MAC0460 - Introdução ao aprendizado de máquina

Back-propagation 2

Felipe Salvatore
https://felipessalvatore.github.io/

Nina S. T. Hirata
https://www.ime.usp.br/~nina/

April 27, 2018

IME-USP: Institute of Mathematics and Statistics, University of São Paulo

Definição de Jacobiano

$$f(x, y) = x + y$$

$$f\left(\boldsymbol{u} = \begin{bmatrix} x \\ y \end{bmatrix}\right) = x + y$$

$$\nabla_{\boldsymbol{u}} f = \frac{\partial f}{\partial \boldsymbol{u}} = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix}$$

Convenção de shape

	$shape(\boldsymbol{x}) = 1 \times 1$	$shape(\boldsymbol{x}) = n \times 1$
$shape(\boldsymbol{f}) = 1 \times 1$	$shape(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}) = 1 \times 1$	$shape(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}) = 1 \times n$
$shape(\boldsymbol{f}) = m \times 1$	$shape(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}) = m \times 1$	$shape(\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}) = m \times n$

Discussão

Dado $\boldsymbol{u} \in \mathbb{R}^n$ e $f : \mathbb{R}^n \to \mathbb{R}$ podemos definir $\nabla_{\boldsymbol{u}} f = \frac{\partial f}{\partial \boldsymbol{u}}$ como um vetor coluna.

- (positivo) $u + \nabla_{\boldsymbol{u}} f$ faz sentido.
- (negativo) quando $\boldsymbol{x} \in \mathbb{R}^n$, $f : \mathbb{R}^k \to \mathbb{R}$, $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^k$, $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x})$ e $z = f(\boldsymbol{y})$, a regra da cadeia tem um formato menos intuitivo.

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

Operações básicas: produto vetor-escalar

$\boldsymbol{x} \in \mathbb{R}^n$ e $\alpha \in \mathbb{R}$

$$\boldsymbol{u} = \boldsymbol{x}\alpha$$

- $\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} = \underbrace{diag(\boldsymbol{1}\alpha)}_{n \times n}$
- $\frac{\partial \boldsymbol{u}}{\partial \alpha} = \underbrace{\boldsymbol{x}}_{n \times 1}$

Operações básicas: soma

$\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$\boldsymbol{u} = \boldsymbol{x} + \boldsymbol{y}$$

- $\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} = diag(1) = \underbrace{\boldsymbol{I}}_{n \times n}$
- $\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{y}} = diag(1) = \underbrace{\boldsymbol{I}}_{n \times n}$

Operações básicas: Hadamard product

$\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$\boldsymbol{u} = \boldsymbol{x} \odot \boldsymbol{y}$$

- $\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} = \underbrace{diag(\boldsymbol{y})}_{n \times n}$
- $\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{y}} = \underbrace{diag(\boldsymbol{x})}_{n \times n}$

Operações básicas: função escalar aplicada em vetor

$\boldsymbol{x} \in \mathbb{R}^n$ e $h : \mathbb{R} \to \mathbb{R}$ é uma função diferenciável.

$$\boldsymbol{u} = h(\boldsymbol{x}) = \begin{bmatrix} h(x_1) \\ h(x_2) \\ \vdots \\ h(x_n) \end{bmatrix}$$

- $\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} = \underbrace{diag(h'(\boldsymbol{x}))}_{n \times n}$ onde $h'(\boldsymbol{x}) = \begin{bmatrix} \frac{dh(x_1)}{dx_1} \\ \frac{dh(x_2)}{dx_2} \\ \vdots \\ \frac{dh(x_n)}{dx_n} \end{bmatrix}$

Operações básicas: redução por soma

$\boldsymbol{x} \in \mathbb{R}^n$

$$u = sum(\boldsymbol{x}) = \sum_{i=1}^{n} x_i$$

- $\frac{\partial u}{\partial \boldsymbol{x}} = \underbrace{\mathbf{1}^\top}_{1 \times n}$

Operações básicas: multiplicação matriz-vetor

$\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{W} \in \mathbb{R}^{d \times n}$

$$\boldsymbol{u} = \boldsymbol{W}\boldsymbol{x}$$

- $\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} = \underbrace{\boldsymbol{W}}_{d \times n}$
- $\underbrace{\dfrac{\partial \boldsymbol{u}}{\partial \boldsymbol{W}}}_{d \times d \times n}$ tal que $\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{W}}_{i,j,k} = \begin{cases} 0, \text{ se } i \neq j \\ x_k, \text{ se } i = j \end{cases}$

