DDA3020 Machine Learning Spring 2023
Produced by TA Teams

Assignment 1

1 Problem 1.1

1.1 Problem 1.1.1

If $X \in \mathbb{R}^{d\times h}, w \in \mathbb{R}^{d\times 1}$, we have

$$X^T w = \begin{bmatrix} x_{1,1}w_1 + x_{2,1}w_2 + \cdots + x_{d,1}w_d \\ x_{1,2}w_1 + x_{2,2}w_2 + \cdots + x_{d,2}w_d \\ \vdots \\ x_{1,h}w_1 + x_{2,h}w_2 + \cdots + x_{d,h}w_d \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_h \end{bmatrix}$$

$$\frac{d(X^T w)}{dw} = \begin{bmatrix} \frac{\partial f_1}{\partial w_1} & \cdots & \frac{\partial f_h}{\partial w_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial w_d} & \cdots & \frac{\partial f_h}{\partial w_d} \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,h} \\ \vdots & \ddots & \vdots \\ x_{d,1} & \cdots & x_{d,d} \end{bmatrix} = X$$

Therefore, for this problem, note that $X \in \mathbb{R}^{h\times d}$ is a matrix, $w \in \mathbb{R}^{d\times 1}$ and $y \in \mathbb{R}^{h\times 1}$ are vectors. We have

$$\frac{d\left(y^T X w\right)}{dw} = \frac{d\left(\left(y^T X\right) w\right)}{dw} = \left(y^T X\right)^T = X^T y$$

1.2 Problem 1.1.2

Note that $w \in \mathbb{R}^{d\times 1}$. We have

$$w^T w = w_1^2 + w_2^2 + \cdots + w_d^2$$

Therefore

$$\frac{d(w^T w)}{dw} = \begin{bmatrix} 2w_1 \\ 2w_2 \\ \vdots \\ 2w_d \end{bmatrix} = 2w$$

1.3 Problem 1.1.3

Note that $X \in \mathbb{R}^{d\times d}$ and $w \in \mathbb{R}^{d\times 1}$, we have

$$\begin{aligned} w^T Xw = & w_1 \left(x_{1,1}w_1 + x_{1,2}w_2 + \cdots + x_{1,d}w_d\right) \\ & + w_2 \left(x_{2,1}w_1 + x_{2,2}w_2 + \cdots + x_{2,d}w_d\right) \\ & + \cdots \\ & + w_d \left(x_{d,1}w_1 + x_{d,2}w_2 + \cdots + x_{d,d}w_d\right). \end{aligned}$$

Define $f = w^T Xw$, we have

$$\frac{d\left(w^T Xw\right)}{dw} = \begin{bmatrix} \frac{\partial f}{\partial w_1} \\ \vdots \\ \frac{\partial f}{\partial w_d} \end{bmatrix} = \begin{bmatrix} x_{1,1}w_1 + \cdots + x_{1,d}w_d + x_{1,1}w_1 + \cdots + x_{d,1}w_d \\ x_{2,1}w_1 + \cdots + x_{2,d}w_d + x_{1,2}w_1 + \cdots + x_{d,2}w_d \\ \vdots \\ x_{d,1}w_1 + \cdots + x_{d,d}w_d + x_{1,d}w_1 + \cdots + x_{d,d}w_d \end{bmatrix} = \left(X + X^T\right) w$$

Alternative Methods:

Use the rule of differential of composite function $\frac{d\left(u^T v\right)}{dw} = \frac{du^T}{dw}v + \frac{dv^T}{dw}u$. (You can find more details in Matrix Cookbook [1] Chapter 2.4).

Define $u = w$ and $v = Xw$, we have

$$\begin{aligned} \frac{d\left(w^T Xw\right)}{dw} &= \frac{d\left(u^T v\right)}{dw} \\ &= \frac{du^T}{dw}v + \frac{dv^T}{dw}u \\ &= \frac{dw^T}{dw}Xw + \frac{d\left(w^T X^T\right)}{dw}w \\ &= Xw + X^T w \\ &= (X + X^T)w, \end{aligned}$$

where I is identity matrix.

2 Problem 1.2

2.1 Problem 1.2.1

$$\begin{aligned} min_{w,b}(Xw-y)^T(Xw-y) + \lambda\overline{w}^T\overline{w} \\ \frac{\partial}{\partial w}(Xw-y)^T(Xw-y) + \lambda\overline{w}^T\overline{w} = 0 \\ 2X^t Xw - 2X^T y + 2\lambda\hat{I}_d w = 0 \\ X^T Xw + \lambda\hat{I}_d w = x^T y \\ (X^T X + \lambda\hat{I}_d)w = x^T y \\ w = (X^T X + \lambda\hat{I}_d)^{-1}X^T y \end{aligned}$$

Note that $X^T X + \lambda \hat{I}_d$ is guaranteed to be invertible given $\lambda > 0$

2.2 Problem 1.2.2

(You can get points if your answer is reasonable.)

