$$J : \mathbb{R}^{n+1} \to \mathbb{R}, \quad \theta \mapsto J(\theta)$$

Sea $j = 0, 1, \dots, n$

$$\frac{\partial J}{\partial \theta_j}(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left[h_\theta(x^{(i)}) - y^{(i)} \right] x_j^{(i)}$$

$$\mu_i := h_\theta(x^{(i)}) = \sigma(\theta^T x^{(i)})$$

$x^{(i)}$ es la i-ésima instancia $(\in \mathbb{R}^{n+1})$

$$\mu := \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_m \end{bmatrix} \qquad y := \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

1

$$\alpha \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} + \beta \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

$$\begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \\ a_3 & b_3 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

$$g := \begin{bmatrix} \frac{\partial J}{\partial \theta_0}(\theta) \\ \vdots \\ \frac{\partial J}{\partial \theta_n}(\theta) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{m} (\mu_i - y^{(i)}) x_0^{(i)} \\ \vdots \\ \sum_{i=1}^{m} (\mu_i - y^{(i)}) x_n^{(i)} \end{bmatrix}$$

$$= \sum_{i=1}^{m} (\mu_i - y^{(i)}) \begin{bmatrix} x_0^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix}$$

$$= \begin{bmatrix} x_0^{(1)} & \cdots & x_0^{(m)} \\ \vdots & & \\ x_n^{(1)} & \cdots & x_n^{(m)} \end{bmatrix}_{(n+1) \times m} \begin{bmatrix} \mu_1 - y^{(1)} \\ \vdots \\ \mu_m - y^{(m)} \end{bmatrix}_{m \times 1}$$

$$= X^T (\mu - y)$$

Sean $j, k = 0, 1, \ldots, n$

$$\frac{\partial^2 J}{\partial \theta_k \partial \theta_j}(\theta) = \frac{\partial}{\partial \theta_k}\left(\frac{\partial J}{\partial \theta_j}(\theta)\right)$$

$$= \frac{\partial}{\partial \theta_k}\left(\sum_{i=1}^{m} (\sigma(\theta^T x^{(i)}) - y_i)\, x_j^{(i)}\right)$$

$$= \sum_{i=1}^{m} \frac{\partial}{\partial \theta_k}\, x_j^{(i)} \sigma(\theta^T x^{(i)})$$

$$= \sum_{i=1}^{m} x_j^{(i)} \sigma(\theta^T x^{(i)})\left[1 - \sigma(\theta^T x^{(i)})\right] x_k^{(i)}$$

$$= \sum_{i=1}^{m} x_j^{(i)} \mu_i (1-\mu_i) x_k^{(i)}$$

$$= \sum_{i=1}^{m} \mu_i (1-\mu_i) x_j^{(i)} \cdot x_k^{(i)}$$

(2)

$$\left[\frac{\partial^2 J}{\partial \theta_k \partial \theta_j}\right]_{jk} = \begin{bmatrix} \cdots & \mu_1(1-\mu_1) & \cdots \\ & \vdots & \\ \cdots & & \cdots \mu_m(1-\mu_m) \end{bmatrix} \begin{bmatrix} x_0^{(1)} & \cdots & x_n^{(1)} \\ \vdots & & \vdots \\ x_0^{(m)} & \cdots & x_n^{(m)} \end{bmatrix}$$

$(n+1) \times m \qquad m \times (n+1)$

$$= \begin{bmatrix} x_0^{(1)} & \cdots & x_0^{(m)} \\ \vdots & & \vdots \\ x_n^{(1)} & \cdots & x_n^{(m)} \end{bmatrix} \begin{bmatrix} \mu_1(1-\mu_1) & & \\ & \ddots & \\ & & \mu_m(1-\mu_m) \end{bmatrix} \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}$$

