Centro de Investigación en Matemáticas (CIMAT). Unidad Monterrey

Tarea 1

Marcelo Alberto Sanchez Zaragoza

27 de agosto de 2021

1. Problema 3

Considera un problema de clasificación multiclase y una red neuronal densa con una capa oculta, como se muestra en la figura 1.1.

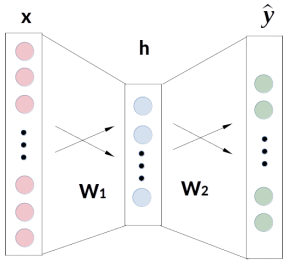

Figura 1.1: Figura 1

Consideraremos también el uso de la función sigmoide como activación de las unidades ocultas, la función softmax para las estimaciones en la capa de salida y cross-entropy como función de costo.

a) Muestra que softmax es invariante a traslaciones (constantes) del vector de entrada, es decir, para cualquier vector $\mathbf{x}$ y cualquier constante c :

$$\text{softmax}(\mathbf{x}) = \text{softmax}(\mathbf{x} + c),$$

donde la operación $\mathbf{x} + c$ se realiza con broadcasting. Recuerda que

$$\text{softmax}(\mathbf{x})_i = \frac{e^{\mathbf{x}_i}}{\sum_j e^{\mathbf{x}_j}}.$$

Lo anterior es útil cuando se escoge $c = -\text{máx}(\mathbf{x})$, es decir, quitando el valor mayor en todos los elementos de $\mathbf{x}$, para estabilidad numérica.

b) Para un escalar $\mathbf{x}$, muestra que el gradiente de la función sigmoide es $\sigma(x)(1 - \sigma(x))$.

c) Muestra que el gradiente en la capa de salida es

$$\frac{\partial L(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{z}} = \hat{\mathbf{y}} - \mathbf{y},$$

donde $\hat{\mathbf{y}} = \text{softmax}(\mathbf{z})$, para algún vector $\mathbf{z}$ que proviene de la capa de salida.

La función de costo, como mencionamos al inicio, es la cross-entropy: $L(\mathbf{y}, \hat{\mathbf{y}}) = -\sum_i y_i \log(\hat{y}_i)$, donde $\mathbf{y}$ es un vector $one-hot$ de las clases y $\hat{\mathbf{y}}$ es el vector de probabilidades estimadas.

d) Considerando los incisos anteriores, obtén los gradientes respecto a los inputs $\mathbf{x}$, es decir, calcula

$$\frac{\partial L(\mathbf{y}, \hat{\mathbf{y}})}{\partial \mathbf{x}}.$$

Recuerda que el paso forward calcula las activaciones: $\mathbf{h} = \sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)$ y $\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}_2\mathbf{h} + \mathbf{b}_2)$.

Recuerda también que la función de activación en un vector (tensor), se aplica entrada por entrada.
Basándote en lo anterior, obtén las ecuaciones de backpropagation para la red neuronal.

Solución

Inciso a)
Para demostrar el primer inciso se parte de la siguiente expresión:

$$\text{softmax}(\mathbf{x})_i = \frac{e^{\mathbf{x}_i}}{\sum_j e^{\mathbf{x}_j}}.$$

donde lo que se esta realizando es la función softmax a cada una de las entradas del vector x, partiendo de lo anterior tenemos:

$$\begin{aligned}\text{softmax}(\mathbf{x})_i = \frac{e^{\mathbf{x}_i}}{\sum_j e^{\mathbf{x}_j}} = \frac{e^{\mathbf{x}_i}}{\sum_j e^{\mathbf{x}_j}} \left(\tfrac{e^c}{e^c}\right) = \frac{e^{\mathbf{x}_i} e^c}{e^c \sum_j e^{\mathbf{x}_j}} = \\ = \frac{e^{\mathbf{x}_i + c}}{\sum_j e^c e^{\mathbf{x}_j}} = \frac{e^{\mathbf{x}_i + c}}{\sum_j e^{\mathbf{x}_j + c}} = \text{softmax}(\mathbf{x}_i + c)\end{aligned}$$

