Text classification

- $X = (w_1, w_2, \cdots, w_T)$, where $w_t \in V$ vocabulary

- $y \in \{1, 2, \cdots, C\}$

- $F(X, y) = \underbrace{w_y}_{\in \mathbb{R}^d}{}^T \phi(X) + \underbrace{b_y}_{\in \mathbb{R}}$ ($\phi(X)$: feature vector)

$$\boxed{p(y \mid X) = \frac{\exp\{w_y^T \phi(X) + b_y\}}{\sum_{y'=1}^{C} \exp\{w_{y'}^T \phi(X) + b_{y'}\}}}$$

$w \in V$

$$\|e(w) - e(u)\| = c \quad \forall w \neq u$$

- one-hot (1-of-K) encoding

$$1h(w) \in \{0,1\}^{|V|} \quad \text{s.t.} \quad 1h^i(w) = \begin{cases} 1, & \text{if } i = \text{hash}(w) \\ 0, & \text{o.w.} \end{cases}$$

$$\text{hash}: V \to \{1, 2, \cdots, |V|\} \quad \text{s.t.} \quad \text{hash}(w) \neq \text{hash}(u) \ \forall w \neq u$$

- (dense) embedding

$$e(w) = \underbrace{W}_{\substack{\text{weight matrix} \\ \in \mathbb{R}^{d' \times |V|}}} \ \underbrace{1h(w)}_{\in \{0,1\}^{|V|}} = \underbrace{W[:, \text{hash}(w)]}_{\text{table lookup}} \in \mathbb{R}^{d'}$$

(W: table)

- $x = (e(w_1), e(w_2) \cdots e(w_T))$

continuous bag-of-words

$\phi(x) = \sum_{t=1}^{T} e(w_t)$

$F(x, y) = u_y^T \phi(x) + b_y$

the y-th coordinate.

$a = \underset{\text{matrix}}{U} \phi(x) + b$, $a \in \mathbb{R}^c$

$U \; \text{th}(y)$

$$U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_c \end{bmatrix} \qquad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_c \end{bmatrix}$$

$F(x, y) = u_y^T \left(\sum_{t=1}^{T} e(w_t) \right) + b_y = \sum_{t=1}^{T} u_y^T e(w_t) + b_y$

assume $b_y = 0$.

$$\log p(y^* \mid x) = \underbrace{u_{y^*}^T \phi(x)}_{= F(y^*, x)} - \underbrace{\log \sum_{y=1}^{C} \exp\{u_y^T \phi(x)\}}_{= \log Z(x)}$$

$$\nabla_{\phi(x)} = u_{y^*} - \sum_{y=1}^{C} \frac{\exp\{u_y^T \phi(x)\}}{\sum_{y'=1}^{C} \exp\{u_{y'}^T \phi(x)\}} u_y$$

$$= u_{y^*} - \mathbb{E}_{y \mid x}[u_y].$$

$$\nabla_{\phi(x)} \frac{\partial \phi(x)}{\partial e(w_t)} = \nabla_{\phi(x)} \quad \text{i.e.} \quad \nabla_{e(w_t)} = \nabla_{\phi(x)} \ldots$$

$$\frac{\partial \log p(y^*|x)}{\partial e(w_t)} = \frac{\partial \log p(y^*|x)}{\partial \phi(x)} \underbrace{\frac{\partial \phi(x)}{\partial e(w_t)}}_{I} = \nabla_{\phi(x)}$$

$$\nabla_{u_{y^*}} = \phi(x) - \frac{\exp\{u_{y^*}^T \phi(x)\}}{\sum_{y'=1}^{C} \exp\{y_{y'}^T \phi(x)\}} \phi(x) = \phi(x) - p(y^*|x)\phi(x)$$

$$= (1 - p(y^*|x))\,\phi(x)$$

$y \neq y^*$

$$\nabla_{u_y} = - \frac{\exp\{u_y^T \phi(x)\}}{\sum_{y'=1}^{C} \exp\{u_{y'}^T \phi(x)\}} \phi(x) = -p(y|x)\phi(x)$$

$$= (0 - p(y|x))\,\phi(X)$$

Nonlinear classification

$V = \{ 0: \text{hate},\ 1: \text{love},\ 2: \text{not} \}$

$y \in \{ 0: \text{positive},\ 1: \text{negative} \}$

$\|e(v)\|_2 = 1$

$e(\text{hate})^T e(\text{not}) = 0, \quad e(\text{love})^T e(\text{not}) = 0$

$e(\text{hate}) = -e(\text{love})$

a hidden layer : 4 neurons $\quad h_i = \overbrace{\max(0, \beta w_i^T e(x))}^{\sigma}$

