DDA2020 Assignment 1

Name: 樊祥　Student ID: 120090414

T1.

$$KL(P_{emp}||q) = \sum_{x\in X} P_{emp}(x)\log\frac{P_{emp}(x)}{q(x|\theta)}$$

$$= \sum_{x\in X} P_{emp}(x)\log P_{emp}(x) - \sum_{x\in X} P_{emp}(x)\log q(x|\theta)$$

$$\operatorname{argmin}_\theta KL(P_{emp}||q) = \operatorname{argmin}_\theta - \sum_{x\in X} P_{emp}(x)\log q(x|\theta)$$

$$= \operatorname{argmax}_\theta \sum_{x\in X} P_{emp}(x)\log q(x|\theta)$$

so we just need to show that:

$$\operatorname*{argmax}_\theta \sum_{x\in X} P_{emp}(x)\log q(x|\theta) = \operatorname*{argmax}_\theta \sum_{x\in X} \log q(x|\theta).$$

Suppose n_x is the number of x samples, N is the total number, by the definition of empirical distribution, we have:

$$P_{emp}(x) = \frac{n_x}{N}$$

$$\therefore \operatorname*{argmax}_\theta \sum_{x\in X} P_{emp}(x)\log q(x|\theta) = \operatorname*{argmax}_\theta \sum_{x\in X} \frac{n_x}{N}\log q(x|\theta)$$

$$= \operatorname*{argmax}_\theta \sum_{x\in X} \log q(x|\theta)^{n_x} = \operatorname*{argmax}_\theta \prod_{x\in X} q(x|\theta)^{n_x}$$

$$= \operatorname*{argmax}_\theta \prod_{x\in X} q(x|\theta) = \operatorname*{argmax}_\theta \sum_{x\in X} \log q(x|\theta)$$

$\therefore$ Q.E.D.

T2,

$$J(w,w_0) = (y^T - w^T X^T - w_0 1^T)(y - Xw - w_0 1) + \lambda w^T w$$

$$= y^T y - y^T Xw - y^T w_0 1 - w^T X^T y + w^T X^T Xw + w^T X^T w_0 1 - w_0 1^T y + w_0 1^T Xw + w_0^2 1^T 1 + \lambda w^T w$$

$$\frac{\partial J(w,w_0)}{\partial w_0} = -y^T 1 + w^T X^T 1 - 1^T y + 1^T Xw + 2w_0 1^T 1 = 0$$

since $\bar{x} = 0 \Rightarrow X^T 1 = 0$, $1^T X = 0$

we also have: $y^T 1 = 1^T y = n\bar{y}$, $1^T 1 = n$

$\therefore \quad -2n\bar{y} + 2nw_0 = 0 \Rightarrow w_0 = \bar{y}$

$$\frac{\partial J(w,w_0)}{\partial w} = 2\lambda I w - 2X^T y + 2X^T Xw + w_0 1^T X + w_0 X^T 1 = 0$$

since $\bar{x} = 0 \Rightarrow X^T 1 = 0$, $1^T X = 0$

$\therefore \quad (\lambda I + X^T X) w = X^T y$

since $X^T X + \lambda I$ is invertible

$\therefore \quad w = (X^T X + \lambda I)^{-1} X^T y$

$\therefore$ Q.E.D.

T3, $\sum_{i=1}^{N} \log P(y_i \mid x_i, w) - \lambda \sum_{c=1}^{C} \|w_c\|_2^2$

$$= \sum_{i=1}^{N} \log \frac{\exp(w_{y_i 0} + w_{y_i}^T x_i)}{\sum_{k=1}^{C} \exp(w_{k0} + w_k^T x_i)} - \lambda \sum_{c=1}^{C} \|w_c\|_2^2$$

$$= \sum_{i=1}^{N} \left(w_{y_i 0} + w_{y_i}^T x_i - \log \sum_{k=1}^{C} \exp(w_{k0} + w_k^T x_i) \right) - \lambda \sum_{c=1}^{C} \|w_c\|_2^2$$

for any y_i, we let $w_{y_i 0} = 0$, and $w_{k0} = 0$

$\therefore J(w) = -\sum_{i=1}^{N} \left(w_{y_i}^T x_i - \log \sum_{k=1}^{C} \exp(w_k^T x_i) \right) + \lambda \sum_{c=1}^{C} \|w_c\|_2^2$

