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

Assignment #2 (Linear Model)

Instructor: Beilun Wang *Name:* Li Haorui, *ID:* 61518407

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 $f(\boldsymbol{\theta} \,|\, \boldsymbol{y})$(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)
The probability density

$$f_X\left(y_i; x, \theta\right)=\left(2 \pi \sigma^2\right)^{-1 / 2} \exp \left(-\frac{1}{2} \frac{\left(y_i-\theta^T \underset{i}{x}\right)^2}{\sigma^2}\right) \tag{1}$$

The joint probability density of the sample xi is

$$f_{\Xi}(\xi ; \theta)=\prod_{i=1}^n f_X\left(x_i ; \mu, \sigma^2\right) \tag{2}$$

The likelihood function is

$$\begin{aligned} L(\theta ; \xi) & =f_{\Xi}(\xi ; \theta) \\ & =\prod_{i=1}^n f_X\left(x_i ; \mu, \sigma^2\right) \\ & =\prod_{i=1}^n\left(2 \pi \sigma^2\right)^{-1 / 2} \exp \left(-\frac{1}{2} \frac{\left(y_i-\theta^T \underset{i}{x}\right)^2}{\sigma^2}\right) \\ & =\left(2 \pi \sigma^2\right)^{-n / 2} \exp \left(-\frac{1}{2 \sigma^2} \sum_{i=1}^n\left(y_i-\theta^T \underset{i}{x}\right)^2\right)\end{aligned} \tag{3}$$

The log-likelihood function is

$$\begin{aligned} l(\theta ; \xi) & =\ln [L(\theta ; \xi)] \\ & =\ln \left[\left(2 \pi \sigma^2\right)^{-n / 2} \exp \left(-\frac{1}{2 \sigma^2} \sum_{i=1}^n\left(y_i-\theta^T \underset{i}{x}\right)^2\right)\right] \\ & =\ln \left[\left(2 \pi \sigma^2\right)^{-n / 2}\right]+\ln \left[\exp \left(-\frac{1}{2 \sigma^2} \sum_{i=1}^n\left(y_i-\theta^T \underset{i}{x}\right)^2\right)\right] \\ & =-\frac{n}{2} \ln \left(2 \pi \sigma^2\right)-\frac{1}{2 \sigma^2} \sum_{i=1}^n\left(y_i-\theta^T \underset{i}{x}\right)^2 \\ & =-\frac{n}{2} \ln (2 \pi)-\frac{n}{2} \ln \left(\sigma^2\right)-\frac{1}{2 \sigma^2} \sum_{i=1}^n\left(y_i-\theta^T \underset{i}{x}\right)^2\end{aligned} \tag{4}$$

(2)
To finding the Maximun Likelihood Estimate of θ, maxmizing L is equivalent to minimizing $\sum_{i=1}^m\left(y^{(i)}-\theta^T x^{(i)}\right)^2$. That is solving argmin $min_\theta \theta\|y-X\theta\|^2$

(3) With the definition:

$$\boldsymbol{x}^T\left(\boldsymbol{X}^T \boldsymbol{X}+\lambda \boldsymbol{I}\right) \boldsymbol{x}=\boldsymbol{x}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{x}+\lambda \boldsymbol{x}^T \boldsymbol{x}=\|\boldsymbol{X} \boldsymbol{x}\|_2+\lambda\|\boldsymbol{x}\|_2>0 \tag{5}$$

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

(4) The problem

$$\operatorname{argmin}_{\boldsymbol{\theta}}\|\boldsymbol{y}-\boldsymbol{X} \boldsymbol{\theta}\|^2+\lambda\|\boldsymbol{\theta}\|^2 \tag{6}$$

just add a small distribance to the classical $\text{argmin}_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2 = \arg\min_{\bar{\theta}^*}(y - X)^T(y - X) = E$ where $\frac{\partial E_{\bar{\theta}}}{\partial \theta} = 2X^T\left(X_{\bar{\theta}}\right)$ When Positive Definite, we have:

$$X^T X\bar{\theta} = X^T y$$
$$\bar{\theta}^* = \left(X^T X\right)^{-1} X^T y$$

Then we add a small distribance λ, get :

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

(5) With the Bayes' formula:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$
$$P(\theta|y) = \frac{P(y|\theta)P(\theta)}{P(y)}$$

