1. MLE minimizes KL divergence to the empirical distribution **(Exercise 2.15 of Murphy's book)** (1 point)

$KL(P\|Q) = \sum_{k=1}^{K} P_k \log \frac{P_k}{q_k} = \sum_{k=1}^{K} P_k \log P_k - \sum_{k=1}^{K} P_k \log q_k$

Since P is empirical distribution, then P_k are the same. So fix $P_k = P$

Then it's $\min_q kP\log P - P\sum_{k=1}^{K} \log q_k$. So we need to maximize $\sum_{k=1}^{K} \log q_k$, which is actually MLE.

2. Centering and ridge regression **(Exercise 7.3 of Murphy's book)** (1 point)

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

$\frac{dJ}{dw} = 2X^T X w + 2X^T w_0 1 - X^T y - X^T y + 2\lambda w = 0 \quad X^T w_0 1 = w_0 \sum_{i=1}^{n} x_i = 0$

$X^T X w + X^T w_0 1 - X^T y + \lambda w = 0 \quad (X^T X + \lambda I) w = X^T y, \; w = (X^T X + \lambda I)^{-1} X^T y$

$\frac{dJ}{dw_0} = 2n w_0 + 2w' \sum (x_1 + \cdots + x_n) - 2(y_1 + \cdots + y_n) = 0$ Since $\bar{X} = 0$,

it's $n w_0 = \sum_{i=1}^{n} y_i, \; w_0 = \bar{y}$.

$\frac{d^2 J}{dw^2} = X^T w + \lambda, \; \frac{d^2 J}{dw_0^2} = 2n. \; \frac{d^2 J}{dw dw_0} = 0 \Rightarrow$ Here J is the local minimum

3. Symmetric version of ℓ_2 regularized multinomial logistic regression **(Exercise 8.5 of Murphy's book)** (1 point)

$f = \sum_{i=1}^{N} \log \frac{\exp(w_{c0} + w_c^T x_i)}{\sum_{k=1}^{C} \exp(w_{k0} + w_k^T x_i)} - \lambda \sum_{c=1}^{C} \|w_c\|_2^2$

$- \log\left(1 + \sum_{k=1}^{C-1} \exp(w_{k0} + w_k^T x_i)\right)$

① $\frac{df}{dw_c} = -2\lambda w_c + \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \frac{x_i \exp(w_{c0} + w_c^T x_i)}{\sum_{k=1}^{C} \exp(w_{k0} + w_k^T x_i)}, \; \frac{df}{dw_k} (1 \le k \le C-1) = -2\lambda w_k - \sum_{i=1}^{N} \frac{x_i \exp(w_{k0} + w_k^T x_i)}{\sum_{k=1}^{C} \exp(w_{k0} + w_k^T x_i)}$

② $\sum_{k=1}^{C} \frac{df}{dw_k} = -2\lambda \left(\sum_{c=1}^{C} w_c\right) + \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i \frac{\sum_{k=1}^{C} \exp(w_{k0} + w_k^T x_i)}{\sum_{k=1}^{C} \exp(w_{k0} + w_k^T x_i)} = -2\lambda \left(\sum_{c=1}^{C} w_c\right)$, At the optimization, $\frac{df}{dw_c} = 0$, $\sum_{k=1}^{C} \frac{df}{dw_c} = 0 = -2\lambda \sum_{c=1}^{C} w_c$, $\sum_{c=1}^{C} w_c = 0$

That means $\forall j, \; \sum_{c=1}^{C} w_{cj} = 0$ for $1 \le j \le D$.

4. Elementary properties of ℓ_2 regularized logistic regression **(Exercise 8.6 of Murphy's book)** (1 point)

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

a. It's a convex function. So just one local minimum, False

b. False, It doesn't encourage reducing a subset of the weights to zero

c. True. If $\lambda = 0$, the model is easy to be overfitting.

d. False. If λ is small, then it can cause overfitting, which produce large $\ell(w, D_{train})$

e. False. Although the increase of λ eclipse the overfitting, which increase $\ell(w, D_{test})$

But if λ is too large, then the model is inflexible and the likelihood function decreases.

5. $\frac{d(\mathbf{X}^\top \mathbf{w})}{d\mathbf{w}} = \mathbf{X}, \quad X = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & & & \\ x_{1n} & & \cdots & x_{nn} \end{bmatrix} \quad X^T w = \begin{bmatrix} x_{11}w_1 + x_{21}w_2 + \cdots + x_{n1}w_n \\ x_{12}w_1 + x_{22}w_2 + \cdots + x_{n2}w_n \\ \vdots \\ x_{1n}w_1 + x_{2n}w_2 + \cdots + x_{nn}w_n \end{bmatrix}$

$\frac{dX^T w}{dw} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ & \ddots & & & \\ & & & & x_{nn} \end{bmatrix} = X$

$\frac{d(\mathbf{y}^\top \mathbf{X}\mathbf{w})}{d\mathbf{w}} = \mathbf{X}^\top \mathbf{y} \quad y^T X = [y_1 x_{11} + \cdots + y_n x_{n1}, \; y_1 x_{12} + \cdots + y_n x_{n2}, \cdots, y_1 x_{1n} + \cdots + y_n x_{nn}]$

$y^T X w = w_1(y_1 x_{11} + \cdots + y_n x_{n1}) + \cdots + w_n(y_1 x_{1n} + \cdots + y_n x_{nn})$

$\frac{dy^T X w}{dw} = \begin{bmatrix} y_1 x_{11} + \cdots + y_n x_{n1} \\ \vdots \\ y_1 x_{1n} + \cdots + y_n x_{nn} \end{bmatrix} = X^T y$

$\frac{d(\mathbf{w}^\top \mathbf{X}\mathbf{w})}{d\mathbf{w}} = (\mathbf{X} + \mathbf{X}^\top)\mathbf{w}$

$w^T X w = w_1(w_1 x_{11} + \cdots + w_n x_{n1}) + \cdots + w_n(w_1 x_{1n} + \cdots + w_n x_{nn})$

$\frac{d(w^T X w)}{dw} = \begin{pmatrix} x_{11} + \cdots + x_{n1} + x_{11} + \cdots + x_{1n} \\ x_{12} + \cdots + x_{n2} + x_{21} + \cdots + x_{2n} \\ \vdots \\ x_{1n} + \cdots + x_{nn} + x_{n1} + \cdots + x_{nn} \end{pmatrix} = Xw + X^T w = (X + X^T) w$