Revisão: regra da cadeia

Dados $x, u_1(x), \ldots, u_n(x) \in \mathbb{R}$ e $f : \mathbb{R}^n \to \mathbb{R}$ temos que cada u_i varia dado uma variação em x. Assim a regra da cadeia para várias variáveis é definida como:

$$\frac{\partial f(u_1, \ldots, u_n)}{\partial x} = \frac{\partial f(u_1, \ldots, u_n)}{\partial u_1}\frac{\partial u_1}{\partial x} + \cdots + \frac{\partial f(u_1, \ldots, u_n)}{\partial u_n}\frac{\partial u_n}{\partial x}$$

$$= \sum_{i=1}^{n} \frac{\partial f(u_1, \ldots, u_n)}{\partial u_i}\frac{\partial u_i}{\partial x}$$

Regra da cadeia, caso vetorial

$\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{f} : \mathbb{R}^k \to \mathbb{R}^m$ e $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^k$.

$$\boldsymbol{g}(\boldsymbol{x}) = \begin{bmatrix} g_1(x_1, \dots, x_n) \\ g_2(x_1, \dots, x_n) \\ \vdots \\ g_k(x_1, \dots, x_n) \end{bmatrix} \qquad \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{x})) = \begin{bmatrix} f_1(g_1, \dots, g_k) \\ f_2(g_1, \dots, g_k) \\ \vdots \\ f_m(g_1, \dots, g_k) \end{bmatrix}$$

Regra da cadeia, caso vetorial

$$
\begin{aligned}
\left.\frac{\partial \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{x}))}{\partial \boldsymbol{x}}\right|_{i,j} &= \frac{\partial f_i}{\partial x_j} \\
&= \frac{\partial f_i(g_1, \ldots, g_k)}{\partial x_j} \\
&= \sum_{s=1}^{k} \frac{\partial f_i(g_1, \ldots, g_k)}{\partial g_s} \frac{\partial g_s}{\partial x_j} \\
&= \left.\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{g}} \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}\right|_{i,j}
\end{aligned}
$$

Regra da cadeia, caso vetorial

$$\underbrace{\frac{\partial \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{x}))}{\partial \boldsymbol{x}}}_{m \times n} = \underbrace{\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{g}}}_{m \times k} \underbrace{\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}}_{k \times n}$$

Regra da cadeia, caso vetorial

Dados $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{u}^{(1)}(\boldsymbol{x}), \ldots, \boldsymbol{u}^{(s)}(\boldsymbol{x}) \in \mathbb{R}^k$ e $\boldsymbol{f} : \mathbb{R}^{s \times k} \to \mathbb{R}^m$:

$$\frac{\partial \boldsymbol{f}\left(\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(s)}\right)}{\partial \boldsymbol{x}} = \sum_{i=1}^{s} \frac{\partial \boldsymbol{f}\left(\boldsymbol{u}^{(1)}, \ldots, \boldsymbol{u}^{(s)}\right)}{\partial \boldsymbol{u}^{(i)}} \frac{\partial \boldsymbol{u}^{(i)}}{\partial \boldsymbol{x}}$$

Novas operações: subtração

$\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$\boldsymbol{u} = \boldsymbol{x} - \boldsymbol{y}$$

$\boldsymbol{u} = \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{y})) = \boldsymbol{x} + \boldsymbol{g}(\boldsymbol{y})$

$\boldsymbol{g}(\boldsymbol{y}) = -\boldsymbol{y}$

- $\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} = \underbrace{\boldsymbol{I}}_{n \times n}$
- $\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{y}} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{g}} \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{y}} = \underbrace{\boldsymbol{I}}_{n \times n} (\underbrace{-\boldsymbol{I}}_{n \times n}) = \underbrace{-\boldsymbol{I}}_{n \times n}$

Novas operações: produto escalar

$\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$u = \boldsymbol{x}^\top \boldsymbol{y}$$

$u = \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y})) = sum(\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}))$

