3)

Data: $(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})$, where $x^{(j)} \in \mathbb{R}^n$

$y^{(j)} \in \{1, \ldots, c\}$

$j = 1, \ldots, m$

m: sample size

n: # of fatures

$\tilde{x} = \begin{bmatrix} x \\ 1 \end{bmatrix}, \quad \tilde{w}_i = \begin{bmatrix} w_i \\ b_i \end{bmatrix}$

$\theta = \{w_i, b_i\}_{i=1, \ldots, c}$

$a_i(x) = \tilde{w}_i^T \tilde{x}$

$$\text{softmax}_i(x) = \frac{e^{w_i^T x + b_i}}{\sum_{k=1}^{c} e^{w_k^T x + b_k}}$$

$$p(x^{(1)}, \ldots, x^{(m)}, y^{(1)}, \ldots, y^{(m)} \mid \theta) = \prod_{i=1}^{m} p(x^{(i)}, y^{(i)} \mid \theta)$$

$$= \prod_{i=1}^{m} p(x^{(i)} \mid \theta) \underbrace{p(y^{(i)} \mid x^{(i)}, \theta)}_{\text{softmax}_j(x^{(i)})}$$

$$\underset{\theta}{\arg\max} \prod_{i=1}^{m} p(x^{(i)} \mid \theta) \, p(y^{(i)} \mid x^{(i)}, \theta)$$

$$= \underset{\theta}{\arg\max} \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}, \theta)$$

Take log:

$$\underset{\theta}{\arg\max} \sum_{i=1}^{m} \log \left[\frac{e^{a_{y^{(i)}}(x^{(i)})}}{\sum_{j=1} e^{a_j(x^{(i)})}} \right]$$

$\arg\max f(\theta) = \underset{\theta}{\arg\min} -f(\theta)$

$$= \underset{\theta}{\arg\min} \frac{1}{m} \sum_{i=1}^{m} \left[\log \left(\sum_{i=1}^{c} e^{a_j(x^{(i)})} \right) - a_{y^{(i)}}(x^{(i)}) \right]$$

$$\nabla_{\tilde{w}_i}\left(\log \sum_{j=1}^{c} e^{a_j(x)}\right)$$

$$= \nabla_{\tilde{w}_i}\left(\log\left[e^{w_1^T\tilde{x}} + e^{w_2^T\tilde{x}} + \dots + e^{w_c^T\tilde{x}}\right]\right)$$

$$= \frac{1}{\sum_{j=1}^{c} e^{a_j(x)}} e^{\tilde{w}_i^T\tilde{x}}\,\tilde{x} \quad \text{, by chain rule}$$

$$= \frac{e^{\tilde{w}_i^T\tilde{x}}}{\sum_{j=1}^{c} e^{a_j(x)}}\tilde{x} = \frac{e^{a_i(x)}}{\sum_{j=1}^{c} e^{a_j(x)}}\tilde{x}$$

$$\nabla_{w_i} a_{y(k)}(x) \Rightarrow \begin{cases} \nabla_{\tilde{w}_i} a_{y(i)}(x) & \text{if } i = y(k) \\ 0 & \text{if } i \neq y(k) \end{cases}$$

If $i = y(k)$

$$\nabla_{\tilde{w}_i} \tilde{w}_i^T\tilde{x} = \tilde{x}$$

$$\nabla_{\tilde{w}_i} \mathcal{L}(\tilde{w}_i) = \frac{1}{m}\sum_{j=1}^{m}\left[\frac{e^{a_i(x(j))}}{\sum_{k=1}^{c} e^{a_k(x(j))}} - \delta_{y(j),i}\right]\tilde{x}(j)$$

$$\boxed{\begin{aligned} \nabla_{w_i} \mathcal{L}(w_i, b_i) &= \frac{1}{m}\sum_{j=1}^{m}\left[\frac{e^{a_i(x(j))}}{\sum_{k=1}^{c} e^{a_k(x(j))}} - \delta_{y(j),i}\right]x(j) \\ \nabla_{b_i} \mathcal{L}(w_i, b_i) &= \frac{1}{m}\sum_{j=1}^{m}\left[\frac{e^{a_i(x(j))}}{\sum_{k=1}^{c} e^{a_k(x(j))}} - \delta_{y(j),i}\right] \end{aligned}}$$

4)

$$D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(k)}, y^{(k)})\}$$

$$x^{(i)} \in \mathbb{R}^d, \quad y^{(i)} \in \{-1, 1\}$$

$$\tilde{w} = \begin{bmatrix} w \\ 1 \end{bmatrix}$$

$$\tilde{x}^{(i)} = \begin{bmatrix} x^{(i)} \\ 1 \end{bmatrix}$$

$$\mathcal{L}(w, b) = \frac{1}{k} \sum_{i=1}^{m} \underbrace{\max(0, 1 - y^{(i)}(w^T x^{(i)} + b))}_{\text{hinge}_{y^{(i)}}(x^{(i)})}$$

$$\frac{\partial \, \text{hinge}_{y^{(i)}}(x^{(i)})}{\partial \tilde{w}} = \begin{cases} 0 & \text{if } y_i(\tilde{w}^T \tilde{x}^{(i)}) \geq 1 \\ -y_i \tilde{x}^{(i)} & \text{if } y_i(\tilde{w}^T \tilde{x}^{(i)}) < 1 \end{cases}$$

$$\frac{\partial \mathcal{L}}{\partial \tilde{w}} = \frac{1}{k} \sum_{i=1}^{m} \frac{\partial \, \text{hinge}_{y^{(i)}}(x^{(i)})}{\partial \tilde{w}} = \frac{1}{k} \sum_{i=1}^{m} \mathbb{1}_{y_i(\tilde{w}^T \tilde{x}^{(i)})} (-y_i \tilde{x})$$

$$\boxed{\begin{aligned} \frac{\partial \mathcal{L}}{\partial w} &= \frac{1}{k} \sum_{i=1}^{m} \mathbb{1}_{y_i(w^T x^{(i)} + b)} (-y_i x^{(i)}) \\ \frac{\partial \mathcal{L}}{\partial b} &= \frac{1}{k} \sum_{i=1}^{m} \mathbb{1}_{y_i(w^T x^{(i)} + b)} (-y_i) \end{aligned}}$$