$i = 1, \cdots, 4$

$w_1 = e(\text{hate})$

$w_2 = e(\text{love})$

$w_3 = e(\text{hate}) + e(\text{not})$

$w_4 = e(\text{love}) + e(\text{not})$

① "hate" $[\beta, 0, \beta, 0]$

$h_1 = \sigma(\beta)$, $h_2 = \sigma(-\beta)$, $h_3 = \sigma(\beta)$, $h_4 = \sigma(0)$

② "not love" $[0, \beta, 0, 2\beta]$

$h_1 = \sigma(-\beta)$, $h_2 = \sigma(\beta)$, $h_3 = \sigma(0)$, $h_4 = \sigma(2\beta)$

③ "not hate" $[\beta, 0, 2\beta, 0]$

$h_1 = \sigma(\beta)$, $h_2 = \sigma(-\beta)$, $h_3 = \sigma(2\beta)$, $h_4 = \sigma(0)$

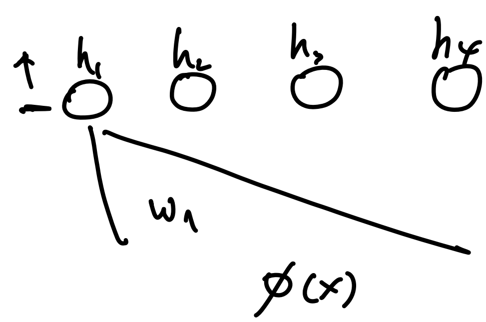
④ "love" $[0, \beta, 0, \beta]$

$h_1 = \sigma(-\beta)$, $h_2 = \sigma(\beta)$, $h_3 = \sigma(0)$, $h_4 = \sigma(\beta)$

class embedding

$u_{\text{negative}} = [1, -1, -\frac{1}{2}, 1]$

$u_{\text{positive}} = [-1, 1, 1, -1/2]$



<u>backpropagation</u>

$F(x, y) = u_y^T \underbrace{\phi(x)}_{\in \mathbb{R}^{d'}} + c_y$

$$\phi(x) = \sigma\Big(\underbrace{W \sum_{t=1}^{T} e(w_t) + b}_{\in \mathbb{R}^d}\Big) \in \mathbb{R}^{d'}$$

∇_{w_i}

...

nonlinearity

$$\nabla_W \ell(x,y) = \frac{\partial \ell}{\underbrace{\partial \phi(a)}_{\text{output for the hidden layer}}} \frac{\partial \phi(a)}{\partial a} \cdot \frac{\partial a}{\partial W}, \quad a = W \sum_{t=1}^{T} e(w_t) + b$$

$$\frac{\partial \ell}{\partial \phi(a)} = u_{y^*} - \mathbb{E}_{y|x}[u_y]$$

$$\frac{\partial \phi(a)}{\partial a} = \mathrm{diag}(\phi'(a)) \qquad \phi_i'(a) = \begin{cases} 1, & \text{if } a_i > 0 \\ 0, & \text{o.w.} \end{cases}$$

$$\frac{\partial a}{\partial W} = \left[\sum_{t=1}^{T} e(w_t)\right]^T$$

$$G : \mathcal{X} \times \Theta \to \mathcal{Y}$$

$$\boxed{\frac{\partial G}{\partial \theta}(x,\theta)} \qquad \boxed{\frac{\partial G}{\partial x}(x,\theta)}$$

$$G(x; W, b) = Wx + b$$

$$G(x; \phi) = \max(0, x)$$

$\vdots$