$(n+1) \times m \qquad m \times m$

$$= X S X$$

$$x^T A x$$

$$H : \mathbb{R}^{n+1} \longrightarrow \mathbb{R}^{n+1} \times \mathbb{R}^{n+1}$$

$$\theta \longmapsto H(\theta)$$

$n+1$

BLANK

$J : \mathbb{R}^{n+1} \to \mathbb{R}$

$H : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1} \times \mathbb{R}^{n+1}$

$H = X^T S X$

donde >0

$S = \begin{bmatrix} \mu_1(1-\mu_1) & & \\ & \ddots & \\ & & \mu_m(1-\mu_m) \end{bmatrix}$ (>0)

$X_{m \times (n+1)}$ $(m > n+1)$

Decir que X es de rango máximo (full rank) ③

es decir que $\text{rank}(X) = n+1$

Sea $a = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \in \mathbb{R}^{n+1}$

Como X es de rango máximo, si

$\alpha_0 \begin{bmatrix} x^0 \end{bmatrix} + \cdots + \alpha_n \begin{bmatrix} x^n \end{bmatrix} = \begin{bmatrix} x^0 \cdots x^n \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_n \end{bmatrix} = Xa = 0$, $\therefore 0 \neq b$

entonces $\alpha_0 = \alpha_1 = \cdots = \alpha_n = 0$, i.e., $a = 0$

Luego, esto es equivalente a

$\forall a \in \mathbb{R}^{n+1} \setminus \{0\} : Xa \neq 0$

Finalmente, veamos que H sea definida positiva.

En efecto, sea $a \in \mathbb{R}^{n+1} \setminus \{0\}$

$a^T H a = a^T X^T S X a$

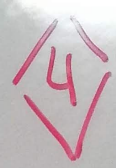
$= (Xa)^T S (Xa)$

$= \sum_{i=1}^{m} \underbrace{\mu_i(1-\mu_i)}_{>0} b_i^2$

> 0

Exercise 8.4

(a) For every $k = 1, 2, \dots, G$ let



$$\nu_k : \mathbb{R}^G \longrightarrow \mathbb{R}$$

$$\begin{bmatrix} \eta_1 \\ \vdots \\ \eta_G \end{bmatrix} = \eta \longmapsto \nu_k(\eta) := \frac{\exp(\eta_k)}{\sum_{c'=1}^{G} \exp(\eta_{c'})} \,.$$

$\sum(\eta)$

Let $k, j = 1, 2, \dots, G$. Show that

$$\frac{\partial \nu_k}{\partial \eta_j}(\eta) = \nu_k(\eta)\left(\delta_{kj} - \nu_j(\eta)\right).$$

Indeed, let $\eta \in \mathbb{R}^G$; then

$$\frac{\partial \nu_k}{\partial \eta_j}(\eta) = \frac{\frac{\partial}{\partial \eta_j}\left(\exp(\eta_k)\right)\Sigma - \exp(\eta_k)\sum_{c'=1}^{G}\frac{\partial}{\partial \eta_j}\left(\exp(\eta_{c'})\right)}{\Sigma^2}$$

$$= \frac{\delta_{kj}\exp(\eta_k)\Sigma - \exp(\eta_k)\exp(\eta_j)}{\Sigma^2}$$

$$= \frac{\exp(\eta_k)}{\Sigma}\left(\delta_{kj} - \frac{\exp(\eta_j)}{\Sigma}\right)$$

$$= \nu_k(\eta)\left(\delta_{kj} - \nu_j(\eta)\right)$$

(b) Let $\ell : \mathbb{R}^{D \times C} \longrightarrow \mathbb{R}$

$$[w_1 \cdots w_C] = W \longmapsto \ell(W) := \sum_{i=1}^{m} \sum_{k=1}^{C} y_{ik} \log\left(\mu_{ik}(W)\right)$$

5

where $y_{ik} := \delta_{y_i k}$, $\mu_{ik} : \mathbb{R}^{D \times C} \longrightarrow \mathbb{R}$

$$W \longmapsto \frac{\exp(w_k^T x_i)}{\sum_{c'=1}^{C} \exp(w_{c'}^T x_i)}$$

and $x_1, x_2, \ldots, x_m$ are the instances.

