

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 30, 2018

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

Definição de Jacobiano

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

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

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

Convenção de shape

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

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

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

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

Operações básicas: produto vetor-escalar

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

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

$$\bullet \frac{\partial \mathbf{u}}{\partial \mathbf{x}} = \underbrace{\text{diag}(\mathbf{1}\alpha)}_{n \times n}$$

$$\bullet \frac{\partial \mathbf{u}}{\partial \alpha} = \underbrace{\mathbf{x}}_{n \times 1}$$

Operações básicas: soma

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

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

$$\bullet \frac{\partial \mathbf{u}}{\partial \mathbf{x}} = \text{diag}(\mathbf{1}) = \underbrace{\mathbf{I}}_{n \times n}$$

$$\bullet \frac{\partial \mathbf{u}}{\partial \mathbf{y}} = \text{diag}(\mathbf{1}) = \underbrace{\mathbf{I}}_{n \times n}$$

Operações básicas: Hadamard product

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

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

$$\bullet \frac{\partial \mathbf{u}}{\partial \mathbf{x}} = \underbrace{\text{diag}(\mathbf{y})}_{n \times n}$$

$$\bullet \frac{\partial \mathbf{u}}{\partial \mathbf{y}} = \underbrace{\text{diag}(\mathbf{x})}_{n \times n}$$

Operações básicas: função escalar aplicada em vetor

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

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

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

Operações básicas: redução por soma

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

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

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

Operações básicas: multiplicação matriz-vetor

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

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

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

Revisão: regra da cadeia

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

$$\begin{aligned}\frac{\partial f(u_1, \dots, u_n)}{\partial x} &= \frac{\partial f(u_1, \dots, u_n)}{\partial u_1} \frac{\partial u_1}{\partial x} + \dots + \frac{\partial f(u_1, \dots, u_n)}{\partial u_n} \frac{\partial u_n}{\partial x} \\ &= \sum_{i=1}^n \frac{\partial f(u_1, \dots, u_n)}{\partial u_i} \frac{\partial u_i}{\partial x}\end{aligned}$$

Regra da cadeia, caso vetorial

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

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

Regra da cadeia, caso vetorial

$$\begin{aligned}\frac{\partial \mathbf{f}(\mathbf{g}(\mathbf{x}))}{\partial \mathbf{x}}_{i,j} &= \frac{\partial f_i}{\partial x_j} \\&= \frac{\partial f_i(g_1, \dots, g_k)}{\partial x_j} \\&= \sum_{s=1}^k \frac{\partial f_i(g_1, \dots, g_k)}{\partial g_s} \frac{\partial g_s}{\partial x_j} \\&= \frac{\partial \mathbf{f}}{\partial \mathbf{g}} \frac{\partial \mathbf{g}}{\partial \mathbf{x}}_{i,j}\end{aligned}$$

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

Regra da cadeia, caso vetorial

Dados $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{u}^{(1)}(\mathbf{x}), \dots, \mathbf{u}^{(s)}(\mathbf{x}) \in \mathbb{R}^k$ e
 $\mathbf{f} : \mathbb{R}^{s \times k} \rightarrow \mathbb{R}^m$:

$$\frac{\partial \mathbf{f}(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(s)})}{\partial \mathbf{x}} = \sum_{i=1}^s \frac{\partial \mathbf{f}(\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(s)})}{\partial \mathbf{u}^{(i)}} \frac{\partial \mathbf{u}^{(i)}}{\partial \mathbf{x}}$$

Novas operações: subtração

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

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

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

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

$$\bullet \quad \frac{\partial \mathbf{u}}{\partial \mathbf{x}} = \underbrace{\mathbf{I}}_{n \times n}$$

$$\bullet \quad \frac{\partial \mathbf{u}}{\partial \mathbf{y}} = \frac{\partial \mathbf{f}}{\partial \mathbf{g}} \frac{\partial \mathbf{g}}{\partial \mathbf{y}} = \underbrace{\mathbf{I}}_{n \times n} \underbrace{(-\mathbf{I})}_{n \times n} = \underbrace{-\mathbf{I}}_{n \times n}$$

Novas operações: produto escalar

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

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

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

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

$$\begin{aligned} \bullet \quad \frac{\partial u}{\partial \mathbf{x}} &= \frac{\partial \mathbf{f}}{\partial \mathbf{g}} \frac{\partial \mathbf{g}}{\partial \mathbf{x}} = \underbrace{\mathbf{1}^\top}_{1 \times n} \underbrace{\text{diag}(\mathbf{y})}_{n \times n} = \underbrace{\mathbf{y}^\top}_{1 \times n} \\ \bullet \quad \frac{\partial u}{\partial \mathbf{y}} &= \frac{\partial \mathbf{f}}{\partial \mathbf{g}} \frac{\partial \mathbf{g}}{\partial \mathbf{y}} = \underbrace{\mathbf{1}^\top}_{1 \times n} \underbrace{\text{diag}(\mathbf{x})}_{n \times n} = \underbrace{\mathbf{x}^\top}_{1 \times n} \end{aligned}$$

