Question 1.

$$\log \frac{P(y=c_1 | x; W)}{P(y=c_2 | x; W)} = \log \frac{\frac{\exp(W_{c_1} \cdot x)}{\cancel{\sum_{y=1}^{C} \exp(W_y \cdot x)}}}{\frac{\exp(W_{c_2} \cdot x)}{\cancel{\sum_{y=1}^{C} \exp(W_y \cdot x)}}} = \log \exp(W_{c_1} - W_{c_2}) \cdot x = (W_{c_1} - W_{c_2}) \cdot x$$

Where $(W_{c_1} - W_{c_2})$ is a constant given any c_1 and c_2.

Thus the log odd is a linear function of x.

When $C=2$,

$$P(y=1 | x; W) = \frac{\exp(W_1 \cdot x)}{\exp(W_1 \cdot x) + \exp(W_2 \cdot x)}$$

To find an alternative setting:

$$\begin{cases} P(y=1 | x; W) + P(y=2 | x; W) = 1 \\ P(y=1 | x; W) = P(y=2 | x; W) \cdot \exp(W_1 - W_2) \cdot x \end{cases}$$

$$\Rightarrow P(y=1 | x, W) = \frac{\exp(W_1 - W_2) \cdot x}{1 + \exp(W_1 - W_2) \cdot x} = \frac{1}{1 + \exp -(W_1 - W_2) \cdot x}$$

$$= \sigma[(W_1 - W_2) \cdot x]$$

$$= \sigma(V \cdot x) \quad \text{where } V = W_1 - W_2$$

$\Rightarrow$ There exists a single parameter $V = (W_1 - W_2)$ such that the above equation holds.

由 扫描全能王 扫描创建

Question 2:

For any $C_k \in \{1, \cdots C\}$

$$P(y=k \mid X; W) = \frac{\exp(W_k \cdot X)}{\sum_{c=1}^{C} \exp(W_c \cdot X)} = \frac{\exp(W_k \cdot X)}{\exp(W_1 \cdot X) + \exp(W_2 \cdot X) + \cdots \exp(W_k \cdot X) + \cdots \exp(W_C \cdot X)}$$

$$= \frac{1}{\exp(W_1 - W_k) \cdot X + \cdots + \underbrace{\exp(W_k - W_k) \cdot X} + \cdots + \exp(W_C - W_k) \cdot X}$$

$$= \frac{1}{1 + \underbrace{\sum_{c=1}^{k-1} (W_c - W_k) \cdot X + \sum_{c=k+1}^{C} (W_c - W_k) \cdot X}_{C-1 \text{ number of parameters}}} \quad (*)$$

setting (*) yeilds the same $p(y|X)$ for every X.

This also shows we only need $C-1$ parameters because there is only $C-1$ parameter in equation (*), where we just substract category k from each category.

由 扫描全能王 扫描创建

Question 3.

For logistic regression.

$$L(w) = \arg\min \log(p(y|X)) = -\sum_{i=1}^{n} y_i \log \sigma(w \cdot x_i) + (1-y_i)\log(1-\sigma(w \cdot x_i))$$

where $\sigma(w \cdot x_i) = \frac{1}{1+e^{-w \cdot x_i}}$

$$\frac{\partial \log \sigma(w \cdot x_i)}{\partial w_j} = \frac{\partial -\log(1+e^{-w \cdot x_i})}{\partial w_j} = + \frac{x_{ij} e^{-w \cdot x_i}}{1+e^{-w \cdot x_i}} = x_{ij}(1-\sigma(w \cdot x_i))$$

$$\frac{\partial \log(1-\sigma(w \cdot x_i))}{\partial w_j} = \frac{-x_{ij} e^{-w \cdot x_i}}{1-\frac{1}{1+e^{-w \cdot x_i}}} = -x_{ij}\sigma(w \cdot x_i)$$