Hence, show that

$$\nabla_W \ell = \begin{bmatrix} \nabla_{w_1} \ell \\ \vdots \\ \nabla_{w_C} \ell \end{bmatrix}$$

where

$$\nabla_{w_c} \ell(W) = \sum_{i=1}^{m} \left(y_{ic} - \mu_{ic}(W)\right) x_i$$

for every $c = 1, 2, \ldots, C$.

Indeed, let $c = 1, 2, \dots, G$. Then

$$\nabla_{w_c} \ell(W) = \begin{bmatrix} \frac{\partial}{\partial w_{c1}} \left(\ell(W) \right) \\ \vdots \\ \frac{\partial}{\partial w_{cD}} \left(\ell(W) \right) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{m} \sum_{k=1}^{G} y_{ik} \frac{1}{\mu_{ik}(W)} \frac{\partial \mu_{ik}}{\partial w_{c1}}(W) \\ \vdots \\ \sum_{i=1}^{m} \sum_{k=1}^{G} y_{ik} \frac{1}{\mu_{ik}(W)} \frac{\partial \mu_{ik}}{\partial w_{cD}}(W) \end{bmatrix}$$

$$= \sum_{i=1}^{m} \sum_{k=1}^{G} \frac{y_{ik}}{\mu_{ik}(W)} \begin{bmatrix} \frac{\partial \mu_{ik}}{\partial w_{c1}}(W) \\ \vdots \\ \frac{\partial \mu_{ik}}{\partial w_{cD}}(W) \end{bmatrix}$$

6

Let $i = 1, 2, \dots, m$. Then, let $g_i : \mathbb{R}^{D \times C} \longrightarrow \mathbb{R}^C$

$$W \longmapsto g_i(W) := \begin{bmatrix} w_1^T x_i \\ \vdots \\ w_C^T x_i \end{bmatrix} =: \begin{bmatrix} \eta_1(W) \\ \vdots \\ \eta_C(W) \end{bmatrix}$$

Now, let $k = 1, 2, \dots, C$ and let ν_k as item (a) above.

Since the following diagram holds:

7

$$\begin{array}{ccc} \mathbb{R}^{D \times C} & \xrightarrow{g_i} & \mathbb{R}^C \\ & \searrow^{\mu_{ik}} & \downarrow \nu_k \\ & & \mathbb{R} \end{array}$$

we can apply 'the chain rule' once again! Then, for every $r = 1, 2, \dots, D$

$$\frac{\partial\, \nu_k \circ g_i}{\partial w_{cr}}(W) = \sum_{j=1}^{C} \frac{\partial \nu_k}{\partial \eta_j}\left(g_i(W)\right) \frac{\partial \eta_j}{\partial w_{cr}}(W)$$

$$= \frac{\partial \nu_k}{\partial \eta_c}\left(g_i(W)\right) \frac{\partial \eta_c}{\partial w_{cr}}(W)$$

$$\overset{\text{item (a)}}{=} \nu_k\left(g_i(W)\right)\left(\delta_{kc} - \nu_c\left(g_i(W)\right)\right) x_{ir}$$

$$= \mu_{ik}(W)\left(\delta_{kc} - \mu_{ic}(W)\right) x_{ir}$$

(cont.)

8

$$\nabla_{w_c} \ell(W) = \sum_{i=1}^{m} \sum_{k=1}^{G} \frac{y_{ik}}{\mu_{ik}(W)} \begin{bmatrix} \mu_{ik}(W)\left(\delta_{kc} - \mu_{ic}(W)\right) x_{i1} \\ \vdots \\ \mu_{ik}(W)\left(\delta_{kc} - \mu_{ic}(W)\right) x_{iD} \end{bmatrix}$$

$$= \sum_{i=1}^{m} \sum_{k=1}^{G} y_{ik}\left(\delta_{kc} - \mu_{ic}(W)\right) \begin{bmatrix} x_{i1} \\ \vdots \\ x_{iD} \end{bmatrix}$$

