Entrenamiento de la red

La función de coste

Para entrenar la red utilizaremos un proceso de optimización. En primer lugar, debemos definir qué función queremos optimizar. Necesitamos una función que, dada una combinación de parámetros de la red, devuelva un valor alto cuando las predicciones con esos parámetros sean malas y un valor bajo cuando estas sean buenas. Esto es lo que denominamos la **función de coste** (J). Definiremos una **función de pérdida** ($\mathcal{L}$) que reciba como entrada una predicción y una etiqueta real y nos indique cómo de desacertada es la predicción. En este caso utilizaremos la entropía cruzada binaria, descrita como:

$$\mathcal{L}(y_{pred}, y_{etiqueta}) = -y_{etiqueta} \log(y_{pred}) - (1 - y_{etiqueta}) \log(1 - y_{pred})$$

La función de coste (J) será la media de la función de pérdida en los m ejemplos del conjunto de entrenamiento:

$$J(\mathbf{W}, \mathbf{b}) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(y_{pred}^{(i)}, y_{etiqueta}^{(i)})$$

Si minimizamos J, nuestras predicciones serán mejores. Para minimizar J utilizaremos **descenso de gradiente**: haremos sucesivos pasos en los que calcularemos el gradiente de J respecto a los distintos parámetros ($\mathbf{W}, \mathbf{b}$) y actualizaremos los parámetros en la dirección del gradiente, con la esperanza de que el siguiente paso obtenga un valor de J menor. Repetiremos este proceso durante un número fijo de pasos.

Por tanto, el algoritmo que debemos aplicar es el siguiente:

1. Calcular la pérdida de las predicciones con los valores actuales de $\mathbf{W}$ y $\mathbf{b}$
2. Calcular el gradiente respecto $\mathbf{W}$ y $\mathbf{b}$.
3. Actualizar $\mathbf{W}$ y $\mathbf{b}$ en la dirección de sus gradientes respectivos.

Gradientes

El segundo paso nos obliga a ser capaces de calcular el gradiente de $J(\mathbf{W}, \mathbf{b})$ respecto a cada uno de los parámetros, es decir, la derivada parcial de $J(\mathbf{W}, \mathbf{b})$ respecto a cada parámetro. Para ello iremos propagando el gradiente hacia atrás, calculando a cada paso el gradiente en el nodo anterior a partir de los posteriores. Calculemos, por ejemplo, el gradiente de $J(\mathbf{W}, \mathbf{b})$ respecto a z_2:

$$\frac{\partial J(\mathbf{W}, \mathbf{b})}{\partial z_2} = \frac{\partial(\frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(y_{pred}^{(i)}, y_{etiqueta}^{(i)}))}{\partial z_2} = \frac{1}{m} \sum_{i=1}^{m} \frac{\partial \mathcal{L}(y_{pred}^{(i)}, y_{etiqueta}^{(i)})}{\partial z_2}$$

El problema se reduce, por tanto, a calcular el gradiente de la pérdida respecto a z_2. Como $y_{etiqueta}$ es una función de z_2, debemos aplicar la regla de la cadena:

$$\frac{\partial \mathcal{L}(y_{pred}, y_{etiqueta})}{\partial z_2} = \frac{\partial \mathcal{L}(y_{pred}, y_{etiqueta})}{\partial y_{pred}} \frac{\partial y_{pred}}{\partial z_2} \tag{1}$$

Aplicando las reglas de derivación, sabemos que:

$$\frac{\partial \mathcal{L}(y_{pred}, y_{etiqueta})}{\partial y_{pred}} = \frac{y_{pred} - y_{etiqueta}}{y_{pred}(1 - y_{pred})}$$

$$\frac{\partial y_{pred}}{\partial z_2} = y_{pred}(1 - y_{pred})$$

Sustituyendo en (1), tenemos:

$$\frac{\partial \mathcal{L}(y_{pred}, y_{etiqueta})}{\partial z_2} = y_{pred} - y_{etiqueta}$$

