Comprehensive Guide to Linear Regression Concepts with Jacobian and Hessian Matrices

Original Problem and Solution

Problem Statement

For a matrix A which has values as (1,1),(1,2),(1,3) and features as age and experience the target column value is salary which has values as \$2000, \$4000 and \$6000. If we take these features in matrix X, how to calculate $(X^T X)^{-1} X^T \mathbf{y}$?

Solution

Given:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 2000 \\ 4000 \\ 6000 \end{bmatrix}$$

Compute $\mathbf{a} = (X^T X)^{-1} X^T \mathbf{y}$:

1. $X^T X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}$

2. $\det(X^T X) = 3 \cdot 14 - 6 \cdot 6 = 42 - 36 = 6$

$$(X^T X)^{-1} = \frac{1}{6} \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix} = \begin{bmatrix} \frac{7}{3} & -1 \\ -1 & \frac{1}{2} \end{bmatrix}$$

3. $X^T \mathbf{y} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 2000 \\ 4000 \\ 6000 \end{bmatrix} = \begin{bmatrix} 12000 \\ 28000 \end{bmatrix}$

4. $\mathbf{a} = \begin{bmatrix} \frac{7}{3} & -1 \\ -1 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 12000 \\ 28000 \end{bmatrix} = \begin{bmatrix} 28000 - 28000 \\ -12000 + 14000 \end{bmatrix} = \begin{bmatrix} 0 \\ 2000 \end{bmatrix}$

Thus, the linear regression model is:

$$\boxed{\text{Salary} = 0 + 2000 \cdot \text{Experience}}$$

Derivation of the Gradient (Jacobian)

Step-by-Step Derivation of $\nabla J(\mathbf{a}) = \frac{1}{m} X^T (X\mathbf{a} - \mathbf{y})$

The cost function for linear regression is:

$$J(\mathbf{a}) = \frac{1}{2m} \sum_{i=1}^{m} (h_{\mathbf{a}}(\mathbf{x}^{(i)}) - y^{(i)})^2 = \frac{1}{2m} \|\mathbf{Xa} - \mathbf{y}\|^2$$

Let's expand this:

$$J(\mathbf{a}) = \frac{1}{2m} (\mathbf{Xa} - \mathbf{y})^T (\mathbf{Xa} - \mathbf{y})$$

Step 1: Expand the quadratic form

$$J(\mathbf{a}) = \frac{1}{2m}\left[(\mathbf{Xa})^T\mathbf{Xa} - (\mathbf{Xa})^T\mathbf{y} - \mathbf{y}^T\mathbf{Xa} + \mathbf{y}^T\mathbf{y}\right]$$

Since $(\mathbf{Xa})^T\mathbf{y} = \mathbf{y}^T\mathbf{Xa}$ (both are scalars):

$$J(\mathbf{a}) = \frac{1}{2m}\left[\mathbf{a}^T\mathbf{X}^T\mathbf{Xa} - 2\mathbf{y}^T\mathbf{Xa} + \mathbf{y}^T\mathbf{y}\right]$$

Step 2: Compute the gradient

Using matrix calculus rules:

$$\frac{\partial}{\partial \mathbf{a}}(\mathbf{a}^T\mathbf{X}^T\mathbf{Xa}) = 2\mathbf{X}^T\mathbf{Xa}$$
$$\frac{\partial}{\partial \mathbf{a}}(2\mathbf{y}^T\mathbf{Xa}) = 2\mathbf{X}^T\mathbf{y}$$
$$\frac{\partial}{\partial \mathbf{a}}(\mathbf{y}^T\mathbf{y}) = \mathbf{0}$$

Therefore:

$$\nabla J(\mathbf{a}) = \frac{1}{2m}\left[2\mathbf{X}^T\mathbf{Xa} - 2\mathbf{X}^T\mathbf{y}\right] = \frac{1}{m}\left[\mathbf{X}^T\mathbf{Xa} - \mathbf{X}^T\mathbf{y}\right]$$

Which can be written as:

$$\nabla J(\mathbf{a}) = \frac{1}{m}X^T(X\mathbf{a} - \mathbf{y})$$

The X^T appears because:

- The derivative of $\mathbf{Xa}$ with respect to $\mathbf{a}$ is $\mathbf{X}^T$
- This follows from the matrix calculus rule: $\frac{\partial}{\partial \mathbf{x}}(\mathbf{Ax}) = \mathbf{A}^T$
- The term $X^T(X\mathbf{a} - \mathbf{y})$ represents the projection of the residuals onto the column space of X

Step 3: Verify with our solution

For our solution $\mathbf{a} = \begin{bmatrix} 0 \\ 2000 \end{bmatrix}$:

$$\nabla J(\mathbf{a}) = \frac{1}{3}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}\left(\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 0 \\ 2000 \end{bmatrix} - \begin{bmatrix} 2000 \\ 4000 \\ 6000 \end{bmatrix}\right) = \frac{1}{3}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

The gradient is zero at the optimal solution, as expected.

Jacobian Matrix: General Definition and Example

General Definition of Jacobian Matrix

The Jacobian matrix generalizes the gradient to vector-valued functions. For a function $\mathbf{f} : R^n \to R^m$, the Jacobian matrix $J_{\mathbf{f}}$ is:

$$J_{\mathbf{f}}(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

In linear regression, we have a scalar-valued function $J : R^n \to R$, so the Jacobian is simply the gradient (a row vector).