Exemplo 1: regressão linear

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

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

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

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

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

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

Novas operações: transformação afim

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

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

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

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

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

Novas operações: ativação

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

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

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

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

$$\bullet \quad \frac{\partial \mathbf{u}}{\partial \mathbf{W}} = \frac{\partial \mathbf{f}}{\partial \mathbf{g}} \frac{\partial \mathbf{g}}{\partial \mathbf{W}}$$

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

Novas operações: ativação

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

$$\bullet \frac{\partial \mathbf{u}}{\partial \mathbf{b}} = \frac{\partial \mathbf{f}}{\partial \mathbf{g}} \frac{\partial \mathbf{g}}{\partial \mathbf{b}}$$

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

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

Novas operações: ativação

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

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

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

Novas operações: softmax

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

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

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

$$g(\mathbf{f}) = \text{sum}(\mathbf{f})$$

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

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

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

$$\bullet \quad \frac{\partial \mathbf{u}}{\partial \mathbf{x}} = \frac{\partial \mathbf{t}}{\partial \mathbf{f}} \frac{\partial \mathbf{f}}{\partial \mathbf{x}} + \frac{\partial \mathbf{t}}{\partial h} \frac{\partial h}{\partial \mathbf{x}}$$

Novas operações: softmax

$$\begin{aligned}\frac{\partial \mathbf{t}}{\partial \mathbf{f}} \frac{\partial \mathbf{f}}{\partial \mathbf{x}} &= \frac{\partial \mathbf{t}}{\partial \mathbf{f}} \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \\&= \underbrace{\text{diag}(\mathbf{1}h)}_{n \times n} \underbrace{\text{diag}(\mathbf{f})}_{n \times n} \\&= \underbrace{\text{diag}(\mathbf{f}h)}_{n \times n} \\&= \underbrace{\text{diag}(\mathbf{s}(\mathbf{x}))}_{n \times n}\end{aligned}$$

Novas operações: softmax

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

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

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

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

Novas operações: softmax

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

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

Novas operações: softmax com entropia cruzada

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

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

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

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

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

$$\mathbf{h}(\mathbf{g}) = \text{sum}(\mathbf{g})$$

$$\mathbf{t}(\mathbf{h}) = -\mathbf{h}$$

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

$$\bullet \quad \frac{\partial u}{\partial \mathbf{x}} = \frac{\partial \mathbf{t}}{\partial \mathbf{h}} \frac{\partial \mathbf{h}}{\partial \mathbf{g}} \frac{\partial \mathbf{g}}{\partial \mathbf{f}} \frac{\partial \mathbf{f}}{\partial \hat{\mathbf{y}}} \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{x}}$$

Novas operações: softmax com entropia cruzada

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

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

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

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

Novas operações: softmax com entropia cruzada

$$\begin{aligned}\frac{\partial u}{\partial \mathbf{x}_i} &= -\frac{\mathbf{y}^\top}{\hat{\mathbf{y}}} \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{x}_{:,i}} \\&= \frac{y_1}{\hat{y}_1} \hat{y}_i \hat{y}_1 + \frac{y_2}{\hat{y}_2} \hat{y}_i \hat{y}_2 + \cdots - \frac{y_i}{\hat{y}_i} \hat{y}_i (1 - \hat{y}_i) + \cdots + \frac{y_n}{\hat{y}_n} \hat{y}_i \hat{y}_n \\&= y_1 \hat{y}_i + y_2 \hat{y}_i + \cdots + (y_i \hat{y}_i - y_i) + \cdots + y_n \hat{y}_i \\&= \sum_{j=1}^n y_j \hat{y}_i - y_i \\&= \hat{y}_i - y_i \text{ (quando } \text{sum}(\mathbf{y}) = 1\text{)}\end{aligned}$$

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

Exemplo 2: Regressão logística

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

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

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

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

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

Exemplo 2: Regressão logística

$$\bullet \frac{\partial L}{\partial \mathbf{W}} = \frac{\partial L_{SCE}}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial \mathbf{W}}$$

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

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

Exemplo 2: Regressão logística

$$\bullet \quad \frac{\partial L}{\partial \mathbf{x}} = \frac{\partial L_{SCE}}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$$

