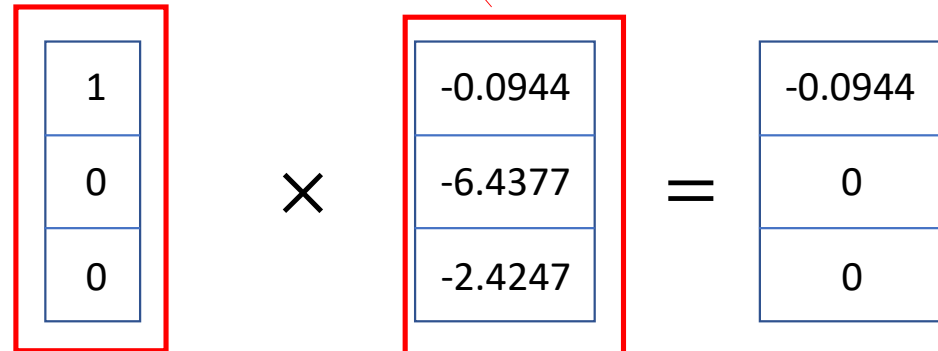
Cross-entropy

$$L(y, y') = - \sum_{i=1}^n y_i \log(y'_i)$$

Por lo tanto,

$$L(y, y') = 0.0944$$

Información que obtienes por conocer la data real



1
0
0

One-hot encoding

×

-0.0944
-6.4377
-2.4247

=

-0.0944
0
0

Ejemplo multi-clase en Pytorch



Resumen

- Softmax + cross entropy loss para problemas multi-clase
- Derivación de gradiente para softmax + cross entropy loss en material suplementario en U-Cursos