6. K-class classification $\rightarrow$ output space $\mathcal{Y} = \{1, 2, \ldots, K\}$

Matrix $W \rightarrow$ represents a hypothesis $h_y()$

$$W = [\vec{w}_1\ \vec{w}_2 \cdots \vec{w}_k \cdots \vec{w}_K]_{(d+1)\times K} \qquad \vec{w}_k \in \mathbb{R}^{d+1}\ \forall k \in \{1, \ldots, K\}$$

$$h_y(\vec{x}) = \frac{e^{\vec{w}_y^T \vec{x}}}{\sum_{k=1}^{K} e^{\vec{w}_k^T \vec{x}}} \xrightarrow{\text{approx.}} P(y \mid \vec{x})$$

$$D = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), \ldots, (\vec{x}_N, y_N)\} \overset{iid}{\sim} p(\vec{x}), \text{ target distribution } p(y \mid \vec{x})$$

$$l(h_y \mid D) = p(D \mid h_y) \propto \prod_{n=1}^{N} h_{y_n}(\vec{x}_n)$$

$$\min -\ln \text{likelihood} \propto \min -\ln \prod_{n=1}^{N} h_{y_n}(\vec{x}_n) = \min \sum_{n=1}^{N} -\ln h_{y_n}(\vec{x}_n) = \min \sum_{n=1}^{N} err(W, \vec{x}_n, y_n)$$

$$= \min \sum_{n=1}^{N} -\ln \frac{e^{\vec{w}_{y_n}^T \vec{x}_n}}{\sum_{k=1}^{K} e^{\vec{w}_k^T \vec{x}_n}}$$

$$= \min \sum_{n=1}^{N} \left[-\vec{w}_{y_n}^T \vec{x}_n + \ln\left(\sum_{k=1}^{K} e^{\vec{w}_k^T \vec{x}_n} \right) \right]$$

$$\frac{\partial\, err(W, \vec{x}, y)}{\partial \vec{w}_i} = \frac{\partial}{\partial \vec{w}_i} \left[-\vec{w}_y^T \vec{x} + \ln\left(\sum_{k=1}^{K} e^{\vec{w}_k^T \vec{x}} \right) \right]$$

$$= \frac{\partial}{\partial \vec{w}_i}(-\vec{w}_y^T \vec{x}) + \frac{\partial}{\partial \vec{w}_i}\left(\sum_{k=1}^{K} e^{\vec{w}_k^T \vec{x}} \right)$$

$$= \begin{cases} -\vec{x} + h_i(\vec{x}) \cdot \vec{x} & y = i \\ h_i(\vec{x}) \cdot \vec{x} & y \neq i \end{cases}$$

$\frac{\partial}{\partial \vec{w}_i}(-\vec{w}_y^T \vec{x}) \rightarrow$ if $y = i$, then $\frac{\partial}{\partial \vec{w}_i}(-\vec{w}_y^T \vec{x})$

$$= \frac{\partial}{\partial \vec{w}_i}(-\vec{w}_i^T \vec{x}) = -\vec{x}$$

if $y \neq i$, then $\frac{\partial}{\partial \vec{w}_i}(-\vec{w}_y^T \vec{x}) = 0$

$$\frac{\partial}{\partial \vec{w}_i}\left(\sum_{k=1}^{K} e^{\vec{w}_k^T \vec{x}} \right) \underset{\text{chain rule}}{\uparrow =} \underbrace{\boxed{\frac{1}{\sum_{k=1}^{K} e^{\vec{w}_k^T \vec{x}}} \cdot e^{\vec{w}_i^T \vec{x}}}}_{= h_i(\vec{x})} \cdot \vec{x}$$

$$= h_i(\vec{x}) \cdot \vec{x}$$

$$\left(\frac{\partial\, err(W, \vec{x}_n, y_n)}{\partial \vec{w}_i} = \begin{cases} -\vec{x}_n + h_i(\vec{x}_n) \cdot \vec{x}_n & y_n = i \\ h_i(\vec{x}_n) \cdot \vec{x}_n & y_n \neq i \end{cases} \right)$$

Let $V = [\vec{v}_1\ \vec{v}_2 \cdots \vec{v}_K]_{(d+1)\times K}$, where $\vec{v}_i \in \mathbb{R}^{d+1}$ $\forall i \in \{1, \ldots, K\}$

update direction for $\vec{w}_1$

$\forall i \in \{1, \ldots, K\}$ $i \neq y_n$, $\vec{v}_i = -h_i(\vec{x}_n) \cdot \vec{x}_n$

For $i = y_n$. $\vec{v}_i = \vec{v}_{y_n} = -[-\vec{x}_n + h_i(\vec{x}_n)\vec{x}_n] = \vec{x}_n - h_i(\vec{x}_n)\vec{x}_n = \vec{x}_n(1 - h_i(\vec{x}_n))$

i.e. $$V = [\vec{x}_n \cdot -h_i(\vec{x}_n) \cdots \boxed{\vec{x}_n(1 - h_i(\vec{x}_n))} \cdots \vec{x}_n \cdot -h_i(\vec{x}_n)]$$

$$= \vec{x}_n [-h_i(\vec{x}_n) \cdots -h_i(\vec{x}_n)\ \underbrace{\boxed{(1 - h_i(\vec{x}_n))}}_{i = y_n}\ -h_i(\vec{x}_n) \cdots -h_i(\vec{x}_n)]$$

$$= \vec{x}_n \cdot \vec{u}^T$$

$$\therefore \vec{u}^T = [-h_i(\vec{x}_n) \cdots (1 - h_i(\vec{x}_n)) \cdots -h_i(\vec{x}_n)]$$

$$\Rightarrow \vec{u} = [-h_i(\vec{x}_n) \cdots (1 - h_i(\vec{x}_n)) \cdots -h_i(\vec{x}_n)]^T \in \mathbb{R}^{K \times 1} \quad \square$$