HOMEWORK 1

1 证明《统计学习方法》习题 1.2:

通过经验风险最小化推导极大似然估计。证明模型是条件概率分布，当损失函数是对数损失函数时，经验风险最小化等价于极大似然估计。

Solution: Let the conditional probability distribution be $P(Y|X)$, and the loss function be the log-loss function $L(Y, P(Y|X)) = -\log P(Y|X)$. The empirical risk is:

$$R_{emp}(P) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, P(y_i|x_i)) = -\frac{1}{N}\sum_{i=1}^{N} \log P(y_i|x_i) \tag{1}$$

Minimizing the empirical risk is equivalent to maximizing:

$$\frac{1}{N}\sum_{i=1}^{N} \log P(y_i|x_i) \tag{2}$$

This is exactly the average log-likelihood. Therefore, minimizing the empirical risk with log-loss function is equivalent to maximum likelihood estimation.

2 请证明下述 Hoeffding 引理:

Lemma 1. *Let X be a random variable with $E(X) = 0$ and $P(X \in [a, b]) = 1$. Then it holds*

$$E\{\exp(sX)\} \leq \exp\{s^2(b-a)^2/8\}. \tag{3}$$

Solution: We begin by noting that the function $f(x) = \exp(sx)$ is convex for any real s. Therefore, for any $x \in [a, b]$, we can bound $f(x)$ by its linear interpolation between the endpoints:

$$\exp(sx) \leq \frac{b-x}{b-a}\exp(sa) + \frac{x-a}{b-a}\exp(sb) \tag{4}$$

Taking the expectation of both sides and using the linearity of expectation:

$$\begin{aligned}\mathbb{E}[\exp(sX)] &\leq \mathbb{E}\left[\frac{b-X}{b-a}\exp(sa)+\frac{X-a}{b-a}\exp(sb)\right] \\ &= \frac{b-\mathbb{E}[X]}{b-a}\exp(sa)+\frac{\mathbb{E}[X]-a}{b-a}\exp(sb)\end{aligned} \tag{5}$$

Since $\mathbb{E}[X]=0$, this simplifies to:

$$\mathbb{E}[\exp(sX)] \leq \frac{b}{b-a}\exp(sa)-\frac{a}{b-a}\exp(sb) \tag{6}$$

Now, let $g(s)=\log(\mathbb{E}[\exp(sX)])$. From the above inequality, we have:

$$g(s) \leq \log\left(\frac{b}{b-a}\exp(sa)-\frac{a}{b-a}\exp(sb)\right) \tag{7}$$

We can observe that $g(0)=0$ and

$$\begin{aligned}g'(s) &= \frac{d}{ds}\log(\mathbb{E}[\exp(sX)]) = \frac{1}{\mathbb{E}[\exp(sX)]}\cdot\frac{d}{ds}\mathbb{E}[\exp(sX)] \\ &= \frac{1}{\mathbb{E}[\exp(sX)]}\cdot\mathbb{E}\Big[\frac{d}{ds}\exp(sX)\Big] = \mathbb{E}[X\exp(sX)].\end{aligned}$$

It implies $g'(0)=\mathbb{E}[X]=0$. To bound $g''(s)$, we use the fact that for any random variable Y:

$$\begin{aligned}\frac{d^2}{ds^2}\log(\mathbb{E}[\exp(sY)]) &= \frac{\mathbb{E}[Y^2\exp(sY)]\mathbb{E}[\exp(sY)]-(\mathbb{E}[Y\exp(sY)])^2}{(\mathbb{E}[\exp(sY)])^2} \\ &\leq \frac{\mathbb{E}[Y^2\exp(sY)]\mathbb{E}[\exp(sY)]}{(\mathbb{E}[\exp(sY)])^2} = \frac{\mathbb{E}[Y^2\exp(sY)]}{\mathbb{E}[\exp(sY)]} \\ &\leq \frac{\mathbb{E}[Y^2]\cdot\mathbb{E}[\exp(sY)]}{\mathbb{E}[\exp(sY)]} = \mathbb{E}[Y^2].\end{aligned} \tag{8}$$