ahora se parte de $\text{softmax}(\mathbf{x}_i + c)$, así tenemos:

$$\begin{aligned}\text{softmax}(\mathbf{x}_i + c) = \frac{e^{\mathbf{x}_i + c}}{\sum_j e^{\mathbf{x}_j + c}} = \frac{e^{\mathbf{x}_i} e^c}{\sum_j e^c e^{\mathbf{x}_j}} = \frac{e^{\mathbf{x}_i} e^c}{e^c \sum_j e^{\mathbf{x}_j}} = \\ = \frac{e^{\mathbf{x}_i}}{\sum_j e^{\mathbf{x}_j}} \left(\tfrac{e^c}{e^c}\right) = \frac{e^{\mathbf{x}_i}}{\sum_j e^{\mathbf{x}_j}} = \text{softmax}(\mathbf{x}_i)\end{aligned}$$

Al final se realiza la misma operación a cada una de las entradas del vector $\mathbf{x}$ y obtenemos que $\text{softmax}(\mathbf{x}) = \text{softmax}(\mathbf{x} + c)$

Inciso b)
Recordemos como es la función sigmoide:

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

Ahora para un escalar $\mathbf{x}$ partimos de lo siguiente:

$$\frac{\partial \sigma(x)}{\partial x} = (-1)(1 + e^{-x})^{-2}(e^{-x}) = \frac{e^{-x}}{(1 + e^{-x})^2} = \left(\frac{e^{-x}}{1+e^{-x}}\right)\left(\frac{1}{1+e^{-x}}\right) = \left(\frac{1-1+e^{-x}}{1+e^{-x}}\right)\left(\frac{1}{1+e^{-x}}\right) = \left(1 - \frac{1}{1+e^{-x}}\right)\left(\frac{1}{1+e^{-x}}\right) = (1 - \sigma(z))\sigma(z)$$

Donde ya hemos encontrado lo que piden demostrar.

Inciso c)
Para la siguiente demostración vamos a comenzar con lo siguiente:

$$\frac{\partial}{\partial z_j} log(\hat{y}_i) = \frac{1}{\hat{y}_i} \frac{\partial \hat{y}_i}{\partial z_j}$$

donde antes vamos a encontrar la derivada de: $y_i \frac{\partial log(\hat{y}_i)}{\partial z_j}$ y $log(\hat{y}_i)$ es:
$log(\hat{y}) = log(\frac{e^{z_i}}{\sum_j e^{z_j}}) = z_i - log(\sum_j e^{z_j})$.
Así tenemos:

$$\frac{\partial}{\partial z_j} log(\hat{y}_i) = \frac{\partial z_i}{\partial z_j} - \frac{\partial}{\partial z_j} log(\sum_j e^{z_j})$$

donde $\frac{\partial z_i}{\partial z_j}$ es 1 si $i = j$ y 0 en otro caso.

$$\frac{\partial z_i}{\partial z_j} - \frac{\partial}{\partial z_j} log(\sum_j e^{z_j}) = 1\{i = j\} - \frac{1}{\sum_j e^{z_j}} (\frac{\partial}{\partial z_j} \sum_j e^{z_j})$$

Observe que:

$$\frac{\partial}{\partial z_j}\sum_j e^{z_j} = \frac{\partial}{\partial z_j}[e^{z_1} + e^{z_2} + ... + e^{z_n}] = e^{z_i}$$

regresando tenemos:

$$\frac{\partial}{\partial z_j} log(\hat{y}_i) = 1\{i = j\} - \frac{e^{z_j}}{\sum_j e^{z_j}} = 1\{i = j\} - \hat{y}_j$$

Ahora nuestra expresión $\hat{y}_i \frac{\partial log(\hat{y}_i)}{\partial z_j} = \hat{y}_i(1\{i = j\} - \hat{y}_j)$.

