CS 480/680

2/10/26

Lecture 10

Back propagation

Sol 1, 2, 3

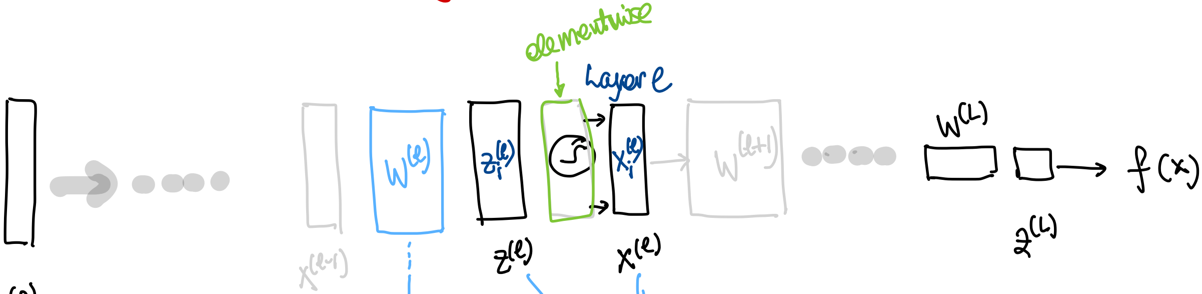
Q1: 11:30 Thu 2/12
15 min

LIV-2 TBPosted

[HW 4 out]

Backpropagation

Training a multilayer network



$x^{(0)} \equiv x$

$m_0 = d$

$m_e \times m_{e-1}$

m_e units

$m_L = 1$

(or $m_L = r$ for r-way classification)

Parameters $\mathcal{W} = \{ W^{(1)}, \dots W^{(L)} \}$ to train

Data $\mathcal{D} = \{ (x^i, y^i), i = 1:n \}$

TRAIN: by G. Descent on loss $\mathcal{L}$

1. Init $\mathcal{W}$ with small random values

$W_{ij}^{(e)} \sim N(0, \sigma^2 \frac{1}{m_e})$

σ^2 small

2. for $t = 1, 2, \dots$

$$\mathcal{W}^{t+1} \leftarrow \mathcal{W}^t - \eta \frac{\partial \mathcal{L}(\mathcal{W}^t)}{\partial \mathcal{W}}$$

step size

y^i, x^{i} ← training ex. i

$x^{(e)}$ ← layer e

x_j } unit

$x^{(0)i}_2$ = input value layer 0, 2nd attribute i-th example

Single unit (Layer L)

$z \equiv f \equiv w^{(L)T} \cdot x^{(0)}$

$\mathbb{W} = \{ w \} \quad w \in \mathbb{R}^d$

$\mathcal{L}_{LS} = \frac{1}{2} (y - f_W(x))^2 \Rightarrow \frac{\partial \mathcal{L}}{\partial f} = f - y = -(y - f)$

target - output

other $\mathcal{L}$: $y \to y_*$ $\quad \frac{\partial \mathcal{L}}{\partial f} = -(y_* - \varphi_{out}) \Rightarrow \cancel{f} = \varphi_{out}(z)$

$f \Rightarrow \varphi_{out}(z)$

Chain Rule:

$n = 1 \quad (x, y) = \mathcal{D}$

$\frac{\partial f}{\partial z} = \varphi'_{out}(z)$

$\frac{\partial z}{\partial w} = x$

$z = w^T x$

$\frac{\partial z}{\partial x} = w$

$\frac{\partial \mathcal{L}}{\partial w} = \frac{\partial \mathcal{L}}{\partial f} \cdot \frac{\partial f}{\partial z} \cdot \frac{\partial z}{\partial w} = -(y_* - \varphi_{out}) \cdot \varphi'_{out}(z) \cdot x$

2 layer network

$$x \to W^{(1)} \to z^{(1)} \to \textcircled{\sigma} \to x^{(1)} \to W^{(2)} \to z^{(2)} \to \varphi^{out}(z^{(2)}) = f$$

$d = m^{(0)}$

$x \uparrow$; $W \uparrow$; $z \uparrow$; $\varphi \uparrow$; $x_k^{(1)} \uparrow x$

$W^{(2)} = W$, $z^{(2)} = z$ ← prev page

$k = 1:m$

$x^{(1)}$ "= f"

$W = \{ W^{(1)}, W^{(2)} \}$

$d \times m \quad m \times 1$

Chain Rule:

$$\frac{\partial f}{\partial z^{(2)}} = \varphi'_{out}(z^{(2)})$$

$$\frac{\partial z^{(2)}}{\partial W^{(2)}} = x^{(1)} \qquad z = W^T x$$

$$\frac{\partial z^{(2)}}{\partial x^{(1)}} = W^{(2)}$$

$$\frac{\partial \mathcal{L}}{\partial W^{(2)}} = \frac{\partial \mathcal{L}}{\partial f} \cdot \frac{\partial f}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial W^{(2)}} = -(y_* - \varphi_{out}) \cdot \varphi'_{out}(z^{(2)}) \cdot x^{(1)}$$

$$\boxed{\frac{\partial f}{\partial x^{(1)}} = \frac{\partial f}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial x^{(1)}} = \underbrace{\varphi'_{out}(z^{(2)}) W^{(2)}}}$$

