$$J = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\log\left(\frac{\exp f_k}{\sum_{c=1}^{K}\exp f_c}\right) + \lambda\sum_{j=1}^{d} w_{kj}^2$$

– (10 pts) Given this formula, show the steps to derive the gradient of J with respect to $\mathbf{w}_k$.

Input: feature vectors $x_1, x_2, \cdots x_N \in R^d$. labels $y_1, y_2, \cdots y_N \in R^k$

$\phi = Wx \in R^k$, here x is one of $x_1, x_2, \cdots x_n$

$$W: k \times d \quad \begin{array}{c c} & d \\ k & \begin{bmatrix} w_1 \text{---} \\ w_2 \text{---} \\ \vdots \\ w_k \text{---} \end{bmatrix} \end{array} \qquad x_i : d \times 1 \quad \begin{array}{c c} 1 & \\ \begin{bmatrix} \\ \\ \end{bmatrix} & d \end{array} \qquad \phi: k \times 1 \quad \begin{array}{c c} 1 & \\ \begin{bmatrix} \\ \\ \end{bmatrix} & k \end{array} \qquad y_i: k \times 1 \quad \begin{array}{c c} 1 & \\ \begin{bmatrix} \\ \\ \end{bmatrix} & k \end{array}$$

Softmax function: $P_k = \dfrac{\exp(\phi_k)}{\sum_{j=1}^{k}\exp(\phi_j)}$ $\qquad \phi_k = W_k x_i$

$\vec{p} = [P_1, P_2, \cdots P_k]^T \in R^k$.

$$J = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\left[\log(\exp(f_k)) - \log\left(\sum_{c=1}^{K}\exp(f_c)\right)\right] + \lambda\sum_{j=1}^{d} W_{kj}^2$$

$$= -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} f_k + \frac{1}{N}\cdot\sum_{i=1}^{N}\underbrace{\sum_{k=1}^{K} y_{ik}}_{=1}\cdot\log\left(\sum_{c=1}^{K}\exp(f_c)\right) + \lambda\sum_{j=1}^{d} W_{kj}^2 \qquad \boxed{\sum_{k=1}^{K} y_{ik} = 1.}$$

$$= -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} f_k + \frac{1}{N}\sum_{i=1}^{N}\log\left(\sum_{c=1}^{K}\exp(f_c)\right) + \lambda\sum_{j=1}^{d} W_{kj}^2$$

$$= H + \lambda\sum_{j=1}^{d} W_{kj}^2 \qquad \text{(set the first two term as } H\text{)}$$

Chain rules:

$$\frac{\partial H}{\partial W_r} = \sum_{g=1}^{K}\frac{\partial f_g}{\partial W_r}\cdot\frac{\partial H}{\partial f_g} \qquad W_r: \text{the r-th row of } W$$

Since $f_g = x^T W_g \Rightarrow \dfrac{\partial H}{\partial W_k} = \sum_{g=1}^{k}\dfrac{\partial f_g}{\partial W_k}\cdot\dfrac{\partial H}{\partial f_g}$

$$= \underbrace{\frac{\partial f_k}{\partial W_k}}_{=x}\cdot\frac{\partial H}{\partial f_k} + \sum_{g\neq k}\underbrace{\frac{\partial f_g}{\partial W_k}}_{=0}\cdot\frac{\partial H}{\partial f_g}$$

Since $\dfrac{\partial f_k}{\partial W_k} = \dfrac{\partial x^T W_k}{\partial W_k} = x$, $\dfrac{\partial f_g}{\partial W_k} = 0$ for $g \neq k$.

that is to say: for the sum, only f related to k left after the derivation.

$\therefore \quad \dfrac{\partial H}{\partial W_k} = x \cdot \dfrac{\partial H}{\partial f_k}$, now we need to calculate $\dfrac{\partial H}{\partial f_k}$

In H: $\dfrac{\partial y_{ik} f_k}{\partial f_k} = y_{ik}$, $\dfrac{\partial \log\left(\sum_{c=1}^{k} \exp(f_c)\right)}{\partial f_k} = \dfrac{\exp(f_k)}{\sum_{c=1}^{k} \exp(f_c)}$

$$\therefore \quad \frac{\partial H}{\partial W_k} = -\frac{1}{N} \sum_{i=1}^{N} x_i \cdot y_{ik} + \sum_{i=1}^{N} x_i \cdot \frac{\exp(f_k)}{\sum_{c=1}^{k} \exp(f_c)}$$

$$= -\frac{1}{N} \sum_{i=1}^{N} x_i \left(y_{ik} - \frac{\exp(f_k)}{\sum_{c=1}^{k} \exp(f_c)} \right)$$

Since $\lambda \sum_{j=1}^{d} w_{kj}^2 = \lambda \cdot \|W_k\|_2^2 \rightarrow$ L-2 norm.

$$\therefore \quad \frac{\partial \left[\lambda \cdot \|W_k\|_2^2 \right]}{\partial W_k} = \lambda \cdot 2W_k = 2\lambda W_k$$

$$\therefore \quad \frac{\partial J}{\partial W_k} = -\frac{1}{N} \sum_{i=1}^{N} x_i \left(y_{ik} - \frac{\exp(f_k)}{\sum_{c=1}^{k} \exp(f_c)} \right) + 2\lambda W_k$$

$\square$

$$\begin{array}{cc} & d \\ n & \begin{bmatrix} x_1 & \text{---} \\ x_2 & \text{---} \\ \vdots & \\ x_n & \text{---} \end{bmatrix} \end{array} \qquad \begin{array}{cc} & k \\ n & \begin{bmatrix} y_1 & \text{---} \\ y_2 & \text{---} \\ \vdots & \\ y_n & \text{---} \end{bmatrix} \end{array}$$

$$\begin{array}{cc} & d \\ k & \begin{bmatrix} w_1 & \text{---} \\ w_2 & \text{---} \\ \vdots & \\ w_k & \text{---} \end{bmatrix} \end{array} \times \begin{array}{cc} & n \\ d & \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix} \end{array} = \begin{array}{cc} & n. \\ k & \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_n. \end{bmatrix} \end{array}$$

$k \times d \qquad d \times n.$

Softmax

$$= \begin{array}{cc} & n. \\ k & \begin{bmatrix} \vec{p}_1, & \vec{p}_2 & \cdots & \vec{p}_n \end{bmatrix} \end{array}$$

log

y_i^T

$$\begin{array}{cc} & n. \\ k & \begin{bmatrix} \vec{y}_1, & \vec{y}_2 & \cdots & \vec{y}_n \end{bmatrix} \end{array}$$

$j=1$

$$H(y_i, p_i) = -\sum_{k=1}^{K} y_{ik} \log(p_{ik})$$

$T_i = p_i - y_i$

$$\overbrace{(p_i - y_i)}^{T_i} x_i^T \qquad k \begin{bmatrix} | \end{bmatrix} \cdot \begin{bmatrix} x_1 & \underset{d}{\text{———}} \end{bmatrix} \in k \times n.$$

$$\begin{array}{cc} & n \\ k & \begin{bmatrix} T_1 & T_2 & & T_n \\ | & | & & | \end{bmatrix} \end{array} \begin{array}{cc} & d \\ n & \begin{bmatrix} x_1 & \text{---} \\ x_2 & \text{---} \\ \vdots & \\ x_n & \text{---} \end{bmatrix} \end{array}$$

$k \times n \quad n \times d \quad k \times d.$

$k \times b \quad b \times d \quad k \times 1 \quad 1 \times d.$