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

Assignment #2 (Linear Model)

Instructor: Beilun Wang *Name:* No Name, *ID:* 12345678

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)

(2)

(3)

(4)

(5)

Problem 2: Gradient Descent

(1)

(2)

(3)

(4)

Problem 3: Kernel functions

(1)

(2)

(3)

(4)