Where $f_X(y_i;\theta)$ is the same with $P(y|\theta)$ and P(y) is independent with $P(\theta)$ so we get:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|y) = \arg\max_{\theta} p(\theta)p(y|\theta) \tag{7}$$

Use the same formula in (4):

$$\frac{p(\theta|X)}{\partial\theta} = \frac{p(\theta)p(X|\theta)}{\partial\theta} = 0 \tag{8}$$

So the the original function has been proved:

$$\bar{\theta}^* = \left(X^T X\right)^{-1} X^T y$$

Problem 2: Gradient Descent

(1) For the convx function f(x) we get:

$$\frac{f'(x) - f'(y)}{x - y} \geqslant 0$$

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

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

Let $x = a < y$, we have

$$'(y) - f'(a) \leqslant \beta(y - a)$$

Then

$$\int_a^b [f'(y) - f'(a)]\, dy \leqslant \int_a^b [\beta(y - a)]dy$$

That is

$$f(b) - f(a) - f'(a)(b - a) \leqslant \beta\left(\frac{1}{2}b^2 - \frac{1}{2}a^2 - ab + a^2\right)$$

Simplifying the inequality above, we have

$$f(b) - f(a) \leqslant f'(a)(b - a) + \frac{\beta}{2}(b - a)^2$$

(2) From (1), let y be x_{k+1} and x be x_k:

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

Give $x_{k+1} = x_k - \eta f'(x_k)$:

$$f\left(x_{k+1}\right) \leqslant f\left(x_k\right) - \eta\left(1 - \frac{\eta\beta}{2}\right)\left(f'\left(x_k\right)\right)^2 \tag{9}$$

(3)

$$\left\|x_{t+1} - x^*\right\|^2 = \left\|x_t - \eta\nabla f\left(x_t\right) - x^*\right\|^2 = \left\|x_t - x^*\right\|^2 - 2\eta\nabla f\left(x_t\right)^T\left(x_t - x^*\right) + \eta^2\left\|\nabla f\left(x_t\right)\right\|^2 \tag{10}$$

At x_t:

$$f\left(x_t\right) - f\left(x^*\right) \leq \nabla f\left(x_t\right)^T\left(x_t - x^*\right) - \frac{1}{2\beta}\left\|\nabla f\left(x_t\right) - \nabla f\left(x^*\right)\right\|^2 \tag{11}$$

define $\delta_t = f\left(x_t\right) - f\left(x^*\right)$: $\delta_{t+1} \leq \delta_t - \frac{1}{2\beta}\left\|\nabla_{x_t} f\right\|^2$ Convexity of f also implies

$$\begin{aligned} \delta_t &\leq \left(\nabla_{x_t} f\right)^\top\left(x_t - x^*\right) \\ &\leq \left\|\nabla_{x_t} f\right\| \cdot \left\|x_t - x^*\right\| \\ &\leq \left\|\nabla_{x_t} f\right\| \cdot D \\ \frac{\delta_t}{D} &\leq \left\|\nabla_{x_t} f\right\| \end{aligned}$$

We know D $\geq \left\|x_1 - x^*\right\|$:

$$\begin{aligned} \delta_{t+1} &\leq \delta_t - \frac{\delta_t^2}{2\beta D^2} \\ \frac{1}{\delta_t} &\leq \frac{1}{\delta_{t+1}} - \frac{\delta_t}{\delta_{t+1}} \cdot \frac{1}{2\beta D^2} \\ \frac{\delta_t}{\delta_{t+1}} \cdot \frac{1}{2\beta D^2} &\leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t} \\ \frac{1}{2\beta D^2} &\leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t} \end{aligned}$$

We may conclude that

$$\frac{1}{\delta_T} \geq \frac{1}{\delta_0} + \frac{T}{2\beta D^2} \geq \frac{T}{2\beta D^2}$$

from which it follows that $\delta_T \leq 2\beta D^2/T$, hence $T = 2\beta D^2\varepsilon^{-1}$ iterations suffice to ensure that $\delta_T \leq \varepsilon$ as claimed. So the original formula has been proved.

