Problem 1 [35pts]

Probabilistic Interpretation of Linear Regression

Given data set $\boldsymbol{X} = (\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(n)})^{\mathrm{T}}$ and $\mathbf{y} = (y^{(1)}, y^{(2)}, \dots, y^{(n)})^{\mathrm{T}}$, where $(\boldsymbol{x}^{(i)\mathrm{T}}, y^{(i)}) = (x_1^{(i)}, x_2^{(i)}, \dots, x_p^{(i)}, y^{(i)})$ is the i-th example. We focus on the model

$$y^{(i)} = \boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + \varepsilon_i,$$

where ε is an error term of unmodeled effects or random noise. Assume that ε follows a Gaussian distribution $\varepsilon \sim N(0, \sigma^2)$, then we have:

$$p(y^{(i)} | \boldsymbol{x}^{(i)}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}\right)^2}{2\sigma^2}\right)$$

(1) [5pts] By the i.i.d. assumption, write down the log-likelihood function of $\mathbf{y}$. You can ignore any unnecessary constants.

(2) [5pts] Based on your answer to (1), show that finding Maximum Likelihood Estimate of $\boldsymbol{\theta}$ is equivalent to solving $\mathrm{argmin}_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2$.

(3) [5pts] Prove that $\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}$ with $\lambda > 0$ is Positive Definite (Hint: definition).

(4) [10pts] Show that $\boldsymbol{\theta}^* = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ is the solution to $\mathrm{argmin}_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2 + \lambda\|\boldsymbol{\theta}\|^2$.

(5) [10pts] Assuming $\theta_i \sim N(0, \tau^2)$ for $i = 1, 2, \dots, p$ in $\boldsymbol{\theta}$ ($\boldsymbol{\theta}$ does not vary in each sample), write down the estimate of $\boldsymbol{\theta}$ by maximizing the conditional distribution $f(\boldsymbol{\theta}|\mathbf{y})$ (Hint: Bayes' formula). You can ignore any unnecessary constants. Also find out the relationship between your estimate and the solution in (4).

(1) The log-likelihood function of $\boldsymbol{y}$ is:

$$\ln \ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln\left(p(y^{(i)} | \boldsymbol{x}^{(i)}; \boldsymbol{\theta})\right) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}\right)^2 - n \ln \sqrt{2\pi}\sigma$$

(2) After ignoring all of the unnecessary constants, the Maximum Likelihood Estimate of $\boldsymbol{\theta}$ is $\hat{\theta} = \arg\min \sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}\right)^2$, then

$$\sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}\right)^2 = \left[y^{(1)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(1)} \cdots\ y^{(n)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(n)}\right] \begin{bmatrix} y^{(1)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(1)} \\ \vdots \\ y^{(n)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(n)} \end{bmatrix}$$

where

$$\begin{bmatrix} y^{(1)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(1)} \\ \vdots \\ y^{(n)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(n)} \end{bmatrix} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{n} \end{bmatrix} - \boldsymbol{\theta}^T \left[\boldsymbol{x}^{(1)} \cdots \boldsymbol{x}^{(n)}\right] = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}$$

Therefore,

$$\sum_{i=1}^{n} \left(y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}\right)^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2$$

In conclusion, finding MLE of $\boldsymbol{\theta}$ is equivalent to solving $\arg\min_{\theta} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2$.

(3) Let $\boldsymbol{v}$ be a non-zero vector, we have

$$\boldsymbol{v}^T(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I})\boldsymbol{v}=\boldsymbol{v}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{v}+\boldsymbol{v}^T\lambda\boldsymbol{I}\boldsymbol{v}=\|\boldsymbol{X}\boldsymbol{v}\|^2+\lambda\|\boldsymbol{v}\|^2$$

Moreover, $\lambda>0$, hence $\|\boldsymbol{X}\boldsymbol{v}\|^2+\lambda\|\boldsymbol{v}\|^2>0$, which implies that $\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I}$ is positive definite.

