Machine Learning **(Due: 8th April)**

Assignment #2 (Linear Model)

Instructor: Beilun Wang *Name:* Qipeng Zhu, *ID:* 58119304

Problem Description:

Problem 1: Linear Regression

Give data set $\boldsymbol{X} = (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \cdots, \boldsymbol{x}^{(n)})^\top$ and $\boldsymbol{y} = (y^{(1)}, y^{(2)}, \cdots, y^{(n)})^\top$ where $(\boldsymbol{x}^{(i)\top}, y^{(i)}) = (x_1^{(i)}, x_2^{(i)}, \cdots, x_p^{(i)}, y^{(i)})$ is the i-th observation. We focus on the model $y = \boldsymbol{\theta}^\top \boldsymbol{x} + \varepsilon$.

(1) Assuming $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, write down the log-likelihood function of $\boldsymbol{y}$. You can ignore any unnecessary constants.

(2) Based on your answer to (1), show that finding Maximum Likelihood Estimate of $\boldsymbol{\theta}$ is equivalent to solving $\operatorname{argmin}_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2$.

(3) Prove that $\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}$ with $\lambda > 0$ is Positive Definite(Hint: definition).

(4) Show that $\boldsymbol{\theta}^* = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ is the solution to $\operatorname{argmin}_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|^2$.

(5) Assuming $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and $\theta_i \sim \mathcal{N}(0, \tau^2)$ for $i = 1, 2, \cdots, p$ in $\boldsymbol{\theta}$($\boldsymbol{\theta}$ does not vary in each sample), write down the estimate of $\boldsymbol{\theta}$ by maximizing the conditional distribution $f(\boldsymbol{\theta} \mid \boldsymbol{y})$(Hint: Bayes' formula). You can ignore any unnecessary constants. Also find out the relationship between your estimate and the solution in (4).

Problem 2: Gradient Descent

Continuously differentiable function $f : \mathbb{R} \mapsto \mathbb{R}$ is called β-**smooth** when its derivative f' is β-**Lipschitz**, which for $\beta > 0$ implies that

$$|f'(x) - f'(y)| \leqslant \beta |x - y|.$$

Now suppose f is β-**smooth** and **convex** as a loss function in a gradient descent problem.

(1) Prove that

$$f(y) - f(x) \leqslant f'(x)(y - x) + \frac{\beta}{2}(y - x)^2.$$

(Hint: Newton-Leibniz formula.)

(2) Give $x_{k+1} = x_k - \eta f'(x_k)$ as one step of GD. Prove that

$$f(x_{k+1}) \leqslant f(x_k) - \eta(1 - \frac{\eta\beta}{2})(f'(x_k))^2.$$

(3) Based on (2), let $\eta = 1/\beta$ and assume the unique global minimum point x^* of f exists. Prove that

$$\lim_{k\to\infty} f'(x_k) = 0, \ \lim_{k\to\infty} x_k = x^*.$$

(Hint: show that for $K \in \mathbb{N}_+$, $\sum_{k=1}^{K} (f'(x_k))^2 \leqslant 2\beta(f(x_1) - f(x_{K+1}))$.)

(4) Recall one of the properties of convex function: $f(y) \geqslant f(x) + f'(x)(y - x)$. Prove that

$$f(y) - f(x) \geqslant f'(x)(y - x) + \frac{1}{2\beta}(f'(y) - f'(x))^2.$$

(Hint: let $z = y - \frac{1}{\beta}(f'(y) - f'(x))$.)

Problem 3: Kernel functions

Kernel function $k : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$ is called **Positive Semi-Definite**(**PSD**) when its Gramian matrix K is PSD, where $K_{ij} = k(\boldsymbol{u}_i, \boldsymbol{u}_j)$ for any group of vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_n \in \mathbb{R}^p$. Let k_1 and k_2 be two PSD kernels.

(1) Give a function $f : \mathbb{R}^p \mapsto \mathbb{R}$. Show that the kernel k defined by $k(\boldsymbol{u}, \boldsymbol{v}) = f(\boldsymbol{u})f(\boldsymbol{v})$ is PSD.