1. Initialize: Choose learning rate β and iteration T, and initialize $t = 0$ and $\mathbf{W}_0$.
2. Update parameters $\mathbf{W}$ with gradient descent
$$\mathbf{W}_{t+1} = \mathbf{W}_t - \beta \frac{\partial J(\mathbf{W})}{\partial \mathbf{W}}$$
3. Repeat the last step for T-times iteration untill convergence.
4. The final parameters are $\mathbf{W}_T$.

3 Problem 1.3

3.1 Problem 1.3.1

Pick x_1, x_2 so that $x_1 \neq x_2$, and pick $\lambda \in (0, 1)$.

$$\begin{aligned} f\left((1-\lambda)x_1 + \lambda x_2\right) &= \left((1-\lambda)x_1 + \lambda x_2\right)^2 \\ &= (1-\lambda)^2 x_1^2 + \lambda^2 x_2^2 + 2(1-\lambda)\lambda x_1 x_2 \end{aligned}$$

What to do from here? Since $x_1 \neq x_2, (x_1 - x_2)^2 > 0$. Expanding, this means that $x_1^2 + x_2^2 > 2x_1x_2$. This means that

$$\begin{aligned} (1-\lambda)^2 x_1^2 + \lambda^2 x_2^2 + 2(1-\lambda)\lambda x_1 x_2 &< (1-\lambda)^2 x_1^2 + \lambda^2 x_2^2 + (1-\lambda)(\lambda)\left(x_1^2 + x_2^2\right) \\ &= \left(1 - 2\lambda - \lambda^2 + \lambda + \lambda^2\right) x_1^2 + \left(\lambda - \lambda^2 + \lambda^2\right) x_2^2 \\ &= (1-\lambda)x_1^2 + \lambda x_2^2 \\ &= (1-\lambda) f(x_1) + \lambda f(x_2) \end{aligned}$$

which proves strict convexity

3.2 Problem 1.3.2

$$\begin{aligned} f\left((1-\lambda)x_1 + \lambda x_2\right) &= a\left((1-\lambda)x_1 + \lambda x_2\right) + b \\ &= a\left((1-\lambda)x_1 + \lambda x_2\right) + ((1-\lambda) + \lambda)b \\ &= (1-\lambda)\left(ax_1 + b\right) + \lambda\left(ax_2 + b\right) \\ &= (1-\lambda) f(x_1) + \lambda f(x_2) \end{aligned}$$

So we see that inequality is in fact satisfed as an equality. So every affine function is convex. However, this means we can't replace the inequality $\leq$ with the strict inequality $<$, so affine functions are not strictly convex.

3.3 Problem 1.3.3

$$\begin{aligned} f\left((1-\lambda)x_1 + \lambda x_2\right) &= |(1-\lambda)x_1 + \lambda x_2| \\ &\leq |(1-\lambda)x_1| + |\lambda x_2| \quad \text{ by the triangle inequality} \\ &= (1-\lambda)\,|x_1| + \lambda\,|x_2| \quad \text{ because } \lambda, 1-\lambda \geq 0 \\ &= (1-\lambda) f\left(x_1\right) + \lambda f\left(x_2\right) \end{aligned}$$

Therefore f is convex. To show that it is strictly convex, we would have to show that the inequality

$$|(1-\lambda)x_1 + \lambda x_2| \leq |(1-\lambda)x_1| + |\lambda x_2|$$

can be replaced by a strict inequality $<$. However, we can't do this: for example, if $x_1 = 1$, $x_2 = 2, \lambda = 0.5$, the left side of the inequality ($|1/2 + 2/2| = 3/2$) is exactly equal to the right side ($|1/2| + |2/2| = 3/2$). So f is not strictly convex.

4 Problem 1.4

The probability density function of Laplace distribution is given by:

$$f(x_i|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|x_i - \mu|}{b}\right)$$

The distribution and the log-likelihood of a mixture of Laplace are given by:

$$p(x|\mu, b) = \prod_{i=1}^{m} \frac{1}{2b} \exp\left(-\frac{|x_i - \mu|}{b}\right)$$

$$L = \log p(x|\mu, b) = -n(\log(2) + \log(b)) - \sum_i \frac{|x_i - \mu|}{b}$$

We can get the MLE of μ and b by computing $\frac{\partial L}{\partial \mu} = 0$ and $\frac{\partial L}{\partial b} = 0$.

$$\begin{aligned} \frac{\partial L}{\partial \mu} &= -\frac{1}{b} \sum_i \frac{\partial}{\partial \mu} |x_i - \mu| \\ &= \frac{1}{b} \sum_i \operatorname{sign}(x_i - \mu) \end{aligned}$$

Therefore, by $\frac{\partial L}{\partial \mu} = 0$, we need to get the sample median to balance the signs of $(x_i - \mu)$.

$$\mu_{MLE} = \text{median}(x_1, x_2, ..., x_n)$$

Next, let's consider the derivative w.r.t. b.

$$\frac{\partial L}{\partial b} = -\frac{n}{b} + \sum_i \frac{|x_i - \mu_{MLE}|}{b^2}$$

By $\frac{\partial L}{\partial b} = 0$:

$$b_{MLE} = \frac{1}{n} \sum_i |x_i - \mu_{MLE}|$$

References

[1] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.