Propagación hacia atrás

De la misma manera que hemos utilizado el gradiente respecto a y_pred para calcular el gradiente respecto a z_2, ahora utilizaremos el gradiente respecto a z_2 para calcular el resto. En esto consiste el **propagar hacia atrás** el gradiente. Recordemos las ecuaciones de nuestro modelo:

$$\mathbf{z}_0 = \mathbf{W}_0^T \mathbf{x} + \mathbf{b}_0$$

$$\mathbf{h}_0 = sigmoide(\mathbf{z}_0)$$

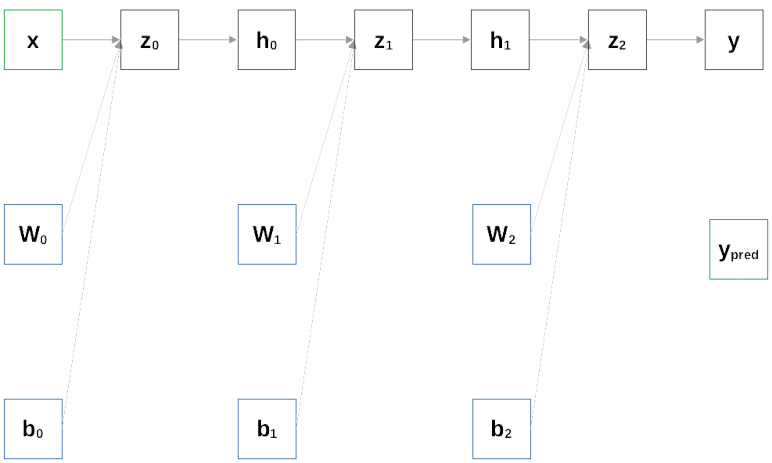
$$\mathbf{z}_1 = \mathbf{W}_1^T \mathbf{h}_0 + \mathbf{b}_1$$

$$\mathbf{h}_1 = sigmoide(\mathbf{z}_1)$$

$$z_2 = \mathbf{W}_2^T \mathbf{h}_1 + \mathbf{b}_2$$

$$y_{pred} = sigmoide(z_2)$$

Estas ecuaciones se pueden resumir en este grafo de operaciones:



Notación matricial

Para generalizar la regla de la cadena a vectores o matrices podemos calcular el gradiente respecto a cada componente del vector independientemente. Si $\mathbf{z} = f(\mathbf{x})$ e $y = g(\mathbf{z})$, la regla de la cadena para una componente x_i del vector $\mathbf{x}$ queda de la siguiente forma:

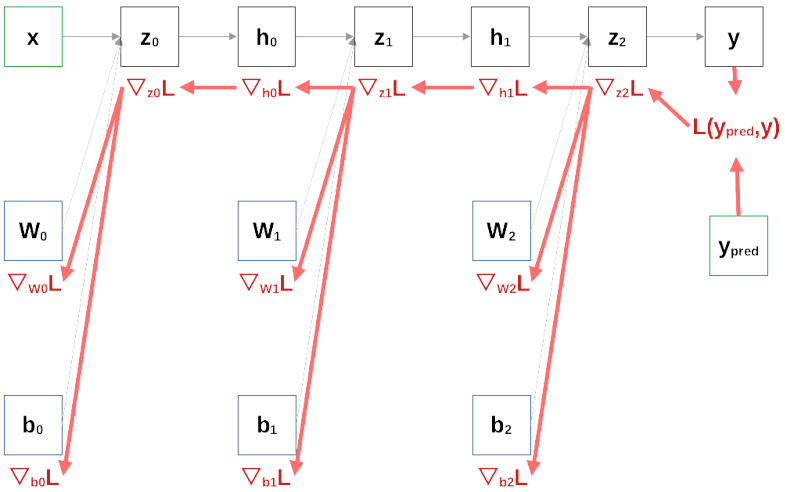
$$\frac{\partial y}{\partial x_i} = \sum_j \frac{\partial y}{\partial z_j} \frac{\partial z_j}{\partial x_i}$$

Por tanto, podemos expresar la regla de la cadena para el gradiente respecto al vector completo $\mathbf{x}$ así (puedes consultar más en detalle cómo se aplica la regla de la cadena sobre vectores y matrices en este enlace: https://explained.ai/matrix-calculus/):

$$\nabla_{\mathbf{x}} \mathbf{y} = \left(\frac{\partial \mathbf{z}}{\partial \mathbf{x}} \right)^\top \nabla_{\mathbf{z}} \mathbf{y}$$