(2) Show that the kernel k defined by $k(\boldsymbol{u}, \boldsymbol{v}) = k_1(\boldsymbol{u}, \boldsymbol{v})k_2(\boldsymbol{u}, \boldsymbol{v})$ is PSD. (Hint: consider about the Hadamard product and eigendecomposition.)

(3) Give P as a polynomial with non-negative coefficients(e.g., $P(x) = \sum_i a_i x^i$ with $a_i \geqslant 0$). Show that the kernel k defined by $k(\boldsymbol{u}, \boldsymbol{v}) = P(k_1(\boldsymbol{u}, \boldsymbol{v}))$ is PSD.

(4) Show that the kernel k defined by $k(\boldsymbol{u}, \boldsymbol{v}) = \exp(k_1(\boldsymbol{u}, \boldsymbol{v}))$ is PSD. (Hint: use the series expansion.)

Answer:

Problem 1: Linear Regression

(1) Because $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, and $y = \boldsymbol{\theta}^\top \boldsymbol{x} + \varepsilon$, we can find $(y - \theta^\top \boldsymbol{x}) \sim \mathcal{N}(0, \sigma^2)$. So

$$f(y|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{2\sqrt{2\pi}\sigma} e^{-\frac{(y-\boldsymbol{\theta}^\top \boldsymbol{x})^2}{2\sigma^2}}$$

$$\begin{aligned} f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) &= \prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \\ &= \frac{1}{(2\sqrt{2\pi}\sigma)^n} e^{-\frac{\sum_{i=1}^{n}(y-\boldsymbol{\theta}^\top \boldsymbol{x})^2}{2\sigma^2}} \\ &= Ce^{-\frac{(\boldsymbol{X\theta}-\boldsymbol{y})^\top(\boldsymbol{X\theta}-\boldsymbol{y})}{2\sigma^2}} \qquad \text{(C is a constant)} \end{aligned}$$

So the log-likelihood function of $\boldsymbol{y}$ is:

$$\begin{aligned} lnf(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) &= -\frac{(\boldsymbol{X\theta}-\boldsymbol{y})^\top(\boldsymbol{X\theta}-\boldsymbol{y})}{2\sigma^2} \\ &= (\boldsymbol{X\theta}-\boldsymbol{y})^\top(\boldsymbol{X\theta}-\boldsymbol{y}) \end{aligned}$$

(2) Suppose

$$L(\boldsymbol{\theta}) = lnf(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})$$

So

$$\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = 2\boldsymbol{X}^T\boldsymbol{X\theta} - 2\boldsymbol{X}^\top \boldsymbol{y}$$

We let

$$\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = 0$$

then we can get

$$\boldsymbol{\theta}_{ML} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

So $\boldsymbol{\theta}_{ML}$ is the solution of $\operatorname{argmin}_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{X\theta}\|^2$

(3) For any $\boldsymbol{v} \neq 0$:

$$\boldsymbol{v}^\top(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{v} = \boldsymbol{v}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v} + \lambda\boldsymbol{v}^\top\boldsymbol{v} = ||\boldsymbol{Xv}||_0^2 > 0$$

So $\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}$ with $\lambda > 0$ is Positive Definite.

(4) Suppose

$$J(\boldsymbol{\theta}) = \|\boldsymbol{y} - \boldsymbol{X\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|^2$$

First we can simplify the optimization:

$$J(\boldsymbol{\theta}) = (\boldsymbol{y} - \boldsymbol{X\theta})^T(\boldsymbol{y} - \boldsymbol{X\theta}) + \lambda\theta^T\theta = \boldsymbol{\theta}^\top\boldsymbol{X\theta} - 2\boldsymbol{\theta}^\top\boldsymbol{X}^\top\boldsymbol{y} + \boldsymbol{y}^\top\boldsymbol{y} + \lambda\boldsymbol{\theta}^\top\boldsymbol{\theta}$$

so its gradient is:

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = 2\boldsymbol{X}^\top\boldsymbol{X\theta} - 2\boldsymbol{X}^\top\boldsymbol{y} + 2\lambda\boldsymbol{\theta}$$

let

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = 0$$

we can get

$$\boldsymbol{\theta}^* = (\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

Because

$$\text{argmin}_{\boldsymbol{\theta}} \ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|^2$$

is a convex optimization, $\boldsymbol{\theta}^*$ is its solution.