Therefore:

$$g''(s) \leq \mathbb{E}[X^2] \leq \frac{(b-a)^2}{4} \tag{9}$$

The last inequality follows from the fact that $X\in[a,b]$, we have $a\leq X\leq b$. Subtracting $\frac{a+b}{2}$ from all parts, we get $-\frac{b-a}{2}\leq X-\frac{a+b}{2}\leq\frac{b-a}{2}$. Squaring this inequality

and taking expectations, we obtain $\mathbb{E}[(X - \frac{a+b}{2})^2] \leq (\frac{b-a}{2})^2$. Expanding the left side and using the fact that $\mathbb{E}[X] = 0$, we get $\mathbb{E}[X^2] + (\frac{a+b}{2})^2 \leq (\frac{b-a}{2})^2$. Rearranging terms leads to $\mathbb{E}[X^2] \leq (\frac{b-a}{2})^2 - (\frac{a+b}{2})^2 \leq \frac{(b-a)^2}{4}$.

Now, by Taylor's theorem with the Lagrange form of the remainder, there exists some $\xi \in [0, s]$ such that:

$$g(s) = g(0) + g'(0)s + \frac{1}{2}g''(\xi)s^2 \leq 0 + 0 + \frac{1}{2} \cdot \frac{(b-a)^2}{4}s^2 = \frac{s^2(b-a)^2}{8} \tag{10}$$

Taking the exponential of both sides completes the proof:

$$\mathbb{E}[\exp(sX)] = \exp(g(s)) \leq \exp\{s^2(b-a)^2/8\} \tag{11}$$

Thus, we have proven Hoeffding's Lemma.

3 已知针对模型 f，使用训练数据集得到的估计记为 $\widehat{f}$，现有独立于训练数据集的 (x_0, y_0)，其中 x_0 为非随机的给定值，$y_0 = f(x_0) + \epsilon$，其中 ϵ 为随机误差项，证明：

$$\mathrm{E}\left(y_0 - \widehat{f}\,(x_0)\right)^2 = \mathrm{Var}\left(\widehat{f}\,(x_0)\right) + \left[\mathrm{Bias}\left(\widehat{f}\,(x_0)\right)\right]^2 + \mathrm{Var}(\epsilon).$$

Solution: Expand the left side:

$$\begin{aligned}
\mathrm{E}\left(y_0 - \widehat{f}\,(x_0)\right)^2 &= \mathrm{E}\left[\left\{f(x_0) + \epsilon - \hat{f}(x_0)\right\}^2\right] \\
&= \mathrm{E}\left[\left\{f(x_0) - \hat{f}(x_0) + \epsilon\right\}^2\right] \\
&= \mathrm{E}\left[\left\{f(x_0) - \hat{f}(x_0)\right\}^2 + 2\epsilon\left\{f(x_0) - \hat{f}(x_0)\right\} + \epsilon^2\right] \\
&= \mathrm{E}\left[\left\{f(x_0) - \hat{f}(x_0)\right\}^2\right] + 2\mathrm{E}\left[\epsilon\left\{f(x_0) - \hat{f}(x_0)\right\}\right] + \mathrm{E}\left(\epsilon^2\right).
\end{aligned}$$

Note that $\mathrm{E}(\epsilon) = 0$ and ϵ is independent of $\hat{f}(x_0)$, so the middle term becomes zero

and $\mathrm{E}(\epsilon^2) = \mathrm{Var}(\epsilon)$. For the first term, add and subtract $\mathrm{E}\big(\hat{f}(x_0)\big)$:

$$\begin{aligned}
\mathrm{E}\Big[\big\{f(x_0) - \hat{f}(x_0)\big\}^2\Big] &= \mathrm{E}\Big[\big\{f(x_0) - \mathrm{E}\big(\hat{f}(x_0)\big) + \mathrm{E}\big(\hat{f}(x_0)\big) - \hat{f}(x_0)\big\}^2\Big] \\
&= \mathrm{E}\Big[\big\{f(x_0) - \mathrm{E}\big(\hat{f}(x_0)\big)\big\}^2\Big] + \mathrm{E}\Big[\big\{\mathrm{E}\big(\hat{f}(x_0)\big) - \hat{f}(x_0)\big\}^2\Big] + \mathcal{C} \\
&= \Big[\mathrm{Bias}\left(\widehat{f}\,(x_0)\right)\Big]^2 + \mathrm{Var}\left(\widehat{f}\,(x_0)\right) + \mathcal{C},
\end{aligned}$$