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

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

Exemplo 3: Rede neural com uma camada escondida

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

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

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

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

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

Exemplo 3: Rede neural com uma camada escondida

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

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

Exemplo 3: Rede neural com uma camada escondida

$$\bullet \quad \frac{\partial L}{\partial \mathbf{W}^{(2)}} = \frac{\partial L_{SCE}}{\partial \mathbf{h}^{(2)}} \frac{\partial \mathbf{h}^{(2)}}{\partial \mathbf{W}^{(2)}}$$

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

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

Exemplo 3: Rede neural com uma camada escondida

$$\begin{aligned}\delta &\leftarrow \frac{\partial L_{SCE}}{\partial \mathbf{h}^{(2)}} \frac{\partial \mathbf{h}^{(2)}}{\partial \mathbf{h}^{(1)}} \\&= \delta \frac{\partial \mathbf{h}^{(2)}}{\partial \mathbf{h}^{(1)}} \\&= (\hat{\mathbf{y}} - \mathbf{y})^\top \mathbf{W}^{(2)}\end{aligned}$$

Exemplo 3: Rede neural com uma camada escondida

- $$\begin{aligned}\frac{\partial L}{\partial \mathbf{b}^{(1)}} &= \frac{\partial L_{SCE}}{\partial \mathbf{h}^{(2)}} \frac{\partial \mathbf{h}^{(2)}}{\partial \mathbf{h}^{(1)}} \frac{\partial \mathbf{h}^{(1)}}{\partial \mathbf{b}^{(1)}} \\ &= \delta \frac{\partial \mathbf{h}^{(1)}}{\partial \mathbf{b}^{(1)}} \\ &= \underbrace{(\hat{\mathbf{y}} - \mathbf{y})^\top}_{1 \times d} \underbrace{\mathbf{W}^{(2)}}_{d \times k} \underbrace{\text{diag}(g'(\mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)}))}_{k \times k}\end{aligned}$$

Exemplo 3: Rede neural com uma camada escondida

$$\begin{aligned} \bullet \quad \frac{\partial L}{\partial \mathbf{W}^{(1)}} &= \frac{\partial L_{SCE}}{\partial \mathbf{h}^{(2)}} \frac{\partial \mathbf{h}^{(2)}}{\partial \mathbf{h}^{(1)}} \frac{\partial \mathbf{h}^{(1)}}{\partial \mathbf{W}^{(1)}} \\ &= \delta \frac{\partial \mathbf{h}^{(1)}}{\partial \mathbf{W}^{(1)}} \\ &= \underbrace{(\hat{\mathbf{y}} - \mathbf{y})^\top}_{1 \times d} \underbrace{\mathbf{W}^{(2)}}_{d \times k} \underbrace{\frac{\partial \mathbf{h}^{(1)}}{\partial \mathbf{W}^{(1)}}}_{k \times k \times n} \end{aligned}$$

Algoritmo de back-propagation (para redes neurais)

Algorithm 1 Back-propagation for a deep neural network

- 1: **Require:** K , network depth
 - 2: **Require:** \mathbf{x} , the input to process
 - 3: **Require:** \mathbf{y} , the target output
 - 4: **Require:** $L_{out}(\cdot, \cdot)$, output with cost function
 - 5: **Require:** $\mathbf{h}^{(i)} = g^{(i)}(\mathbf{W}^{(i)}\mathbf{h}^{(i-1)} + \mathbf{b}^{(i)})$, $i \in \{1, \dots, K\}$, activation function for the i -th layer where $\mathbf{h}^{(0)} = \mathbf{x}$ and $g^{(K)}$ is the identity function.
 - 6: $L \leftarrow L_{out}(\mathbf{h}^{(K)}, \mathbf{y})$
 - 7: $\delta \leftarrow \frac{\partial L}{\partial \mathbf{h}^{(K)}}$
 - 8: **for** $i = K$ down to 1 **do**
 - 9: $\frac{\partial L}{\partial \mathbf{b}^{(i)}} \leftarrow \delta \frac{\partial \mathbf{h}^{(i)}}{\partial \mathbf{b}^{(i)}}$
 - 10: $\frac{\partial L}{\partial \mathbf{W}^{(i)}} \leftarrow \delta \frac{\partial \mathbf{h}^{(i)}}{\partial \mathbf{W}^{(i)}}$
 - 11: $\delta \leftarrow \delta \frac{\partial \mathbf{h}^{(i)}}{\partial \mathbf{h}^{(i-1)}}$
 - 12: **end for**
-



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.