we have $w_{y_i}^T x_i = \sum_{c=1}^{C} I(y_i = c) w_c^T x_i$

$$\frac{\partial \sum_{c=1}^{C} I(y_i = c) w_c^T x_i}{\partial w_c} = I(y_i = c) x_i$$

$$\frac{\partial \log \sum_{k=1}^{C} \exp(w_k^T x_i)}{\partial w_c} = \frac{1}{\sum_{k=1}^{C} \exp(w_k^T x_i)} \cdot \frac{\partial \sum_{k=1}^{C} \exp(w_k^T x_i)}{\partial (w_c^T x_i)} \cdot \frac{\partial (w_c^T x_i)}{\partial w_c}$$

$$= \frac{\exp(w_c^T x_i)}{\sum_{k=1}^{C} \exp(w_k^T x_i)} \cdot x_i$$

$$\frac{\partial \sum_{c=1}^{C} \|w_c\|_2^2}{\partial w_c} = \frac{\partial w_c^T w_c}{\partial w_c} = 2 w_c$$

$$\frac{\partial J(w)}{\partial w_c} = -\sum_{i=1}^{N} I(y_i = c) x_i + \sum_{i=1}^{N} \frac{\exp(w_c^T x_i)}{\sum_{k=1}^{C} \exp(w_k^T x_i)} x_i + 2\lambda w_c$$

$\sum_{c=1}^{C} \frac{\partial J(w)}{\partial w_c} = 0$, $\sum_{c=1}^{C} \sum_{i=1}^{N} I(y_i = c) x_i = \sum_{i=1}^{N} x_i$

and in fact: $\frac{\exp(w_c^T x_i)}{\sum_{k=1}^{C} \exp(w_k^T x_i)} = P(y_i = c \mid x_i, w)$

$\therefore \sum_{c=1}^{C} \sum_{i=1}^{N} P(y_i = c \mid x_i, w) \cdot x_i = \sum_{i=1}^{N} x_i$ $\therefore \sum_{c=1}^{C} \frac{\partial J(w)}{\partial w_c} = 2\lambda \sum_{c=1}^{C} w_c = 0$

$\therefore \sum_{c=1}^{C} \hat{w}_c = 0 \Rightarrow \sum_{c=1}^{C} \hat{w}_{cj} = 0$ for $j = 1 : D$ 、 Q.E.D.

T4.

a, F.

$$J(w) = -\frac{1}{|D|}\sum_{i\in D}\log\sigma(y_i X_i^T w) + \lambda\|w\|_2^2$$

$$= -\frac{1}{|D|}\sum_{i\in D}\log\frac{1}{1+e^{-y_i X_i^T w}} + \lambda w^T w$$

$$= \frac{1}{|D|}\sum_{i\in D}\log(1+e^{-y_i X_i^T w}) + \lambda w^T w$$

$$\frac{\partial(w^T w)}{\partial w} = 2w \quad , \quad \frac{\partial^2(w^T w)}{\partial w^2} = 2I$$

$$\frac{\partial\log(1+e^{-y_i X_i^T w})}{\partial w} = \frac{\partial\log(1+e^{-y_i X_i^T w})}{\partial(1+e^{-y_i X_i^T w})}\cdot\frac{\partial(1+e^{-y_i X_i^T w})}{\partial w}$$

$$= \frac{1}{1+e^{-y_i X_i^T w}}\cdot e^{-y_i X_i^T w}\cdot(-y_i X_i) = -\frac{y_i X_i}{1+e^{-y_i X_i^T w}}\cdot e^{-y_i X_i^T w}$$

$$\frac{\partial^2\log(1+e^{-y_i X_i^T w})}{\partial w^2} = -y_i X_i\cdot\left(\frac{-y_i X_i e^{-y_i X_i^T w}(1+e^{-y_i X_i^T w}) - (-y_i X_i e^{-y_i X_i^T w})e^{-y_i X_i^T w}}{(1+e^{-y_i X_i^T w})^2}\right)^T$$