$$\frac{\partial x_k^{(1)}}{\partial z_k^{(1)}} = \varphi'(z_k^{(1)}) \qquad \frac{\partial x^{(1)}}{\partial z^{(1)}} = \varphi'(z^{(1)}) \in \mathbb{R}^m$$

$$z_k^{(1)} = W_k^{(1)T} x^{(0)} \qquad \text{row } k \text{ of } W^{(1)}$$

$$\frac{\partial z_k^{(1)}}{\partial x^{(0)}} = W_k^{(1)}$$

$$\frac{\partial z_k^{(1)}}{\partial W_k^{(1)}} = x^{(0)}$$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial W_k^{(1)}} = \frac{\partial \mathcal{L}}{\partial f} \cdot \boxed{\frac{\partial f}{\partial x_k^{(1)}}} \cdot \frac{\partial x_k^{(1)}}{\partial z_k^{(1)}} \cdot \frac{\partial z_k^{(1)}}{\partial W_k^{(0)}} = -(y_* - \varphi_{out})\, \varphi'_{out}(z^{(2)}) \cdot \underbrace{\ }_{W_k^{(2)}} \varphi'(z_k^{(1)})\, x^{(0)}$$

$\mathcal{L}$ 2 layer network

$x \to W^{(1)} \to z^{(1)} \to \textcircled{\sigma} \to x^{(1)} \to W^{(2)} \to z^{(2)} \to \varphi^{out}(z^{(2)}) = f$

$x^{(\ell-1)}$ $W^{(\ell)}$ $z^{(\ell)}$ φ (element wise) $x^{(\ell)}$ . . . .

$d = m_0$ m_ℓ ↑ x; W; z; φ; x; $x_k^{(1)}$ $k = 1:m$ "f"

W = W, z = z ← prev page

$\mathbb{W} = \{ W^{(1)}, W^{(2)} \}$

$d \times m$ $\quad m \times 1$

NEW

$$\frac{\partial f}{\partial x^{(\ell+1)}} =$$

$$\frac{\partial f}{\partial x^{(\ell)}} = \frac{\partial f}{\partial x^{(\ell+1)}} \cdot \frac{\partial x^{(\ell+1)}}{\partial x^{(\ell)}}$$

$$\frac{\partial x_j^{(\ell+1)}}{\partial x^{(\ell)}} = \varphi'(z_j^{(\ell+1)}) \cdot \underbrace{\frac{\partial z_j^{(\ell+1)}}{\partial x^{(\ell)}}}_{W_j^{(\ell)}}$$

$$z_j^{(\ell+1)} = W_j^{(\ell+1)T} \cdot x^{(\ell)}$$

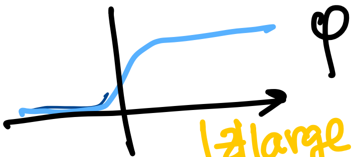
Chain Rule:

$$\frac{\partial f}{\partial z^{(2)}} = \varphi'_{out}(z^{(2)})$$

$$\frac{\partial z^{(2)}}{\partial W^{(2)}} = x^{(1)}$$

$z = W^T x$

$$\frac{\partial z^{(2)}}{\partial x^{(1)}} = W^{(2)}$$



$|z|$ large $\Rightarrow \varphi' \approx 0$

$$\frac{\partial \mathcal{L}}{\partial W^{(2)}} = \frac{\partial \mathcal{L}}{\partial f} \cdot \frac{\partial f}{\partial z^{(2)}} \cdot \frac{\partial z^{(2)}}{\partial W^{(2)}} = -(y_* - \varphi_{out}) \cdot \varphi'_{out}(z^{(2)}) \cdot x^{(1)}$$

$$\boxed{\frac{\partial f}{\partial x^{(\ell)}}} = \frac{\partial f}{\partial z^{(\ell+1)}} \cdot \underbrace{\frac{\partial z_j^{(\ell+1)}}{\partial x^{(\ell)}}}_{W_j^{(\ell)}} =$$

$$\frac{\partial x_k^{(1)}}{\partial z_k^{(1)}} = \varphi'(z_k^{(1)}) \qquad \frac{\partial x^{(1)}}{\partial z^{(1)}} = \varphi'(z^{(1)}) \in \mathbb{R}^m$$

$z_k^{(1)} = W_k^{(1)T} x^{(0)}$ row k of $W^{(1)}$

$$\frac{\partial z_k^{(1)}}{\partial x^{(0)}} = W_k^{(1)}$$

$$\frac{\partial z_k^{(1)}}{\partial W_k^{(1)}} = x^{(0)}$$

$$\Rightarrow \frac{\partial \mathcal{L}}{\partial W_k^{(\ell)}} = \underset{\checkmark}{\frac{\partial \mathcal{L}}{\partial f}} \cdot \boxed{\frac{\partial f}{\partial x_k^{(\ell)}}} \cdot \underset{\checkmark\ \varphi'}{\frac{\partial x_k^{(\ell)}}{\partial z_k^{(\ell)}}} \cdot \underset{x^{(\ell)}}{\frac{\partial z_k^{(\ell)}}{\partial W_k^{(\ell)}}} = \underset{\checkmark}{-(y_* - \varphi_{out})}\, \varphi'_{out}(z^{(2)}) \cdot W_k^{(2)} \cdot \underset{\checkmark}{\varphi'(z_k^{(1)})}\, x^{(\ell)}$$