where $\mathcal{C}$ is the cross term:

$$\begin{aligned}
\mathcal{C} &= 2 \times \mathrm{E}\Big[\big\{f(x_0) - \mathrm{E}\big(\hat{f}(x_0)\big)\big\}\big\{\mathrm{E}\big(\hat{f}(x_0)\big) - \hat{f}(x_0)\big\}\Big] \\
&= 2 \times \big\{f(x_0) - \mathrm{E}\big(\hat{f}(x_0)\big)\big\} \times \mathrm{E}\Big\{\mathrm{E}\big(\hat{f}(x_0)\big) - \hat{f}(x_0)\Big\} \\
&= 2 \times \big\{f(x_0) - \mathrm{E}\big(\hat{f}(x_0)\big)\big\} \times 0 = 0.
\end{aligned}$$

Finally, we obtain

$$\mathrm{E}\left(y_0 - \widehat{f}\,(x_0)\right)^2 = \mathrm{Var}\left(\widehat{f}\,(x_0)\right) + \Big[\mathrm{Bias}\left(\widehat{f}\,(x_0)\right)\Big]^2 + \mathrm{Var}(\epsilon).$$

4 Please read the background and then prove the following results.

Background:

Let $\mathbf{y} = \Psi(\mathbf{x})$, where $\mathbf{y}$ is an $m \times 1$ vector, and $\mathbf{x}$ is an $n \times 1$ vector. Denote

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}^\top} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \tag{12}$$

Prove the results:

(a) Let $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{y}$ is m×1, $\mathbf{x}$ is n×1, $\mathbf{A}$ is m×n, and $\mathbf{A}$ does not depend on $\mathbf{x}$,

then

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}^\top} = \mathbf{A} \tag{13}$$

Solution: Let's expand the matrix multiplication $\mathbf{y} = \mathbf{A}\mathbf{x}$:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

This means that for each i from 1 to m:

$$y_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = \sum_{j=1}^{n} a_{ij}x_j$$

Now, let's compute $\frac{\partial y_i}{\partial x_j}$ for any i and j:

$$\frac{\partial y_i}{\partial x_j} = \frac{\partial}{\partial x_j} \sum_{k=1}^{n} a_{ik}x_k = a_{ij}$$

This is because $\frac{\partial x_k}{\partial x_j} = 1$ if $k = j$, and 0 otherwise, and a_{ij} does not depend on x_j. Therefore, we have:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}^\top} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = \mathbf{A}$$

This completes the proof of part (a).

(b) Let the scalar α be defined by $\alpha = \mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, where $\mathbf{y}$ is m×1, $\mathbf{x}$ is n×1, $\mathbf{A}$ is m×n, and $\mathbf{A}$ is independent of $\mathbf{x}$ and $\mathbf{y}$, then

$$\frac{\partial \alpha}{\partial \mathbf{x}^{\top}} = \mathbf{x}^{\top}\mathbf{A}^{\top}\left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}^{\top}}\right) + \mathbf{y}^{\top}\mathbf{A} \tag{14}$$

Solution: Let's expand $\alpha = \mathbf{y}^{\top}\mathbf{A}\mathbf{x}$:

$$\alpha = [y_1, y_2, ..., y_m] \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{m}\sum_{j=1}^{n} y_i a_{ij} x_j$$

Now, let's apply the product rule to differentiate with respect to x_k:

$$\begin{aligned} \frac{\partial \alpha}{\partial x_k} &= \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\frac{\partial y_i}{\partial x_k} a_{ij} x_j + y_i a_{ij}\frac{\partial x_j}{\partial x_k}\right) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\partial y_i}{\partial x_k} a_{ij} x_j + \sum_{i=1}^{m} y_i a_{ik} \end{aligned}$$

Now, we can write this in matrix form:

$$\begin{aligned} \frac{\partial \alpha}{\partial \mathbf{x}^{\top}} &= \left[\frac{\partial \alpha}{\partial x_1}, \frac{\partial \alpha}{\partial x_2}, ..., \frac{\partial \alpha}{\partial x_n}\right] \\ &= \mathbf{x}^{\top}\mathbf{A}^{\top}\left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}^{\top}}\right) + \mathbf{y}^{\top}\mathbf{A} \end{aligned}$$

This completes the proof of part (b).