Resolviendo llegamos a:

$$\frac{\partial L(Y, \hat{Y})}{\partial z_j} = -\frac{\partial}{\partial z_j}\sum_i Y_i log(\hat{y}_i) = -\sum_i Y_i \frac{\partial}{\partial z_j} log(\hat{y}_i) =$$
$$-\sum_i \frac{Y_i}{\hat{y}_i}\frac{\partial \hat{y}_i}{\partial z_j} = -\sum_i \frac{y_i}{\hat{y}_i}\hat{y}_i(1\{i = j\} - \hat{y}_j)$$

Observe que se puede rescribir como:

$$\frac{\partial L(Y, \hat{Y})}{\partial z_j} = [-\sum_{i \neq j} y_i(1\{i = j\} - \hat{y}_j)] - y_j(1\{i = j\} - \hat{y}_j)$$

$$\frac{\partial L(Y, \hat{Y})}{\partial z_j} = [\sum_{i \neq j} y_i \hat{y}_j] - y_i(1 - \hat{y}_j) = [\sum_{i \neq j} y_i \hat{y}_j] - y_j + y_j \hat{y}_j$$
$$\sum_i y_i \hat{y}_j - y_j = \hat{y}_j \sum_i y_i - y_j = \hat{y}_j - y_j$$

Observe que $\sum_i y_i$ es igual a 1.

Inciso d)

Ahora encontremos $\frac{\partial L(Y,\hat{Y})}{\partial x}$:
Sabemos que $h = \sigma(W_1 x + b_1)$, $\hat{Y} = softmax(W_2 x + b_2)$ y $Z = W_2 x + b_2$.

$$\frac{\partial L(Y,\hat{Y})}{\partial x} = \frac{\partial L(Y,\hat{Y})}{\partial z}(\frac{\partial z}{\partial h})(\frac{\partial h}{\partial x}) = (\hat{y}_j - y_j)(W_2)\frac{\partial h}{\partial x}$$

Sabemos que $\frac{\partial \sigma}{\partial x} = (1 - \sigma(x)\sigma(x))$, vamos a sustituir y tendremos que $\sigma(W_1 x + b_1) = \frac{1}{1+e^{-(w1x+b_1)}}$, derivamos para finalmente sustituir, así:

$$\frac{\partial \sigma}{\partial x} = (-1)(1 + e^{-(w_1 x + b_1)})(-w_1) = (1 - \sigma(w_1 x + b_1))\sigma(w_1 x + b_1)(w_1)$$

Finalmente tenemos:

$$\begin{aligned}\frac{\partial L(Y,\hat{Y})}{\partial x} &= (\hat{y}_j - y_j)(w_2)(1 - \sigma(w_1 x + b_1))\sigma(w_1 x + b_1)(w_1) \\ &= w_1 w_2(\hat{y}_j - y_j)(1 - \sigma(w_1 x + b_1))\sigma(w_1 x + b_1)\end{aligned}$$

Las ecuaciones que nos van a servir para el paso backpropagation son las siguientes:

$$\frac{\partial L(y_j,\hat{y}_j)}{\partial w_2} = (\frac{\partial L}{\partial z})(\frac{\partial z}{\partial w_2}) = (\hat{y}_i - y_i)h$$

$$\frac{\partial L(y_i,\hat{y}_i)}{\partial b_2} = (\frac{\partial L}{\partial z})(\frac{\partial z}{\partial b_2}) = (\hat{y}_i - y_i)(1) = (\hat{y}_i - y_i)$$

$$\frac{\partial L(y_j,\hat{y}_j)}{\partial w_1} = (\frac{\partial L}{\partial z})(\frac{\partial z}{\partial h})(\frac{\partial h}{\partial w_1}) =$$

$$(\hat{y}_i - y)(w_2)[x(1 - \sigma(w_1 x + b_1))\sigma(w_1 x + b_1)]$$

$$\frac{\partial L(y_j,\hat{y}_j)}{\partial b_1} = (\frac{\partial L}{\partial z})(\frac{\partial z}{\partial h})(\frac{\partial h}{\partial b_1}) =$$

$$(\hat{y}_i - y)(w_2)[(1 - \sigma(w_1 x + b_1))\sigma(w_1 x + b_1)]$$