(4) First simplify the objective function,

$$\begin{aligned}\boldsymbol{F}(\boldsymbol{\theta})&=\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|^2+\lambda\|\boldsymbol{\theta}\|^2\\&=(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta})+\lambda\boldsymbol{\theta}^T\boldsymbol{\theta}\\&=\boldsymbol{y}^T\boldsymbol{y}-\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\theta}-\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{y}+\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta}+\lambda\boldsymbol{\theta}^T\boldsymbol{\theta}\end{aligned}$$

Then, find the partial derivative with respect to $\boldsymbol{\theta}$ and set it equal to 0.

$$\frac{\partial\boldsymbol{F}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}=2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\theta}-2\boldsymbol{X}^T\boldsymbol{y}+2\lambda\boldsymbol{I}\boldsymbol{\theta}=0$$

Solve this equation, we can get $\boldsymbol{\theta}^*=(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$

Then find the second order partial derivative with respect to $\boldsymbol{\theta}$

$$\frac{\partial^2\boldsymbol{F}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}=2(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I})>0$$

Therefore, $\boldsymbol{\theta}^*$ is the solution of $\arg\min\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|^2+\lambda\|\boldsymbol{\theta}\|^2$

(5) According to Bayes' formula

$$f(\boldsymbol{\theta}|\boldsymbol{y})=\frac{f(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta})}{f(\boldsymbol{y})}$$

Write the distribution of θ_i in vector form

$$f(\boldsymbol{\theta})=\frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}\right),\text{where }\boldsymbol{\Sigma}=\begin{bmatrix}\tau^2&0&\cdots&0\\0&\tau^2&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&\tau^2\end{bmatrix}=\tau^2\boldsymbol{I}$$

Then

$$\begin{aligned}f(\boldsymbol{\theta}|\boldsymbol{y})&\propto f(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta})\\&\propto\exp\left(-\frac{\|\boldsymbol{\theta}\|^2}{2\tau^2}\right)\exp\left(-\frac{\left(y^{(i)}-\boldsymbol{\theta}^T\boldsymbol{x}^{(i)}\right)^2}{2\sigma^2}\right)\end{aligned}$$

Let $\ell_0(\boldsymbol{\theta})$ be the log-likelihood function of $f(\boldsymbol{\theta}|\boldsymbol{y})$, and logarithmization

$$\ln\ell_0(\boldsymbol{\theta})\propto\frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\theta}\|^2}{2\sigma^2}+\frac{\|\boldsymbol{\theta}\|^2}{2\tau^2}$$

Find the partial derivative with respect to $\boldsymbol{\theta}$ and set it equal to 0, solve the equation, we get

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

Compare with the solution in (4), we find that $\widehat{\boldsymbol{\theta}}$ is equal to $\boldsymbol{\theta}^*$ when $\lambda = \frac{\sigma^2}{\tau^2}$.

Problem 2 [40pts]

Multi-Class Logistic Regression

Given data set $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$, where $\mathbf{x}_i = (x_{i1}; x_{i2}; \ldots; x_{id})$, $y \in \{1, 2, \ldots, K\}$, please extend Logistic Regression to multiclass classification problem.

(1) **[20pts]** Write down the log-likelihood function of the multiclass Logistic Regression model;

(2) **[20pts]** Write down the gradient of log-likelihood function.

Hint 1: To arrive at the multinomial logit model, for K possible outcomes, we can run $K-1$ independent binary logistic regression models, in which one outcome is chosen as a "pivot" and then the other $K-1$ outcomes are separately regressed against the pivot outcome.

$$\begin{aligned}
\ln\frac{p(y=1|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_1^{\mathrm{T}}\mathbf{x} + b_1 \\
\ln\frac{p(y=2|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_2^{\mathrm{T}}\mathbf{x} + b_2 \\
&\cdots \\
\ln\frac{p(y=K-1|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_{K-1}^{\mathrm{T}}\mathbf{x} + b_{K-1}
\end{aligned}$$

Hint 2: Define the indicator function $\mathbb{I}(\cdot)$,

$$\mathbb{I}(y=j) = \begin{cases} 1 & \text{if } y = j \\ 0 & \text{if } y \neq j \end{cases}$$

(1) Let's start from logistic function $y = 1/(1+e^{-z})$, where $z = \boldsymbol{w}_k^T\boldsymbol{x} + b_k, k = 1, \ldots, K$. After logarithmization, we can get

$$\ln\frac{y}{1-y} = \boldsymbol{w}_k^T\boldsymbol{x} + b_k$$

According to the Hint 1, we can run $K-1$ independent binary logistic regression models, in which one output is chosen as a main category and other $K-1$ outputs are separately regressed against the main category.