(5) From problem(1), we can know

$$f(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) = \frac{1}{(2\sqrt{2\pi}\sigma)^n} e^{-\frac{(\boldsymbol{X}\boldsymbol{\theta}-\boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{\theta}-\boldsymbol{y})}{2\sigma^2}}$$

Because

$$\theta_i \sim \mathcal{N}(0,\tau^2)\text{for i=1,2,}\cdots\text{,p}$$

we can know

$$f(\boldsymbol{\theta}) = \prod_{i=1}^{p} f(\theta_i) = \frac{1}{(2\sqrt{2\pi}\tau)^p} e^{-\frac{\boldsymbol{\theta}^T\boldsymbol{\theta}}{2\tau^2}}$$

Then

$$f(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})}{\int_{-\infty}^{+\infty} f(\boldsymbol{y}|\boldsymbol{\theta})f(\boldsymbol{\theta})dx} = \alpha e^{-(\frac{(\boldsymbol{X}\boldsymbol{\theta}-\boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{\theta}-\boldsymbol{y})}{2\sigma^2}+\frac{\boldsymbol{\theta}^T\boldsymbol{\theta}}{2\tau^2})}$$

$$\text{lnf}(\boldsymbol{\theta}|\boldsymbol{y}) = \text{ln}\alpha - (\frac{(\boldsymbol{X}\boldsymbol{\theta}-\boldsymbol{y})^T(\boldsymbol{X}\boldsymbol{\theta}-\boldsymbol{y})}{2\sigma^2} + \frac{\boldsymbol{\theta}^T\boldsymbol{\theta}}{2\tau^2})$$

Let

$$\nabla_{\boldsymbol{\theta}} lnf(\boldsymbol{\theta}|\boldsymbol{y}) = -2(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta} - \boldsymbol{X}^T\boldsymbol{y} + (\frac{\sigma^2}{\tau^2})\boldsymbol{\theta}) = 0$$

we can get

$$(\boldsymbol{X}^T\boldsymbol{X} + \frac{\sigma^2}{\tau^2}\boldsymbol{I})\boldsymbol{\theta} = \boldsymbol{X}^T\boldsymbol{y}$$

Finally, we can get

$$\boldsymbol{\theta}^* = (\boldsymbol{X}^T\boldsymbol{X} + \frac{\sigma^2}{\tau^2}\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

When we suppose $\lambda = \frac{\sigma^2}{\tau^2}$, the solution is the same as the solution of problem(4)

Problem 2: Gradient Descent

(1) From Lagrange theorem:$\forall x, y \in \mathbb{R}, \ \exists\xi$ between x and y,s.t

$$f''(\xi) = \frac{f'(y) - f'(x)}{y - x}$$

So

$$|f''(\xi)| = |\frac{f'(y) - f'(x)}{y - x}| \leqslant \beta$$

For f(y), we can use taylor expansion at x, we can get

$$f(y) = f(x) + f'(x)(y-x) + \frac{f''(\xi)}{2}(y-x)^2(\xi \in (x,y))$$

So

$$f(y) - f(x) \leqslant f'(x)(y-x) + \frac{|f''(\xi)|}{2}(y-x)^2 \leqslant f'(x)(y-x) + \frac{\beta}{2}(y-x)^2$$

(2) From problem(1), we can easily get:

$$f(x_{k+1}) - f(x_k) \leqslant f'(x_k)(x_{k+1} - x_k) + \frac{\beta}{2}(x_{k+1} - x_k)^2$$

$$= f'(x_k)(-\eta f'(x_k)) + \frac{\beta}{2}\eta^2(f'(x_k))^2 = -\eta(1-\frac{\eta\beta}{2})(f'(x_k))^2$$

(3) From problem to we can get,

$$\exists n \in N_+, s.t.\ f(x_{n+1}) \leq f(x_n) - \frac{1}{2\beta}(f'(x_n))^2$$

$$\begin{cases} f(x_{n+1}) \leq f(x_n) - \frac{1}{2\beta}(f'(x_n))^2 & \text{(n)} \\ f(x_n) \leq f(x_{n-1}) - \frac{1}{2\beta}(f'(x_{n-1}))^2 & \text{(n-1)} \\ \qquad\vdots & \\ f(x_2) \leq f(x_1) - \frac{1}{2\beta}(f'(x_1))^2 & \text{(1)} \end{cases}$$