$\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} \odot \boldsymbol{y}$

- $\frac{\partial u}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{g}} \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}} = \underbrace{\boldsymbol{1}^\top}_{1 \times n} \underbrace{diag(\boldsymbol{y})}_{n \times n} = \underbrace{\boldsymbol{y}^\top}_{1 \times n}$
- $\frac{\partial u}{\partial \boldsymbol{y}} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{g}} \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{y}} = \underbrace{\boldsymbol{1}^\top}_{1 \times n} \underbrace{diag(\boldsymbol{x})}_{n \times n} = \underbrace{\boldsymbol{x}^\top}_{1 \times n}$

Exemplo 1: regressão linear

$\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^n$ e $y \in \mathbb{R}$

$$L = f(g(\hat{y}))$$

$\hat{y} = \boldsymbol{w}^\top \boldsymbol{x}$

$g(\hat{y}) = \hat{y} - y$

$f(g(\hat{y})) = g(\hat{y})^2$

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{w}} &= \frac{\partial f}{\partial g} \frac{\partial g}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \boldsymbol{w}} \\
&= \underbrace{2(\hat{y} - y)}_{1\times 1} \underbrace{1}_{1\times 1} \underbrace{\boldsymbol{x}^\top}_{1\times n} \\
&= \underbrace{2(\hat{y} - y)\boldsymbol{x}^\top}_{1\times n}
\end{aligned}$$

Novas operações: transformação afim

$\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{W} \in \mathbb{R}^{d \times n}$, $\boldsymbol{b} \in \mathbb{R}^d$

$$\boldsymbol{u} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$$

$\boldsymbol{u} = \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{W}, \boldsymbol{x})) = \boldsymbol{g}(\boldsymbol{W}, \boldsymbol{x}) + \boldsymbol{b}$

$\boldsymbol{g}(\boldsymbol{W}, \boldsymbol{x}) = \boldsymbol{W}\boldsymbol{x}$

- $\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{g}} \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}} = \underbrace{\boldsymbol{I}}_{d \times d} \underbrace{\boldsymbol{W}}_{d \times n} = \underbrace{\boldsymbol{W}}_{d \times n}$
- $\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{b}} = \underbrace{\boldsymbol{I}}_{d \times d}$
- $\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{W}} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{g}} \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{W}} = \underbrace{\boldsymbol{I}}_{d \times d} \underbrace{\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{W}}}_{d \times d \times n} = \underbrace{\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{W}}}_{d \times d \times n}$

Novas operações: ativação

$\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{W} \in \mathbb{R}^{d \times n}$, $\boldsymbol{b} \in \mathbb{R}^d$, $h : \mathbb{R} \to \mathbb{R}$

$$\boldsymbol{u} = h(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$$

$\boldsymbol{u} = \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{W}, \boldsymbol{x}, \boldsymbol{b})) = h(\boldsymbol{g}(\boldsymbol{W}, \boldsymbol{x}, \boldsymbol{b}))$

$\boldsymbol{g}(\boldsymbol{W}, \boldsymbol{x}, \boldsymbol{b}) = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$

- $$\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{W}} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{g}} \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{W}}$$

$$= \underbrace{diag(h'(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}))}_{d \times d} \underbrace{\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{W}}}_{d \times d \times n}$$

Novas operações: ativação

$$\underbrace{\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{W}}}_{d \times d \times n} \text{ tal que } \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{W}}_{i,j,k} = \begin{cases} 0, \text{ se } i \neq j \\ h'(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})_i x_k, \text{ se } i = j \end{cases}$$

- $$\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{b}} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{g}} \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{b}}$$

$$= \underbrace{diag(h'(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}))}_{d \times d} \underbrace{\boldsymbol{I}}_{d \times d}$$

$$= \underbrace{diag(h'(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}))}_{d \times d}$$

Novas operações: ativação

- $$\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{g}} \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{x}}$$

$$= \underbrace{diag(h'(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}))}_{d \times d} \underbrace{\boldsymbol{W}}_{d \times n}$$

$$= \begin{bmatrix} h'(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})_1 w_{1,1} & \dots & h'(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})_1 w_{1,n} \\ \vdots & \ddots & \vdots \\ h'(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})_d w_{d,1} & \dots & h'(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})_d w_{d,n} \end{bmatrix}_{d \times n}$$