$$= -\frac{y_i^2 X_i X_i^T e^{-y_i X_i^T w}}{(1+e^{-y_i X_i^T w})^2}$$

$$\frac{\partial^2 J(w)}{\partial w^2} = \frac{1}{|D|}\sum_{i\in D}\frac{y_i^2 X_i X_i^T e^{-y_i X_i^T w}}{(1+e^{-y_i X_i^T w})^2} + 2\lambda I \quad , y_i\in\{-1,1\}$$

we have $\alpha_i = \frac{y_i^2 e^{-y_i X_i^T w}}{(1+e^{-y_i X_i^T w})^2} > 0$ $\therefore \frac{\partial^2 J(w)}{\partial w^2} = \frac{1}{|D|}\sum_{i\in D}\alpha_i X_i X_i^T + 2\lambda I$

$X_i X_i^T$ is positive semidefinite $\therefore J(w)$ is convex. Q.E.D.

b. F, l_2 regularized won't let $\hat{w}$ be sparse since it's "smooth" property

c. T, if the training data is linearly separable, the greater w_j is, the greater likelihood is.

d. F, when λ increases, since we minimize $J(w)$, weights become smaller, the likelihood decreases.

e. F, at first it can reduce overfitting and increase $L(w, D_{test})$ but Later $L(w, D_{test})$ decrease since weights get smaller.

T5.

$$X^T W = [x_1, x_2, \cdots, x_d] \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix} = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d$$

$$= \begin{bmatrix} w_1 x_{11} + w_2 x_{21} + \cdots + w_d x_{d1} \\ \vdots \\ w_1 x_{1n} + w_2 x_{2n} + \cdots + w_d x_{dn} \end{bmatrix}$$

$$\therefore \frac{d(X^T W)}{dw} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ & \vdots & & \\ x_{d1} & x_{d2} & \cdots & x_{dn} \end{bmatrix} = X$$

$$y^T X = [y_1, \cdots, y_d] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = y_1 x_1 + \cdots + y_d x_d$$

$$= [y_1 x_{11} + y_2 x_{21} + \cdots + y_d x_{d1}, \cdots, y_1 x_{1d} + y_2 x_{2d} + \cdots + y_d x_{dd}]$$

$$y^T X w = w_1 (y_1 x_{11} + \cdots + y_d x_{d1}) + \cdots + w_d (y_1 x_{1d} + \cdots + y_d x_{dd})$$

$$\frac{d(y^T X w)}{dw} = \begin{bmatrix} y_1 x_{11} + \cdots + y_d x_{d1} \\ \vdots \\ y_1 x_{1d} + \cdots + y_d x_{dd} \end{bmatrix} = (y^T X)^T = X^T y$$

$$w^T X = [w_1, w_2, \cdots, w_d] \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d$$

$$= [w_1 x_{11} + w_2 x_{21} + \cdots + w_d x_{d1}, \cdots, w_1 x_{1d} + w_2 x_{2d} + \cdots + w_d x_{dd}]$$

$$w^T X w = w_1^2 x_{11} + w_1 (w_2 x_{21} + \cdots + w_d x_{d1}) + \cdots +$$

$$w_d (w_1 x_{1d} + w_2 x_{2d} + \cdots + w_{d-1} x_{d-1d}) + w_d^2 x_{dd}$$

$$\therefore \frac{d(w^T X w)}{dw} = \begin{bmatrix} 2w_1 x_{11} + w_2 x_{21} + \cdots + w_d x_{d1} + w_2 x_{12} + \cdots + w_d x_{1d} \\ w_1 x_{21} + 2w_2 x_{22} + w_1 x_{12} + \cdots + w_d x_{d2} + \cdots + w_d x_{2d} \\ \vdots \\ w_1 x_{d1} + \cdots + w_1 x_{1d} + \cdots + w_{d-1} x_{d-1d} + 2w_d x_{dd} \end{bmatrix}$$

$$= \left(\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ - & - & \ddots & - \\ - & - & - & - \\ x_{d1} & x_{d2} & \cdots & x_{dd} \end{bmatrix} + \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{d1} \\ - & - & \ddots & - \\ \ddots & - & - & - \\ x_{1d} & x_{2d} & \cdots & x_{dd} \end{bmatrix} \right) \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

$$= (X + X^T) w$$