Obtención de los gradientes

Usando esta fórmula, podemos calcular los gradientes que nos interesan siguiendo el grafo de operaciones:



Aplicando la regla de la cadena y las reglas de derivación, obtenemos las siguientes expresiones para los parámetros de la capa C_2:

$$\nabla_{\mathbf{b}_2} \mathcal{L}(y_{pred}, y_{etiqueta}) = \left(\frac{\partial z_2}{\partial \mathbf{b}_2} \right)^\top \nabla_{\mathbf{z}_2} \mathcal{L}(y_{pred}, y_{etiqueta}) = \vec{1} \, \frac{\partial \mathcal{L}(y_{pred}, y_{etiqueta})}{\partial z_2}$$

$$\nabla_{\mathbf{W}_2} \mathcal{L}(y_{pred}, y_{etiqueta}) = \left(\frac{\partial z_2}{\partial \mathbf{W}_2} \right)^\top \nabla_{\mathbf{z}_2} \mathcal{L}(y_{pred}, y_{etiqueta}) = \mathbf{h}_1^T \, \frac{\partial \mathcal{L}(y_{pred}, y_{etiqueta})}{\partial z_2}$$

.

Calculando $\nabla_{\mathbf{h}_1} \mathcal{L}(y_{pred}, y_{etiqueta})$ y, a partir de él, $\nabla_{\mathbf{z}_1} \mathcal{L}(y_{pred}, y_{etiqueta})$ (al que llamaremos **dLdz1** para simplificar la notación) podemos propagar hacia atrás y obtener los gradientes respecto a los parámetros de la capa C_1 ($\mathbf{b}_1$ y $\mathbf{W}_1$):

$$\nabla_{\mathbf{h}_1} \mathcal{L}(y_{pred}, y_{etiqueta}) = \vec{1} \, \frac{\partial \mathcal{L}(y_{pred}, y_{etiqueta})}{\partial z_2} \, \mathbf{W}_2^T$$

$$\nabla_{\mathbf{z}_1} \mathcal{L}(y_{pred}, y_{etiqueta}) = \nabla_{\mathbf{h}_1} \mathcal{L}(y_{pred}, y_{etiqueta}) \, \text{diag}(\mathbf{h}_1) \, \text{diag}(\vec{1} - \mathbf{h}_1) = \mathbf{dLdz1}$$

$$\nabla_{\mathbf{b}_1} \mathcal{L}(y_{pred}, y_{etiqueta}) = \vec{1} \sum_{i=1}^{5} dLdz1_i$$

$$\nabla_{\mathbf{W}_1} \mathcal{L}(y_{pred}, y_{etiqueta}) = \text{diag}(\mathbf{dLdz1}) \, \mathbf{h}_0^T$$

Por último, con $\nabla_{\mathbf{h}_0} \mathcal{L}(y_{pred}, y_{etiqueta})$ y, a partir de él, $\nabla_{\mathbf{z}_0} \mathcal{L}(y_{pred}, y_{etiqueta})$ (al que llamaremos $\mathbf{dLdz0}$ para simplificar la notación), podemos propagar hacia atrás y obtener los gradientes respecto a los parámetros de la capa C_0 ($\mathbf{b}_0$ y $\mathbf{W}_0$):

$$\nabla_{\mathbf{h}_0} \mathcal{L}(y_{pred}, y_{etiqueta}) = \text{diag}(\mathbf{dLdz1}) \, \mathbf{W}_1$$

$$\nabla_{\mathbf{z}_0} \mathcal{L}(y_{pred}, y_{etiqueta}) = \nabla_{\mathbf{h}_0} \mathcal{L}(y_{pred}, y_{etiqueta}) \, \text{diag}(\mathbf{h}_0) \, \text{diag}(\vec{1} - \mathbf{h}_0) = \mathbf{dLdz0}$$

$$\nabla_{\mathbf{b}_0} \mathcal{L}(y_{pred}, y_{etiqueta}) = \vec{1} \sum_{i=1}^{5} dLdz0_i$$

$$\nabla_{\mathbf{W}_0} \mathcal{L}(y_{pred}, y_{etiqueta}) = \text{diag}(\mathbf{dLdz0}) \, \mathbf{x}^T$$