Novas operações: softmax

$\boldsymbol{x} \in \mathbb{R}^n$

$$\boldsymbol{u} = \boldsymbol{s}(\boldsymbol{x})$$

$\boldsymbol{f}(\boldsymbol{x}) = exp(\boldsymbol{x})$

$g(\boldsymbol{f}) = sum(\boldsymbol{f})$

$h(g) = g^{-1}$

$\boldsymbol{t}(\boldsymbol{f}, h) = \boldsymbol{f} h$

$\boldsymbol{u} = \boldsymbol{s}(\boldsymbol{x}) = \boldsymbol{t}(\boldsymbol{f}, h)$

- $$\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{t}}{\partial \boldsymbol{f}} \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} + \frac{\partial \boldsymbol{t}}{\partial h} \frac{\partial h}{\partial \boldsymbol{x}}$$

Novas operações: softmax

$$\frac{\partial \boldsymbol{t}}{\partial \boldsymbol{f}}\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{t}}{\partial \boldsymbol{f}}\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$$

$$= \underbrace{diag(\mathbf{1}h)}_{n\times n}\underbrace{diag(\boldsymbol{f})}_{n\times n}$$

$$= \underbrace{diag(\boldsymbol{f}h)}_{n\times n}$$

$$= \underbrace{diag(\boldsymbol{s}(\boldsymbol{x}))}_{n\times n}$$

Novas operações: softmax

$$\frac{\partial \boldsymbol{t}}{\partial h}\frac{\partial h}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{t}}{\partial h}\frac{\partial h}{\partial g}\frac{\partial g}{\partial \boldsymbol{f}}\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}}$$

$$= (\underbrace{\boldsymbol{f}}_{n\times 1}\underbrace{-\frac{1}{g^2}}_{1\times 1})\underbrace{\mathbf{1}^\top}_{1\times n}\underbrace{diag(\boldsymbol{f})}_{n\times n}$$

$$= \underbrace{(\boldsymbol{f} - \frac{1}{g^2})}_{n\times 1}\underbrace{\boldsymbol{f}^\top}_{1\times n}$$

$$= \begin{bmatrix} -\boldsymbol{s}_1^2 & -\boldsymbol{s}_1\boldsymbol{s}_2 & \dots & -\boldsymbol{s}_1\boldsymbol{s}_n \\ \vdots & \vdots & \ddots & \vdots \\ -\boldsymbol{s}_1\boldsymbol{s}_n & -\boldsymbol{s}_n\boldsymbol{s}_2 & \dots & -\boldsymbol{s}_n^2 \end{bmatrix}_{n\times n}$$

Novas operações: softmax

$$\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} = diag(\boldsymbol{s}(\boldsymbol{x})) + \frac{\partial \boldsymbol{t}}{\partial h}\frac{\partial h}{\partial \boldsymbol{x}}$$

$$= \begin{bmatrix} \boldsymbol{s}_1(1-\boldsymbol{s}_1) & -\boldsymbol{s}_1\boldsymbol{s}_2 & \dots & -\boldsymbol{s}_1\boldsymbol{s}_n \\ -\boldsymbol{s}_2\boldsymbol{s}_1 & \boldsymbol{s}_2(1-\boldsymbol{s}_2) & \dots & -\boldsymbol{s}_2\boldsymbol{s}_n \\ \vdots & \vdots & \ddots & \vdots \\ -\boldsymbol{s}_n\boldsymbol{s}_1 & -\boldsymbol{s}_n\boldsymbol{s}_2 & \dots & \boldsymbol{s}_n(1-\boldsymbol{s}_n) \end{bmatrix}_{n\times n}$$

Novas operações: softmax com entropia cruzada

$\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$

$$u = L_{SCE}(\boldsymbol{x}, \boldsymbol{y})$$

$\hat{\boldsymbol{y}}(\boldsymbol{x}) = \boldsymbol{s}(\boldsymbol{x})$

$\boldsymbol{f}(\hat{\boldsymbol{y}}) = log(\hat{\boldsymbol{y}})$

$\boldsymbol{g}(\boldsymbol{f}) = \boldsymbol{y} \odot \boldsymbol{f}$

$h(\boldsymbol{g}) = sum(\boldsymbol{g})$

$t(h) = -h$

$u = L_{SCE}(\boldsymbol{x}, \boldsymbol{y}) = t$

- $\dfrac{\partial u}{\partial \boldsymbol{x}} = \dfrac{\partial t}{\partial h} \dfrac{\partial h}{\partial \boldsymbol{g}} \dfrac{\partial \boldsymbol{g}}{\partial \boldsymbol{f}} \dfrac{\partial \boldsymbol{f}}{\partial \hat{\boldsymbol{y}}} \dfrac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{x}}$