Add up these n formulas, we can get:

$$\sum_{k=1}^{n}(f'(x_k))^2 \leqslant 2\beta(f(x_1) - f(x_{n+1}))$$

because $f(x^*)$ is the minimum, so $f(x_{n+1}) \geqslant f(x^*)$, then

$$\sum_{k=1}^{n}(f'(x_k))^2 \leqslant 2\beta(f(x_1) - f(x^*)) = c(\text{c is a constant})$$

It's easy to know $\sum_{k=1}^{n}(f'(x_k))^2$ is monotonic increasing, and bounded, which means it can converge. So

$$\lim_{k\to\infty} f'(x_k) = 0$$

$$\text{And } \lim_{k\to\infty} x_k \text{ is solution of f'(x)=0,}$$

we can get

$$\lim_{k\to\infty} x_k$$

is extreme point. Because f(x) is convex, so its local minimum is global minimum. So f(x)

$$\lim_{k\to\infty} x_k = x^*$$

.

(4) Suppose:

$$z = y - \frac{1}{\beta}(f'(y) - f'(x))$$

Because f(x) is convex:

$$f(x) - f(z) \leqslant f'(x)(x-z) = f'(x)(x-y) + f'(x)(y-z) \qquad (1)$$

From proble(1), we can know:

$$f(z) - f(y) \leqslant f'(y)(z-y) + \frac{\beta}{2}(z-y)^2 = -f'(y)(y-z) + \frac{\beta}{2}(y-z)^2 \qquad (2)$$

Let (1) + (2), we can get:

$$f(x) - f(y) \leqslant f'(x)(x-y) + (f'(x) - f'(y))(y-z) + \frac{\beta}{2}(y-z)^2$$

Substituting $y - z = \frac{1}{\beta}(f'(y) - f'(x))$

$$f(x) - f(y) \leqslant f'(x)(x-y) + \frac{1}{\beta}(f'(x) - f'(y))(f'(y) - f'(x)) - \frac{1}{2\beta}(f'(y) - f'(x))^2$$

So

$$f(y) - f(x) \leqslant f'(x)(y-x) + \frac{1}{2\beta}(f'(y) - f'(x))^2$$

Problem 3: Kernel functions

(1) Suppose $\boldsymbol{K}$ is the Gramian matrix of $k(\boldsymbol{u}, \boldsymbol{v})$, We can easily know:

$$\mathbf{K} = \begin{pmatrix} f(\boldsymbol{u}_1)f(\boldsymbol{u}_1) & f(\boldsymbol{u}_1)f(\boldsymbol{u}_2) & \dots \\ f(\boldsymbol{u}_2)f(\boldsymbol{u}_1) & f(\boldsymbol{u}_2)f(\boldsymbol{u}_2) & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

$$= (f(\boldsymbol{u}_1), f(\boldsymbol{u}_2), \dots, f(\boldsymbol{u}_n))^T \cdot (f(\boldsymbol{u}_1, f(\boldsymbol{u}_2), \dots, f(\boldsymbol{u}_n))$$

For any given $\boldsymbol{v} \in \mathbb{R}^p$, we have:

$$\boldsymbol{v}\boldsymbol{K}\boldsymbol{v}^T = \boldsymbol{v}(f(\boldsymbol{u}_1), f(\boldsymbol{u}_2), \dots, f(\boldsymbol{u}_n))^T \cdot (f(\boldsymbol{u}_1, f(\boldsymbol{u}_2), \dots, f(\boldsymbol{u}_n))\boldsymbol{v}^T$$

$$= ||\boldsymbol{v} \cdot (f(\boldsymbol{u}_1), f(\boldsymbol{u}_2), \dots, f(\boldsymbol{u}_n)^T||_2^2 \geqslant 0$$

So k$(\boldsymbol{u}, \boldsymbol{v})$ is PSD kernels.