(4) Let

$$f(y) - f(x) = f(y) - f(z) + f(z) - f(x)$$

With convex:

$$f(z) - f(x) \geqslant f'(x)(z - x) = f'(x)(z - y) + f'(x)(y - x)$$

And:

$$f(y) - f(z) \geqslant f'(y)(y - z) - \frac{\beta}{2}(y - z)^2$$

And:

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

So we have:

$$f(y) - f(x) \geqslant f'(x)(y - x) + \frac{1}{\beta}\left[f'(y) - f'(x)\right]^2 - \frac{\beta}{2} \times \frac{1}{\beta}\left[f'(y) - f'(x)\right]^2 = f'(x)(y - x) + \frac{1}{2\beta}\left[f'(y) - f'(x)\right]^2$$

Problem 3: Kernel functions

(1) For every z:

$$\begin{aligned}
z^T K z &= \sum_{i=1}^{m} \sum_{j=1}^{m} z_i K_{ij} z_j \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} z_i f\left(x^{(i)}\right)^T f\left(x^{(j)}\right) z_j \\
&= \left(\sum_{i=1}^{m} z_i f\left(x^{(i)}\right)\right)^T \left(\sum_{j=1}^{m} z_j f\left(x^{(j)}\right)\right) \\
&= \left(\sum_{i=1}^{m} z_i f\left(x^{(i)}\right)\right)^2 \\
&\geqslant 0
\end{aligned}$$

So K is PSD.
(2) We will start by showing that "if matrices A and B are PSD, then $C_{ij} = A_{ij} \times B_{ij}$ is PSD"
Suppose we have PSD matrix Q, then we can prove $Q^{\frac{1}{2}}$ is PSD matrix(where cov() return co-variance matrix):

$$\operatorname{cov}\left(Q^{\frac{1}{2}} \mathbf{x}\right) = Q^{\frac{1}{2}} \operatorname{cov}(\mathbf{x})\Big) Q^{\frac{1}{2}} = Q^{\frac{1}{2}} \mathbf{I} Q^{\frac{1}{2}} = Q$$

We also know that any covariance matrix is PSD. So given A and B PSD, we know that they are covariance matrices.
We want to show that C is also a covariance matrix and therefore PSD.
Let $u = (u_1, \ldots, u_n)^T \sim N\left(0_p, A\right)$ and $v = (v_1, \ldots, v_n)^T \sim N\left(0_p, B\right)$ where $0 + p$ is a p-dimensional vector of zeros Define the vector $w = (u_1 v_1, \ldots, u_n v_n)^T$

$$\operatorname{cov}(w) = E\left[\left(w - \mu^w\right)\left(w - \mu^w\right)^T\right] = E\left[w w^T\right]$$

This is because $\mu_i^w = 0$ for all i. This is because u and v are independent so $\mu^w = \mu^u \times \mu^v = 0_p$

$$\begin{aligned}
\operatorname{cov}(w)_{i,j} &= E\left[w_i w_j^T\right] = E\left[\left(u_i v_i\right)\left(u_j v_j\right)\right] = E\left[\left(u_i u_j\right)\left(v_i v_j\right)\right] \\
&= E\left[u_i u_j\right] E\left[v_i v_j\right]
\end{aligned}$$

This is again because u and v are independent.

$$\operatorname{cov}(w)_{i,j} = E\left[u_i u_j\right] E\left[v_i v_j\right] = A_{i,j} \times B_{i,j} = C_{i,j}$$

Therefore C is a covariance matrix and therefore PSD.
Therefore any kernel matrix created from $k = k_1 k_2$ is PSD.
(3) First, we will show that $c_1 * k_1\left(x, x'\right) + c_2 * k_2\left(x, x'\right)$, where $c_1, c_2 \geq 0$ is a valid Kernel.
K is PSD because any $v \in \mathbb{R}^n$ $v^T\left(c_1 K_1 + c_2 K_2\right) v = c_1\left(v^T K_1 v\right) + c_2\left(v^T K_2 v\right) \geq 0$ as $v^T K_1 v \geq 0$ and $v^T K_2 v \geq 0$ follows from K_1 and K_2 being positive semi definite.
So k is a valid kernel.
So any Non-negative weighted sum of k will be PSD.
(4) We have:

$$\exp(x) = \lim_{i \to \infty}\left(1 + x + \cdots + \frac{x^i}{i!}\right)$$

with (4) we know any Non-negative weighted sum of k is PSD, so $k(\boldsymbol{u}, \boldsymbol{v}) = \exp(k_1(\boldsymbol{u}, \boldsymbol{v}))$ is PSD.