(c) For the special case in which the scalar α is given by the quadratic form $\alpha = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$ where $\mathbf{x}$ is n×1, $\mathbf{A}$ is n×n, and $\mathbf{A}$ does not depend on $\mathbf{x}$, then

$$\frac{\partial \alpha}{\partial \mathbf{x}^{\top}} = \mathbf{x}^{\top}(\mathbf{A} + \mathbf{A}^{\top}) \tag{15}$$

Solution: Let's expand $\alpha = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$:

$$\alpha = [x_1, x_2, ..., x_n] \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i a_{ij} x_j$$

Now, let's differentiate with respect to x_k:

$$\begin{aligned} \frac{\partial \alpha}{\partial x_k} &= \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\partial x_i}{\partial x_k} a_{ij} x_j + x_i a_{ij}\frac{\partial x_j}{\partial x_k}\right) \\ &= \sum_{j=1}^{n} a_{kj} x_j + \sum_{i=1}^{n} x_i a_{ik} \\ &= \sum_{j=1}^{n} (a_{kj} + a_{jk}) x_j \end{aligned}$$

Now, we can write this in matrix form:

$$\frac{\partial \alpha}{\partial \mathbf{x}^{\top}} = [\frac{\partial \alpha}{\partial x_1}, \frac{\partial \alpha}{\partial x_2}, ..., \frac{\partial \alpha}{\partial x_n}] = \mathbf{x}^{\top}(\mathbf{A} + \mathbf{A}^{\top})$$

This completes the proof of part (c).

(d) Let the scalar α be defined by $\alpha = \mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, where $\mathbf{y}$ is $m \times 1$, $\mathbf{x}$ is $n \times 1$, $\mathbf{A}$ is $m \times n$, and both $\mathbf{y}$ and $\mathbf{x}$ are functions of the vector $\mathbf{z}$, where $\mathbf{z}$ is a $q \times 1$ vector and $\mathbf{A}$ does not depend on $\mathbf{z}$. Then

$$\frac{\partial \alpha}{\partial \mathbf{z}^{\top}} = \mathbf{x}^{\top}\mathbf{A}^{\top}\left(\frac{\partial \mathbf{y}}{\partial \mathbf{z}^{\top}}\right) + \mathbf{y}^{\top}\mathbf{A}\left(\frac{\partial \mathbf{x}}{\partial \mathbf{z}^{\top}}\right) \tag{16}$$

Solution: Let's expand $\alpha = \mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{x}$:

$$\alpha = [y_1, y_2, ..., y_m] \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{m} \sum_{j=1}^{n} y_i a_{ij} x_j$$

Now, let's differentiate with respect to z_k, applying the chain rule:

$$\begin{aligned} \frac{\partial \alpha}{\partial z_k} &= \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{\partial y_i}{\partial z_k} a_{ij} x_j + y_i a_{ij} \frac{\partial x_j}{\partial z_k} \right) \\ &= \sum_{j=1}^{n} \sum_{i=1}^{m} \frac{\partial y_i}{\partial z_k} a_{ij} x_j + \sum_{i=1}^{m} \sum_{j=1}^{n} y_i a_{ij} \frac{\partial x_j}{\partial z_k} \end{aligned}$$

Now, we can write this in matrix form:

$$\frac{\partial \alpha}{\partial \mathbf{z}^\top} = \left[\frac{\partial \alpha}{\partial z_1}, \frac{\partial \alpha}{\partial z_2}, ..., \frac{\partial \alpha}{\partial z_q} \right] = \mathbf{x}^\top \mathbf{A}^\top \left(\frac{\partial \mathbf{y}}{\partial \mathbf{z}^\top} \right) + \mathbf{y}^\top \mathbf{A} \left(\frac{\partial \mathbf{x}}{\partial \mathbf{z}^\top} \right)$$

This completes the proof of part (d).