Example: Jacobian of a Vector Function

Consider $\mathbf{f}(x, y) = \begin{bmatrix} x^2 + y^2 \\ e^{xy} \end{bmatrix}$. The Jacobian is:

$$J_{\mathbf{f}}(x, y) = \begin{bmatrix} \frac{\partial}{\partial x}(x^2 + y^2) & \frac{\partial}{\partial y}(x^2 + y^2) \\ \frac{\partial}{\partial x}(e^{xy}) & \frac{\partial}{\partial y}(e^{xy}) \end{bmatrix} = \begin{bmatrix} 2x & 2y \\ ye^{xy} & xe^{xy} \end{bmatrix}$$

Hessian Matrix: General Definition and Example

General Definition of Hessian Matrix

The Hessian matrix contains all second-order partial derivatives of a scalar-valued function. For $f : R^n \to R$:

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Hessian in Linear Regression

For linear regression:

$$\nabla^2 J(\mathbf{a}) = \frac{1}{m} X^T X$$

For our problem:

$$\nabla^2 J(\mathbf{a}) = \frac{1}{3} \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & \frac{14}{3} \end{bmatrix}$$

The Hessian is positive definite (eigenvalues are positive), confirming that our solution is a minimum.

Example: Hessian of a Quadratic Function

Consider $f(x, y) = x^2 + 2xy + 3y^2$. The Hessian is:

$$\nabla^2 f(x, y) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 6 \end{bmatrix}$$

Geometric Interpretation of Coefficients in X-Y Plane

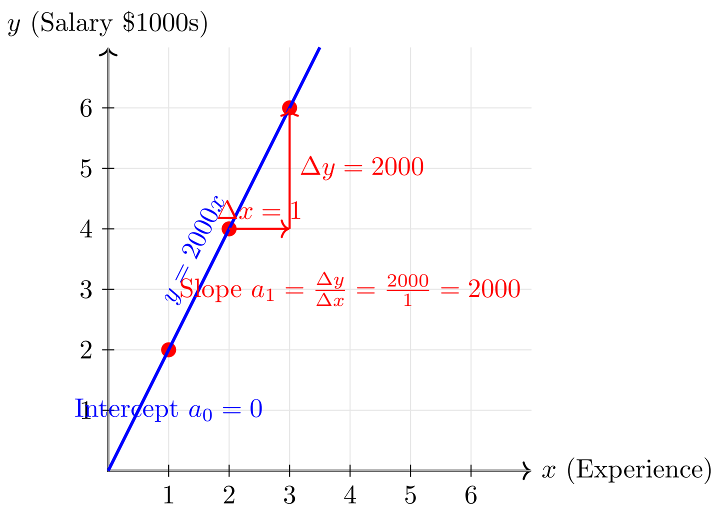
Representation of Coefficients and Slope

In the X-Y plane, the linear regression equation $y = a_0 + a_1 x$ represents a straight line where:

- a_0 is the **y-intercept** (where the line crosses the y-axis)
- a_1 is the **slope** (steepness of the line)

For our solution $\mathbf{a} = \begin{bmatrix} 0 \\ 2000 \end{bmatrix}$:

- **Intercept** $a_0 = 0$: The line passes through the origin (0,0)
- **Slope** $a_1 = 2000$: For each unit increase in Experience, Salary increases by $2000



Clarification of Symbols and Function

Question

"So capital X is a feature matrix small a is a function and why is the output the value of this function is a vector am I correct?"

Answer

- $\mathbf{X}$: The **feature matrix** (input data). **Correct**.
- $\mathbf{a}$: The **parameter vector** (learned weights). This is *not* a function. **Incorrect**.
- $\mathbf{y}$: The **true target output vector**. **Correct**.
- $f(\mathbf{X})$: The **prediction function**. Its output is the vector of predictions, $\hat{\mathbf{y}}$. **Correct**.

The core equation of the linear model is:

$$\hat{\mathbf{y}} = f(\mathbf{X}) = \mathbf{X}\mathbf{a}$$

For our calculated values:

$$\hat{\mathbf{y}} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 2000 \end{bmatrix} = \begin{bmatrix} 2000 \\ 4000 \\ 6000 \end{bmatrix} = \mathbf{y}$$

The value of $\mathbf{a}$ is:

$$\boxed{\mathbf{a} = \begin{bmatrix} 0 \\ 2000 \end{bmatrix}}$$

The prediction function for a new data point is:

$$f(\mathbf{x}) = \mathbf{x} \cdot \mathbf{a} = a_0 + a_1 \cdot (\text{Experience}) = 2000 \cdot (\text{Experience})$$

Meaning of X^TX, X^Ty, and $(X^TX)^{-1}$

Question

"What is the meaning of X transpose X and X transpose Y and X transpose X inverse, related with covariance correlation etc."

Answer

The Gram Matrix: $X^T X$

- **Size:** $(n \times n)$
- **Meaning:** Proportional to the **covariance matrix** of the features. The off-diagonals indicate feature correlation (multicollinearity).

$$\text{Covariance Matrix} \propto \frac{1}{m} X_c^T X_c$$

where X_c is the mean-centered feature matrix.

The Covariance Vector: $X^T y$

- **Size:** $(n \times 1)$
- **Meaning:** Proportional to the **covariance** between each feature and the target variable y.

$$\text{Covariance}(X_j, y) \propto (X^T y)_j$$

The Inverse Gram Matrix: $(X^T X)^{-1}$

- **Meaning:** The **precision matrix**. It adjusts for correlations between features, isolating their unique contributions.

The Complete Solution: $\mathbf{a} = (X^T X)^{-1} X^T y$

This formula calculates the optimal coefficients by:

1. Measuring the raw relationship between features and target ($X^T y$).
2. Adjusting this relationship for the internal covariance structure of the features themselves ($(X^T X)^{-1}$).