(2) Suppose $\boldsymbol{A}, \boldsymbol{B} and \boldsymbol{C}$ is respectively the Gramian matrix of k_1, k_2, k, because:

$$k(\boldsymbol{u}, \boldsymbol{v}) = k_1(\boldsymbol{u})k_2(\boldsymbol{v})$$

We can easily know $c_{ij} = a_{ij}b_{ij}$, so we can get:

$$\mathbf{C} = \mathbf{A} \circ \mathbf{B}$$

According the the nature of Hadamard product:

$$\lambda_{min}(\mathbf{C}) = \lambda_{min}(\mathbf{A} \circ \mathbf{B}) \geqslant \lambda_{min}(\mathbf{A})\lambda_{min}(\mathbf{B})$$

Because $\mathbf{A}$ and $\mathbf{B}$ are PSD, so $\lambda_{min}(\mathbf{A}) \geqslant 0, \lambda_{min}(\mathbf{B}) \geqslant 0$, so we can get:

$$\lambda_{min}(\mathbf{C}) \geqslant 0$$

So $\boldsymbol{C}$ is PSD, and $k(\boldsymbol{u}, \boldsymbol{v})$is PSD.

(3) Suppose $m_i(\boldsymbol{u}, \boldsymbol{v}) = k_1^i(\boldsymbol{u}, \boldsymbol{v})$:

$$\begin{cases} m_1(\boldsymbol{u}, \boldsymbol{v}) = k_1(\boldsymbol{u}, \boldsymbol{v}) \\ m_2(\boldsymbol{u}, \boldsymbol{v}) = k_2^2(\boldsymbol{u}, \boldsymbol{v}) \\ \vdots \\ m_n(\boldsymbol{u}, \boldsymbol{v}) = k_1^n(\boldsymbol{u}, \boldsymbol{v}) \end{cases}$$

When i = 1, we can easily get $m_1(\boldsymbol{u}, \boldsymbol{v})$ is PSD.
Suppose $m_n(\boldsymbol{u}, \boldsymbol{v})$ is PSD;
When i = n+1; For $m_{n+1}(\boldsymbol{u}, \boldsymbol{v})$, we have:

$$m_{n+1}(\boldsymbol{u}, \boldsymbol{v}) = m_n(\boldsymbol{u}, \boldsymbol{v}) * m_1(\boldsymbol{u}, \boldsymbol{v})$$

From problem(2) and hypothesis, we can know $m_{n+1}(\boldsymbol{u}, \boldsymbol{v})$ is PSD. So for any i $\in \mathbb{N}$, we have $m_i(\boldsymbol{u}, \boldsymbol{v})$ is PSD. Suppose $\boldsymbol{M}_i$ is the Gramian matrix of m_i, so M_i is PSD, for any $i \in \mathbb{N}$. Then we have:

$$\lambda_{min}(M_i) \geqslant 0, \text{for any i} \in \mathbb{N}$$

Because
$$k(\boldsymbol{u},\boldsymbol{v}) = P(k_1(\boldsymbol{u},\boldsymbol{v})) = \sum_i a_i m_i(\boldsymbol{u},\boldsymbol{v})$$

Suppose $\boldsymbol{K}$ is the Gramian matrix of $k(\boldsymbol{u},\boldsymbol{v})$, so
$$\boldsymbol{K} = \sum_i a_i \boldsymbol{M}_i \Longrightarrow \lambda(\boldsymbol{K}) = \sum_i a_i \lambda(\boldsymbol{M}_i) \geqslant 0$$

So $\boldsymbol{K}$ is PSD, and $k(\boldsymbol{u},\boldsymbol{v})$ is PSD.

(4) From Maclaurin series expansion,
$$exp(k_1(\boldsymbol{u},\boldsymbol{v})) = \sum_{i=0}^{\infty} \frac{f^{(i)}(x^*)}{i!} x^i = \sum_{i=0}^{\infty} \frac{exp(x^*)}{i!} x^i \quad (x^* = k_1(\boldsymbol{u},\boldsymbol{v}))$$

It's clear that $exp(x^*) \geqslant 0$ for any $i \in \mathbb{N}$. So from problem(3), we can get $k(\boldsymbol{u},\boldsymbol{v})$ is PSD.