Novas operações: softmax com entropia cruzada

$$\frac{\partial u}{\partial \boldsymbol{x}} = \frac{\partial t}{\partial h}\frac{\partial h}{\partial \boldsymbol{g}}\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{f}}\frac{\partial \boldsymbol{f}}{\partial \hat{\boldsymbol{y}}}\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{x}}$$

$$= \underbrace{-1}_{1\times 1}\underbrace{\mathbf{1}^\top}_{1\times n}\underbrace{diag(\boldsymbol{y})}_{n\times n}\underbrace{diag(\hat{\boldsymbol{y}}^{-1})}_{n\times n}\underbrace{\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{x}}}_{n\times n}$$

$$= \underbrace{-\mathbf{1}^\top}_{1\times n}\underbrace{diag(\frac{\boldsymbol{y}}{\hat{\boldsymbol{y}}})}_{n\times n}\underbrace{\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{x}}}_{n\times n}$$

$$= \underbrace{-\frac{\boldsymbol{y}}{\hat{\boldsymbol{y}}}^\top}_{1\times n}\underbrace{\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{x}}}_{n\times n}$$

Novas operações: softmax com entropia cruzada

$$\frac{\partial u}{\partial \boldsymbol{x}_i} = -\frac{\boldsymbol{y}}{\hat{\boldsymbol{y}}}^\top \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{x}_{:,i}}$$

$$= \frac{y_1}{\hat{y}_1}\hat{y}_i\hat{y}_1 + \frac{y_2}{\hat{y}_2}\hat{y}_i\hat{y}_2 + \cdots - \frac{y_i}{\hat{y}_i}\hat{y}_i(1-\hat{y}_i) + \cdots + \frac{y_n}{\hat{y}_n}\hat{y}_i\hat{y}_n$$

$$= y_1\hat{y}_i + y_2\hat{y}_i + \cdots + (y_i\hat{y}_i - y_i) + \cdots + y_n\hat{y}_i$$

$$= \sum_{j=1}^{n} y_j\hat{y}_i - y_i$$

$$= \hat{y}_i - y_i \text{ (quando } sum(\boldsymbol{y}) = 1)$$

Assim, daqui em diante $\frac{\partial u}{\partial \boldsymbol{x}} = (\hat{\boldsymbol{y}} - \boldsymbol{y})^\top$

Exemplo 2: Regressão logística

$\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{W} \in \mathbb{R}^{d \times n}$ e $\boldsymbol{b}, \boldsymbol{y} \in \mathbb{R}^d$

$$L = L_{SCE}(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{y})$$

$\boldsymbol{u} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}$

$\hat{\boldsymbol{y}} = \boldsymbol{s}(\boldsymbol{u})$

- $\frac{\partial L}{\partial \boldsymbol{b}} = \frac{\partial L_{SCE}}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{b}} = \underbrace{(\hat{\boldsymbol{y}} - \boldsymbol{y})^\top}_{1 \times d} \underbrace{\boldsymbol{I}}_{d \times d} = \underbrace{(\hat{\boldsymbol{y}} - \boldsymbol{y})^\top}_{1 \times d}$

Exemplo 2: Regressão logística

- $$\frac{\partial L}{\partial \boldsymbol{W}} = \frac{\partial L_{SCE}}{\partial \boldsymbol{u}} \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{W}}$$

$$= \underbrace{(\hat{\boldsymbol{y}} - \boldsymbol{y})^\top}_{1 \times d} \underbrace{\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{W}}}_{d \times d \times n}$$

$$= \begin{bmatrix} (\hat{y}_1 - y_1)x_1 & (\hat{y}_1 - y_1)x_2 & \dots & (\hat{y}_1 - y_1)x_n \\ (\hat{y}_2 - y_2)x_1 & (\hat{y}_2 - y_2)x_2 & \dots & (\hat{y}_2 - y_2)x_n \\ \vdots & \vdots & \ddots & \vdots \\ (\hat{y}_d - y_d)x_1 & (\hat{y}_d - y_d)x_2 & \dots & (\hat{y}_d - y_d)x_n \end{bmatrix}_{1 \times d \times n}$$