Treat y in the formula as a class posterior probability estimate $P(y=j|\boldsymbol{x}), j = 1, \ldots, K-1$, and then rewrite the above formula to get

$$\ln\frac{P(y=j|\boldsymbol{x})}{P(y=K|\boldsymbol{x})} = \boldsymbol{w}_j^T\boldsymbol{x} + b_j \Rightarrow P(y=j|\boldsymbol{x}) = P(y=K|\boldsymbol{x})e^{\boldsymbol{w}_j^T\boldsymbol{x}+b_j}$$

For all categories, we have $\sum_{k=1}^{K} P(y=k|\boldsymbol{x}) = 1$, hence

$$P(y=K|\boldsymbol{x}) = 1 - \sum_{n=1}^{K-1} P(y=K|\boldsymbol{x}) e^{\boldsymbol{w}_n^T x + b_n} = \frac{1}{1+\sum_{n=1}^{K-1} e^{\boldsymbol{w}_n^T x + b_n}}$$

Next, only need to substitute $P(y=K|\boldsymbol{x})$ into $P(y=j|\boldsymbol{x})$, then we can calculate all the probabilities $P(y=k|\boldsymbol{x})$.

Finally, the log-likelihood function of multiclass logistic regression model is

$$\ln \ell(\boldsymbol{w}_k, b_k) = \sum_{i=1}^{m} \ln p(y_i|\boldsymbol{x}_i; \boldsymbol{w}_k, b_k)$$

Let $\boldsymbol{\beta}_k = (\boldsymbol{w}_k; b_k), \hat{\boldsymbol{x}} = (\boldsymbol{x}; 1)$, then $\boldsymbol{w}_k^T \boldsymbol{x} + b_k$ can be written as $\boldsymbol{\beta}_k^T \hat{\boldsymbol{x}}$ in short. And let $p_k(\hat{\boldsymbol{x}}; \boldsymbol{\beta}_k) = p(y=k|\hat{\boldsymbol{x}}; \boldsymbol{\beta}_{\boldsymbol{k}})$, then likelihood term in $\ln \ell(\boldsymbol{w}, b)$ can be rewritten as

$$\begin{aligned} p(y_i|\boldsymbol{x}_i; \boldsymbol{w}_k, b_k) &= \mathbb{I}(y_i = k) p_k(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}_k) \\ &= \mathbb{I}(y_i = j) p_j(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}_k) + \mathbb{I}(y_i = K) p_K(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}_k) \\ &= \mathbb{I}(y_i = j) \frac{e^{\boldsymbol{\beta}_j \hat{\boldsymbol{x}}_i}}{1+\sum_{n=1}^{K-1} e^{\boldsymbol{\beta}_n \hat{\boldsymbol{x}}_i}} + \mathbb{I}(y_i = K) \frac{1}{1+\sum_{n=1}^{K-1} e^{\boldsymbol{\beta}_n \hat{\boldsymbol{x}}_i}} \end{aligned}$$

Finally, we can obtain

$$\ln \ell(\boldsymbol{\beta}_k) = \sum_{i=1}^{m} \sum_{j=1}^{K-1} \mathbb{I}(y_i = j) \boldsymbol{\beta}_j^T \hat{\boldsymbol{x}}_i - \ln\left(1 + \sum_{n=1}^{K-1} e^{\boldsymbol{\beta}_n^T \hat{\boldsymbol{x}}_i}\right)$$

(2) The gradient of log-likelihood function is

$$\frac{\partial \ln \ell(\boldsymbol{\beta}_k)}{\partial \boldsymbol{\beta}_k} = \begin{cases} \sum_{i=1}^{m} \mathbb{I}(y_i = k) \hat{\boldsymbol{x}}_i - \dfrac{\hat{\boldsymbol{x}}_i e^{\boldsymbol{\beta}_k^T \hat{\boldsymbol{x}}_i}}{1+\sum_{n=1}^{K-1} e^{\boldsymbol{\beta}_n^T \hat{\boldsymbol{x}}_i}}, k = 1, \dots, K-1 \\ 0, k = K \end{cases}$$

Problem 3 [45pts]: 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) [10pts] Prove that

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

(Hint: Newton-Leibniz formula.)