(e) Let $\mathbf{A}$ be a nonsingular, $m \times m$ matrix whose elements are functions of the scalar parameter α. Then

$$\frac{\partial \mathbf{A}^{-1}}{\partial \alpha} = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial \alpha} \mathbf{A}^{-1} \tag{17}$$

提示：可利用乘法求导法则进行证明：

$$\frac{\partial AB}{\partial x} = \frac{\partial A}{\partial x} B + A \frac{\partial B}{\partial x}$$

其中 A, B 为矩阵，x 为标量。

Solution: We start with the identity $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$, where $\mathbf{I}$ is the $m \times m$ identity matrix. Differentiating both sides with respect to α:

$$\frac{\partial}{\partial \alpha} (\mathbf{A}\mathbf{A}^{-1}) = \frac{\partial}{\partial \alpha} (\mathbf{I}).$$

Applying the product rule:

$$\frac{\partial \mathbf{A}}{\partial \alpha}\mathbf{A}^{-1} + \mathbf{A}\frac{\partial \mathbf{A}^{-1}}{\partial \alpha} = \mathbf{0},$$

where $\mathbf{0}$ is the $m \times m$ zero matrix. Multiplying both sides by $\mathbf{A}^{-1}$ from the left:

$$\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial \alpha}\mathbf{A}^{-1} + \frac{\partial \mathbf{A}^{-1}}{\partial \alpha} = \mathbf{0}$$

Rearranging:

$$\frac{\partial \mathbf{A}^{-1}}{\partial \alpha} = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial \alpha}\mathbf{A}^{-1}$$

This completes the proof of part (e).

5 Please write $\hat{\mathbf{a}}$ as the solution of the minimization problem:

$$\min_{\mathbf{a}} \|\mathbf{X}\mathbf{a} - \mathbf{y}\|_2 \tag{18}$$

where $\mathbf{X}$ is a $n \times p$ matrix, $\mathbf{y}$ is a $n \times 1$ vector and $\mathbf{a}$ is a $p \times 1$ vector. $\mathbf{X}^\top\mathbf{X}$ is nonsingular.

Solution: To solve this problem, we first observe that minimizing the ℓ_2 norm is equivalent to minimizing its square. Thus, we can reformulate our problem as:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{X}\mathbf{a} - \mathbf{y}\|_2^2$$

Expanding the squared ℓ_2 norm, we obtain:

$$\|\mathbf{X}\mathbf{a} - \mathbf{y}\|_2^2 = (\mathbf{X}\mathbf{a} - \mathbf{y})^\top(\mathbf{X}\mathbf{a} - \mathbf{y}) = \mathbf{a}^\top\mathbf{X}^\top\mathbf{X}\mathbf{a} - 2\mathbf{a}^\top\mathbf{X}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y}$$

To find the minimum, we differentiate this expression with respect to $\mathbf{a}$ and set the result to zero:

$$\frac{\partial}{\partial \mathbf{a}}(\mathbf{a}^\top\mathbf{X}^\top\mathbf{X}\mathbf{a} - 2\mathbf{a}^\top\mathbf{X}^\top\mathbf{y} + \mathbf{y}^\top\mathbf{y}) = 2\mathbf{X}^\top\mathbf{X}\mathbf{a} - 2\mathbf{X}^\top\mathbf{y} = \mathbf{0}$$

Simplifying this equation, we obtain:

$$\mathbf{X}^\top \mathbf{X}\mathbf{a} = \mathbf{X}^\top \mathbf{y}$$

Given that $\mathbf{X}^\top \mathbf{X}$ is nonsingular, we can multiply both sides by its inverse to solve for $\mathbf{a}$:

$$\hat{\mathbf{a}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

This solution, known as the Ordinary Least Squares (OLS) estimator, provides the vector $\hat{\mathbf{a}}$ that minimizes $\|\mathbf{X}\mathbf{a} - \mathbf{y}\|_2$. To confirm that this critical point is indeed a minimum, we can verify that the Hessian matrix of the objective function, $2\mathbf{X}^\top \mathbf{X}$, is positive definite, which is guaranteed by the assumption that $\mathbf{X}^\top \mathbf{X}$ is nonsingular. In conclusion, the solution to the minimization problem $\min_{\mathbf{a}} \|\mathbf{X}\mathbf{a} - \mathbf{y}\|_2$ is given by $\hat{\mathbf{a}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$, under the condition that $\mathbf{X}^\top \mathbf{X}$ is nonsingular.