Exemplo 2: Regressão logística

- $\dfrac{\partial L}{\partial \boldsymbol{x}} = \dfrac{\partial L_{SCE}}{\partial \boldsymbol{u}} \dfrac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}}$

$$= \underbrace{(\hat{\boldsymbol{y}} - \boldsymbol{y})^\top}_{1 \times d} \underbrace{\boldsymbol{W}}_{d \times n}$$

$$= \left[\sum_{j=1}^{d} (\hat{y}_j - y_j) w_{j,1}, \quad \ldots, \quad \sum_{j=1}^{d} (\hat{y}_j - y_j) w_{j,1} \right]_{1 \times n}$$

Exemplo 3: Rede neural com uma camada escondida

$\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{W}^{(1)} \in \mathbb{R}^{k \times n}$ e $\boldsymbol{b}^{(1)} \in \mathbb{R}^k$, $\boldsymbol{W}^{(2)} \in \mathbb{R}^{d \times k}$ $\boldsymbol{b}^{(2)} \in \mathbb{R}^d$, $\boldsymbol{y} \in \mathbb{R}^d$

$$\boldsymbol{h}^{(1)} = g\big(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}\big)$$

$$\boldsymbol{h}^{(2)} = \boldsymbol{W}^{(2)}\boldsymbol{h}^{(1)} + \boldsymbol{b}^{(2)}$$

$$\hat{\boldsymbol{y}} = \boldsymbol{s}\big(\boldsymbol{h}^{(2)}\big)$$

$$L = L_{SCE}\big(\boldsymbol{h}^{(2)}, \boldsymbol{y}\big)$$

Exemplo 3: Rede neural com uma camada escondida

$$\boldsymbol{\delta} \leftarrow \frac{\partial L_{SCE}}{\partial \boldsymbol{h}^{(2)}}$$

- $\frac{\partial L}{\partial \boldsymbol{b}^{(2)}} = \frac{\partial L_{SCE}}{\partial \boldsymbol{h}^{(2)}} \frac{\partial \boldsymbol{h}^{(2)}}{\partial \boldsymbol{b}^{(2)}} = \boldsymbol{\delta} \frac{\partial \boldsymbol{h}^{(2)}}{\partial \boldsymbol{b}^{(2)}} = \underbrace{(\hat{\boldsymbol{y}} - \boldsymbol{y})}_{1\times d}^{\top} \underbrace{\boldsymbol{I}}_{d\times d} = \underbrace{(\hat{\boldsymbol{y}} - \boldsymbol{y})^{\top}}_{1\times d}$

Exemplo 3: Rede neural com uma camada escondida

- $\dfrac{\partial L}{\partial \boldsymbol{W}^{(2)}} = \dfrac{\partial L_{SCE}}{\partial \boldsymbol{h}^{(2)}} \dfrac{\partial \boldsymbol{h}^{(2)}}{\partial \boldsymbol{W}^{(2)}}$

$$= \underbrace{\boldsymbol{\delta}}_{1\times d} \underbrace{\frac{\partial \boldsymbol{h}^{(2)}}{\partial \boldsymbol{W}^{(2)}}}_{d\times d\times k}$$

$$= \begin{bmatrix} (\hat{y}_1 - y_1)h_1^1 & (\hat{y}_1 - y_1)h_2^1 & \dots & (\hat{y}_1 - y_1)h_k^1 \\ (\hat{y}_2 - y_2)h_1^1 & (\hat{y}_2 - y_2)h_2^1 & \dots & (\hat{y}_2 - y_2)h_k^1 \\ \vdots & \vdots & \ddots & \vdots \\ (\hat{y}_d - y_d)h_1^1 & (\hat{y}_d - y_d)h_2^1 & \dots & (\hat{y}_d - y_d)h_k^1 \end{bmatrix}_{1\times d\times k}$$