(2) [5pts] 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) [20pts] Based on (2), let $\eta = 1/\beta$ and assume the unique global minimum point x^* of f exists. Prove that

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

(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) [10pts] 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 $z = y - \frac{1}{\beta}(f'(y) - f'(x))$.)

(1) From the Newton-Leibniz formula, it is apparent that

$$f(y) - f(x) = \int_0^1 f'\big(x + t(y - x)\big)(y - x)dt$$

Then by the property of convex function and β-smoothness

$$\begin{aligned}
f(y) - f(x) - f'(x)(y - x) &= \int_0^1 \Big(f'\big(x + t(y - x)\big) - f'(x)\Big)(y - x)dt \\
&\leq \left|\int_0^1 \Big(f'\big(x + t(y - x)\big) - f'(x)\Big)(y - x)dt\right| \\
&\leq \int_0^1 \big|f'\big(x + t(y - x)\big) - f'(x)\big||y - x|dt \\
&\leq \int_0^1 \beta|t(y - x)||y - x|dt \\
&= \frac{\beta}{2}(y - x)^2
\end{aligned}$$

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

(2) Based on (1)

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

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

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

$$\Rightarrow\ f(x_{k+1}) \le f(x_k) - \eta(1 - \frac{\eta\beta}{2})\big(f'(x_k)\big)^2$$

(3) Based on (2), let $\eta = 1/\beta$, we have

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

$$\sum_{k=1}^{K}\big(f'(x_k)\big)^2 \le 2\beta\big(f(x_k) - f(x_{k+1}) + \cdots + f(x_1) - f(x_2)\big)$$

$$= 2\beta(f(x_1) - f(x_{K+1}))$$

$\sum_{k=1}^{K}\big(f'(x_k)\big)^2$ is converge, hence $\lim\limits_{k\to\infty}\big(f'(x_k)\big)^2 = \lim\limits_{k\to\infty} f'(x_k) = 0$.

Next, we can prove $\lim\limits_{k\to\infty} x_k = x^*$ by contradiction. Assume that $\lim\limits_{k\to\infty} x_k \ne x^*$, which means

$$\exists\delta > 0, \forall N \in N_+, \exists m > N, |x_m - x^*| \ge \delta$$

From $\lim\limits_{k\to\infty} f'(x_k) = 0$, we know

$$\forall\epsilon > 0, \exists N' \in N_+, \forall n > N', |f'(x_n)| < \epsilon$$

Note that x^* is the unique global minimum point and $f'(x)$ is monotony, hence take $x_1 \in (x^* - \delta, x^*)$ and $x_2 \in (x^*, x^* + \delta)$, we get $f'(x_1) < f'(x^*) = 0 < f'(x_2)$. Let $\epsilon_0 = \frac{1}{2}\min\{|f'(x_1)|, |f'(x_2)|\}$, $\exists N = N_0, \forall m > N_0, |f'(x_m)| < \epsilon_0$. However, for the same $N = N_0$, there should be $\exists m > N, |x_m - x^*| \ge \delta$, which means $|f'(x_m)| \ge \min\{|f'(x_1)|, |f'(x_2)|\} = 2\epsilon_0$. This is contradictory. In conclusion, $\lim\limits_{k\to\infty} x_k = x^*$.

(4) Let $z = y - \frac{1}{\beta}\big(f'(y) - f'(x)\big)$, based on (1), we have

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

Therefore,

$$f(y) - f(z) \ge f'(y)(y - z) - \frac{\beta}{2}(y - z)^2 = \frac{1}{\beta}f'(y)\big(f'(y) - f'(x)\big) - \frac{\beta}{2}(y - z)^2$$

From the property of convex function

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

Add the above two formulas together

$$\begin{aligned} f(y) - f(x) &\geq f'(x)(y - x) + \frac{1}{\beta}\big(f'(y) - f'(x)\big)^2 - \frac{\beta}{2}(y - z)^2 \\ &= f'(x)(y - x) + \frac{1}{\beta}\big(f'(y) - f'(x)\big)^2 - \frac{1}{2\beta}\big(f'(y) - f'(x)\big)^2 \\ &= f'(x)(y - x) + \frac{1}{2\beta}\big(f'(y) - f'(x)\big)^2 \end{aligned}$$