$$= \sum_{i=1}^{m} \sum_{k=1}^{G} y_{ik}\left(\delta_{kc} - \mu_{ic}(W)\right) x_i$$

$$= \sum_{i=1}^{m} \left[y_{ic} - y_{ic}\mu_{ic}(W) - \sum_{k \neq c} y_{ik}\mu_{ic}(W) \right] x_i$$

$$= \sum_{i=1}^{m} \left[y_{ic} - \left(\sum_{k=1}^{G} y_{ik}\right)\mu_{ic}(W) \right] x_i$$

$$= \sum_{i=1}^{m} \left(y_{ic} - \mu_{ic}(W)\right) x_i$$

(9) (c) Show that the Hessian of ℓ : $H\ell$ is given by

$$H\ell(W) = \begin{bmatrix} H_{1,1}\ell(W) & \cdots & H_{1,G}\ell(W) \\ \vdots & & \vdots \\ H_{G,1}\ell(W) & \cdots & H_{G,G}\ell(W) \end{bmatrix}$$

where

$$H_{c,c'}\ell(W) = -\sum_{i=1}^{m} \mu_{ic}(W)\left(\delta_{c,c'} - \mu_{i,c'}(W)\right) x_i x_i^T$$

for every $c,c' = 1,2,\ldots,G$.

Indeed, let $c,c' = 1,2,\ldots,G$. Then, by definition of the Hessian and item (b), the element (j,k) of the matrix $H_{c,c'}\ell(W)$ is given by

$$\frac{\partial}{\partial w_{c',k}}\left[\sum_{i=1}^{m}\left(y_{ic} - \mu_{ic}(W)\right)x_{ij}\right] = -\sum_{i=1}^{m} x_{ij}\frac{\partial\, \nu_c \circ g_i}{\partial w_{c',k}}(W)$$

(y_{ic}: Does NOT depend on $w_{c',k}$; x_{ij}: does not depend on $w_{c',k}$)

$\mu_{ic} = \nu_c \circ g_i$, $g_i(W) = W^T x_i$

$\eta_{i,s}(W) = w_s^T x_i$

'chain rule' again!

$$= -\sum_{i=1}^{m} x_{ij}\left[\sum_{s=1}^{G}\frac{\partial \nu_c}{\partial \eta_{i,s}}(g_i(W))\frac{\partial \eta_{i,s}}{\partial w_{c',k}}(W)\right]$$

$$= -\sum_{i=1}^{m} x_{ij}\left[\frac{\partial \nu_c}{\partial \eta_{i,c'}}(g_i(W))\frac{\partial \eta_{i,c'}}{\partial w_{c',k}}(W)\right]$$

10

$$= -\sum_{i=1}^{m} x_{ij} \left[\frac{\partial \nu_c}{\partial \eta_{i,c'}} (g_i(W)) \, x_{i,K} \right]$$

by item (a)

$$= -\sum_{i=1}^{m} x_{ij} \, \nu_c(g_i(W)) \left(\delta_{c,c'} - \nu_{c'}(g_i(W))\right) x_{i,K}$$

$$= -\sum_{i=1}^{m} \mu_{i,c}(W) \left(\delta_{c,c'} - \mu_{i,c'}(W)\right) x_{ij} \, x_{iK}$$

Then, since $\mu_{i,c}(W)\left(\delta_{c,c'} - \mu_{i,c'}(W)\right)$ does NOT depend on j neither K,

$j, K \in \{1, 2, \ldots, D\}$ and $x_i x_i^T = \begin{bmatrix} x_{i1}x_{i1} & \cdots & x_{i1}x_{iD} \\ \vdots & & \vdots \\ x_{iD}x_{i1} & \cdots & x_{iD}x_{iD} \end{bmatrix}$,

$$H_{c,c'} \ell(W) = -\sum_{i=1}^{m} \mu_{i,c}(W) \left(\delta_{c,c'} - \mu_{i,c'}(W)\right) x_i x_i^T.$$

□

9

In
the