Exemplo 3: Rede neural com uma camada escondida

$$\boldsymbol{\delta} \leftarrow \frac{\partial L_{SCE}}{\partial \boldsymbol{h}^{(2)}} \frac{\partial \boldsymbol{h}^{(2)}}{\partial \boldsymbol{h}^{(1)}}$$

$$= \boldsymbol{\delta} \frac{\partial \boldsymbol{h}^{(2)}}{\partial \boldsymbol{h}^{(1)}}$$

$$= (\hat{\boldsymbol{y}} - \boldsymbol{y})^\top \boldsymbol{W}^{(2)}$$

Exemplo 3: Rede neural com uma camada escondida

- $$\frac{\partial L}{\partial \boldsymbol{b}^{(1)}} = \frac{\partial L_{SCE}}{\partial \boldsymbol{h}^{(2)}} \frac{\partial \boldsymbol{h}^{(2)}}{\partial \boldsymbol{h}^{(1)}} \frac{\partial \boldsymbol{h}^{(1)}}{\partial \boldsymbol{b}^{(1)}}$$

$$= \boldsymbol{\delta} \frac{\partial \boldsymbol{h}^{(1)}}{\partial \boldsymbol{b}^{(1)}}$$

$$= \underbrace{(\hat{\boldsymbol{y}} - \boldsymbol{y})^\top}_{1 \times d} \underbrace{\boldsymbol{W}^{(2)}}_{d \times k} \underbrace{diag(g'(\boldsymbol{W}^{(1)} \boldsymbol{x} + \boldsymbol{b}^{(1)}))}_{k \times k}$$

Exemplo 3: Rede neural com uma camada escondida

- $$\frac{\partial L}{\partial \boldsymbol{W}^{(1)}} = \frac{\partial L_{SCE}}{\partial \boldsymbol{h}^{(2)}} \frac{\partial \boldsymbol{h}^{(2)}}{\partial \boldsymbol{h}^{(1)}} \frac{\partial \boldsymbol{h}^{(1)}}{\partial \boldsymbol{W}^{(1)}}$$

$$= \boldsymbol{\delta} \frac{\partial \boldsymbol{h}^{(1)}}{\partial \boldsymbol{W}^{(1)}}$$

$$= \underbrace{(\hat{\boldsymbol{y}} - \boldsymbol{y})^\top}_{1 \times d} \underbrace{\boldsymbol{W}^{(2)}}_{d \times k} \underbrace{\frac{\partial \boldsymbol{h}^{(1)}}{\partial \boldsymbol{W}^{(1)}}}_{k \times k \times n}$$

Algoritmo de back-propagation (para redes neurais)

Algorithm 1 Back-propagation for a deep neural network

Require: K, network depth
Require: $\boldsymbol{x}$, the input to process
Require: $\boldsymbol{y}$, the target output
Require: $L_{out}(\cdot, \cdot)$, output with cost function
Require: $\boldsymbol{h}^{(i)} = g^{(i)}(\boldsymbol{W}^{(i)}\boldsymbol{h}^{(i-1)} + \boldsymbol{b}^{(i)})$, $i \in \{1, \ldots, K\}$, activation function for the i-th layer where $\boldsymbol{h}^{(0)} = \boldsymbol{x}$
$L \leftarrow L_{out}(\boldsymbol{h}^{(K)}, \boldsymbol{y})$
$\boldsymbol{\delta} \leftarrow \frac{\partial L}{\partial \boldsymbol{h}^{(K)}}$
for $i = K$ down to 1 **do**
    $\frac{\partial L}{\partial \boldsymbol{b}^{(i)}} \leftarrow \boldsymbol{\delta} \frac{\partial \boldsymbol{h}^{(i)}}{\partial \boldsymbol{b}^{(i)}}$
    $\frac{\partial L}{\partial \boldsymbol{W}^{(i)}} \leftarrow \boldsymbol{\delta} \frac{\partial \boldsymbol{h}^{(i)}}{\partial \boldsymbol{W}^{(i)}}$
    $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta} \frac{\partial \boldsymbol{h}^{(i)}}{\partial \boldsymbol{h}^{(i-1)}}$
end for

Referências I

- Computational Graphs, and Backpropagation (course notes for nlp by michael collins).
  http://www.cs.columbia.edu/~mcollins/ff2.pdf.
- Vector Calculus (in mathematics for machine learning).
  https://mml-book.github.io/book/chapter05.pdf.
- I. Goodfellow, Y. Bengio, and A. Courville.
  ***Deep Learning*.**
  MIT Press, 2017.
- T. Parr and J. Howard.
  The matrix calculus you need for deep learning.
  CoRR, abs/1802.01528, 2018.