Assignment1 Report

Author: 余永琦 Student Number: ID 120090761

Contents

Part1: Programming Question

In this assignment, we only need to calculate the least squares solution for w. According to the equation we gain in the class, $w = (X^TX)^{-1}X^Ty$. Since we have the input X and y already, we can use np.transpose(X) to get XT, np.dot(XT,X) to get XTX and np.linalg.inv(XTX) to get InvXTX. Finally, we can multiple these to get the result w, and then return the three outputs the problem required.

Part2: Written Questions

Name: 原永琦

ID: 120090761

Assignment 1.

Written Problems.

1. $\underset{q}{\operatorname{argmin}} KL(p_{emp} \| q)$

$= \underset{q}{\operatorname{argmin}} \sum_{k=1}^{K} p_{emp\,k}(x) \log p_{emp\,k}(x) - \sum_{k=1}^{K} p_{emp\,k}(x) \cdot \log q(x)$

$\because p_{emp}(x)$ is a ~~constant~~ term that we can remove because it's not related to the parameter $q(x)$.

$\therefore \underset{q}{\operatorname{argmin}} KL(p_{emp} \| q) = \underset{q}{\operatorname{argmin}} -\sum_{k=1}^{K} \log q(x) = \underset{q}{\operatorname{argmin}} -\log q(x)$

$= \underset{q}{\operatorname{argmax}} \log q(x)$

$\therefore$ We just need to maximize $\cancel{q(x)} \cdot \log q(x)$

And for $q(x)$, we have the MLE $\hat{\theta} = \cancel{\underset{\theta}{\operatorname{argmax}} \log q(x;\theta)}$, which maximize $q(x)$.

$\underset{\theta}{\operatorname{argmax}} \log q(x|\theta)$

$\therefore \underset{q}{\operatorname{argmin}} KL(p_{emp} \| q)$ is obtain by $q(x) = q(x;\hat{\theta})$ where $\hat{\theta}$ is the MLE.

2. ① $\frac{\partial}{\partial w_0}\left[(y - Xw - w_0 \cdot 1)^T (y - Xw - w_0 1) + \lambda w^T w\right] = 0$

For the term without w_0, it will be a 0.

Then we have: $\frac{\partial}{\partial w_0}\left(m w_0^2 \cancel{\sum_{i=1}^{m} y_i} - 2w_0 \sum_{i=1}^{m} y_i + \cancel{2w_0 \sum_{i=1}^{m}\sum_{j=1}^{d} x_{ij} w_j}\ 2w_0 \sum_{i=1}^{m} x_i w \right) = 0$

$$2mw_0 - 2\sum_{i=1}^{m} y_i + \cancel{2w_0 \sum_{i=1}^{m}\sum_{j=1}^{d} x_{ij} w_j} + 2w_0 \sum_{i=1}^{m} x_i w = 0$$

Since $\bar{x} = 0$, we can know that $\cancel{\sum_{i=1}^{m}\sum_{j=1}^{d} x_{ij} w_j} \quad \sum_{i=1}^{m} x_i w = 0$

$$\therefore\ 2mw_0 = 2\sum_{i=1}^{m} y_i$$

$$w_0 = \frac{1}{m}\sum_{i=1}^{m} y_i = \bar{y}$$, where m is the number of data points.

② $\frac{\partial}{\partial w}\left[(y - Xw - w_0 1)^T (y - Xw - w_0 \cdot 1) + \lambda w^T w\right] = 0$

Also we drop the term without w.

Then, we got: $\frac{\partial}{\partial w}\left(w^T X^T X w - y^T X w - w^T X^T y + w^T X^T 1 \cdot w_0 + 1^T X w \cdot w_0 + \lambda w^T w\right) = 0$

$$2X^T X w - 2X^T y + 2w_0 \cdot X^T \cdot 1 + 2\lambda I w = 0$$

$\because \bar{x} = 0$

$\therefore X^T \cdot 1 = 0$

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

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

3. Let $y_{ic} = \mathbb{I}(y_i = c)$, $\mu_{ic} = P(y_i = c \mid x, W) = \frac{\exp(w_0 + w_c^T x)}{\sum_{k=1}^{C} \exp(w_{k0} + w_k^T x)}$

$$\sum_{i=1}^{N} \log p(y_i \mid x_i, W) = \cancel{\sum_{i=1}^{N} y_{ic} \log \mu_{ic}} \quad \sum_{i=1}^{N} \sum_{c=1}^{C} y_{ic} \log \mu_{ic}$$

$$= \sum_{i=1}^{N} \left[\sum_{c=1}^{C} y_{ic} w_c^T x_i - \log \left(\sum_{c=1}^{C} \exp(w_c^T x_i) \right) \right]$$