F.O.C: $$\frac{\partial L(w)}{\partial w_j} = -\sum_{i=1}^{n} y_i x_{ij}(1-\sigma(w \cdot x_i)) - (1-y_i) x_{ij} \sigma(w \cdot x_i)$$

$$= -\sum_{i=1}^{n} y_i x_{ij} - \cancel{y_i x_{ij} \sigma(w \cdot x_i)} - x_{ij}\sigma(w \cdot x_i) + \cancel{y_i x_{ij} \sigma(w \cdot x_i)}$$

$$= \sum_{i=1}^{n} x_{ij}\sigma(w \cdot x_i) - y_i x_{ij}$$

$$= \sum_{i=1}^{n} x_{ij}[\sigma(w \cdot x_i) - y_i]$$

F.O.C gives the gradient $= X^T[\sigma(w \cdot x_i) - Y]$

S.O.C: $$\frac{\partial L(w)}{\partial w_j \partial w_k} = \sum_{i=1}^{n} x_{ij} x_{ik} \sigma(w \cdot x_i)[1-\sigma(w \cdot x_i)]$$

$$= z_j^T \beta z_k$$

where $z_j^T = [x_{1j}, \quad x_{nj}]$ $\quad \beta = \mathrm{diag}(\sigma(w \cdot x_1)[1-\sigma(w \cdot x_1)], \cdots \sigma(w \cdot x_n)[1-\sigma(w \cdot x_n)])$

$$z_k = \begin{bmatrix} x_{1k} \\ \vdots \\ \\ x_{nk} \end{bmatrix} \qquad \sigma(w \cdot x_i)[1-\sigma(w \cdot x_i)] = \frac{1}{1+e^{-w \cdot x_i}} \frac{e^{-w \cdot x_i}}{1+e^{-w \cdot x_i}} \in [0,1] \quad (*)$$

Because z_j is the j^{th} column of X, z_k is the k^{th} column of X

$z_j^T \beta z_k$ is the $j\text{-}k^{th}$ entry of H, thus $H = X^T \beta X$

Because all entrys in $\beta \geq 0$ (proved by (*)), we can re-write H as

$$H = X^T \beta^{\frac{1}{2}} \beta^{\frac{1}{2}} X = (\beta^{\frac{1}{2}} X)^T (\beta^{\frac{1}{2}} X) \Longrightarrow H \text{ is positive definite} \Rightarrow L \text{ is convex}$$

Taylor's expansion

$$L(w) \approx L(a) + g^T(w-a) + \frac{1}{2}(w-a)^T H (w-a)$$

$$= \frac{1}{2} w^T H w + b^T w + C \qquad \text{where } b = g - Ha$$

$0 = Hw + b \Longrightarrow w^* = H^{-1} b = -H^{-1} g + a \Longrightarrow$ Therefore the minimum is unique.

由 扫描全能王 扫描创建

Question 4

$$W = \begin{bmatrix} w_{10} & w_{11} & \cdots & w_{1d-1} \\ w_{20} & w_{21} & \cdots & w_{2d-1} \\ \vdots & & & \\ w_{c0} & w_{c1} & \cdots & w_{cd-1} \end{bmatrix} w_c \quad c \times d \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1d-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2d-1} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nd-1} \end{bmatrix} x_i \quad n \times d.$$

an sample x_i belongs to class c, has 1 in the c^{th} position, and 0 in the rest of cells.

for each row (sample), $\hat{p}(y_i = c | x_i, W) = \dfrac{\exp(w_i \cdot x_i)}{\sum_{y=1}^{c} \exp(w_y \cdot x_i)}$

$$W^* = \arg\min L_i(x_i, y_i, W) = -\frac{1}{N}\sum_{i=1}^{n} t_i \log \hat{p}(y_i | x_i, w) + \lambda \|W\|^2.$$

$$= -\frac{1}{N}\sum_{i=1}^{n} t_i \log \hat{p}_i + \lambda \|W\|^2.$$

$$\frac{\partial L_i}{\partial w_{ij}} = \sum_{i=1}^{n} \frac{\partial L}{\partial w_i x_i} \frac{\partial w_i x_i}{\partial w_{ij}} + \lambda \sum_i \sum_j \frac{\partial w_{ij}^2}{\partial w_{ij}}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\hat{p}_i - t_i\right] x_{ij} + 2\lambda \sum_{i=1}^{n} w_{ij}$$

$$= \frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - t_i) x_i + 2\lambda W_i$$

Stochastic gradient. $W_{t+1} = \eta \frac{1}{N}(X^T(\hat{p} - t)) + 2\lambda W_t$

由 扫描全能王 扫描创建