$$\therefore \nabla_{w_c} \left(\sum_{i=1}^{N} \log p(y_i \mid x_i, W) - \lambda \sum_{c=1}^{C} \|w_c\|_2^2 \right)$$

$$= \sum_{i=1}^{N} (y_{ic} - \mu_{ic}) x_i - 2\lambda w_c.$$

$\because$ At the optimum, we have $\sum_{c=1}^{C} \nabla_{w_c} \left(\sum_{i=1}^{N} \log p(y_i \mid x_i, W) - \lambda \sum_{c=1}^{C} \|w_c\|_2^2 \right) = 0$

$$\therefore \sum_{i=1}^{N} \sum_{c=1}^{C} (y_{ic} - \mu_{ic}) x_i - \cancel{2\lambda w_c} \, 2\lambda \sum_{c=1}^{C} w_c = 0.$$

$\because \sum_{c=1}^{C} y_{ic} = 1$, $\sum_{c=1}^{C} \mu_{ic} = 1$

$\cancel{\therefore -2\lambda w_c = 0 \Rightarrow w_c = 0}$ $\therefore -2\lambda \sum_{c=1}^{C} w_c = 0 \Rightarrow \sum_{c=1}^{C} w_c = 0$

$\therefore \sum_{c=1}^{C} w_{cj} = 0$ for $j = 1:D$.

4. a. F b. F c. T d. F e. F

5. ① Suppose the size of X is $d \times h$ and w is $d \times 1$

$$X^T w = \begin{bmatrix} X_{11}W_1 + X_{21}W_2 + \cdots + X_{d1}W_d \\ X_{12}W_1 + X_{22}W_2 + \cdots + X_{d2}W_d \\ \vdots \\ X_{1h}W_1 + X_{2h}W_2 + \cdots + X_{dh}W_d \end{bmatrix}$$

By definition, $$\frac{d(X^T w)}{d(w)} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1h} \\ X_{21} & X_{22} & \cdots & X_{2h} \\ \vdots & \vdots & \ddots & \vdots \\ X_{d1} & X_{d2} & \cdots & X_{dh} \end{bmatrix} = X$$

② Let the size of y is $m \times 1$, X is $m \times d$. w is $d \times 1$.

Then, $$y^T X = \begin{bmatrix} y_1 X_{11} + y_2 X_{21} \cdots + y_m X_{m1} \\ y_1 X_{12} + y_2 X_{22} \cdots + y_m X_{m2} \\ \vdots \\ y_1 X_{1d} + y_2 X_{2d} + \cdots + y_m X_{md} \end{bmatrix}^T$$

$$y^T X w = W_1 (y_1 X_{11} + \cdots + y_m X_{m1}) + W_2 (y_1 X_{12} + \cdots + y_m X_{m2}) + \cdots + W_d (y_1 X_{1d} + \cdots + y_m X_{md})$$

By definition, $$\frac{d(y^T X w)}{dw} = \begin{bmatrix} y_1 X_{11} + y_2 X_{21} + \cdots + y_m X_{m1} \\ y_1 X_{12} + y_2 X_{22} + \cdots + y_m X_{m2} \\ \vdots \\ y_1 X_{1d} + y_2 X_{2d} + \cdots y_m X_{md} \end{bmatrix} = (y^T X)^T = X^T y$$

③ Suppose size of X is $d \times d$ and w is $d \times 1$.

$$w^T X w = X_{11} W_1^2 + X_{12} W_1 W_2 + \cdots + X_{dd} W_d^2 = \sum_{i=1}^{d} \sum_{j=1}^{d} X_{ij} W_i W_j$$

By definition, $$\frac{d(w^T X w)}{d(w)} = \begin{bmatrix} 2X_{11} W_1 + X_{12} W_2 + X_{21} W_2 + \cdots + X_{1d} W_d + X_{d1} W_d \\ \vdots \\ 2X_{dd} W_d + \cdots + X_{1d} W_1 + X_{d1} W_1 \end{bmatrix}$$

$$= \begin{bmatrix} X_{11} W_1 + X_{12} W_2 + \cdots + X_{1d} W_d \\ \vdots \\ X_{d1} W_1 + X_{d2} W_2 + \cdots + X_{dd} W_d \end{bmatrix} + \begin{bmatrix} X_{11} W_1 + X_{21} W_2 + \cdots + W_{d1} W_d \\ \vdots \\ X_{1d} W_1 + X_{2d} W_2 + \cdots + W_{dd} W_d \end{bmatrix} = (X + X^T) w$$