Introduction to Multivariate Calculus

1 Introduction

Understanding multivariate calculus requires a solid grasp of linear algebra because of the heavy use of vectors, matrices, and higher-dimensional spaces. Concepts like gradients, Jacobians, and Hessians are all expressed in terms of matrices and vectors, making linear algebra essential for mastering this area of calculus.

2 Functions of Multiple Variables

A function of multiple variables takes several inputs and produces a single output. For example, consider the function:

$$f(x, y) = x^2 + y^2$$

This function takes two inputs, x and y, and maps them to a single output $f(x, y)$. Such functions can represent a wide range of phenomena, such as the temperature at different points on a surface, or the height of a landscape at different geographical coordinates.

Example:

Consider the function $f(x, y) = x^2 - 2xy + y^2$. The output depends on the values of both x and y. If we plug in specific values for x and y, say $x = 1$ and $y = 2$, we can compute:

$$f(1, 2) = 1^2 - 2(1)(2) + 2^2 = 1 - 4 + 4 = 1$$

3 Partial Derivatives

Partial derivatives measure the rate of change of a function with respect to one variable, while holding the other variables constant. For a function $f(x, y)$, the partial derivatives with respect to x and y are denoted as:

$$\frac{\partial f}{\partial x} \quad \text{and} \quad \frac{\partial f}{\partial y}$$

These represent how the function f changes when x or y are varied independently.

Example:

Given the function $f(x, y) = x^2 + y^2$, the partial derivatives are:

$$\frac{\partial f}{\partial x} = 2x, \quad \frac{\partial f}{\partial y} = 2y$$

Thus, if we want to know the rate of change of f with respect to x when $x = 1$ and $y = 2$, we compute:

$$\frac{\partial f}{\partial x}(1, 2) = 2(1) = 2$$

Similarly, the partial derivative with respect to y at the same point is:

$$\frac{\partial f}{\partial y}(1, 2) = 2(2) = 4$$

4 Critical Points

Critical points occur where the gradient of a function is zero. These points often correspond to local maxima, minima, or saddle points.

For a function $f(x, y)$, the gradient is the vector of partial derivatives:

$$\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right)$$

To find the critical points, we solve:

$$\nabla f = 0$$

Example:

Consider the function $f(x, y) = x^2 + y^2$. The gradient is:

$$\nabla f = (2x, 2y)$$

Setting the gradient to zero, we get the system of equations:

$$2x = 0, \quad 2y = 0$$

Solving these, we find the critical point at $(0, 0)$.

5 Vectorized Gradients

In higher dimensions, the gradient is a vector consisting of the partial derivatives with respect to each variable. For a function $f(x_1, x_2, \ldots, x_n)$, the gradient is:

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n} \right)$$

This vector points in the direction of the steepest ascent of the function.

Example:

For the function $f(x, y, z) = x^2 + y^2 + z^2$, the gradient is:

$$\nabla f = (2x, 2y, 2z)$$

6 Hessian Matrix

The Hessian matrix extends the concept of second-order derivatives to multivariate functions. It provides information about the curvature of the function in multiple directions, and is particularly useful in optimization to classify critical points as local minima, maxima, or saddle points.

For a function $f(x_1, x_2, \ldots, x_n)$, the Hessian is the square matrix of second-order partial derivatives:

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

The Hessian provides critical information about the local behavior of a function. For example, the eigenvalues of the Hessian can tell us whether a point is a local minimum (positive eigenvalues), local maximum (negative eigenvalues), or a saddle point (mixed signs).

Example:

Consider the function $f(x, y) = x^2 + xy + y^2$. The Hessian matrix is computed by taking the second-order partial derivatives of f:

$$\frac{\partial^2 f}{\partial x^2} = 2, \quad \frac{\partial^2 f}{\partial y^2} = 2, \quad \frac{\partial^2 f}{\partial x \partial y} = 1$$

Thus, the Hessian matrix for this function is:

$$H(f) = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

The Hessian matrix provides information about the curvature of the function at any given point. For this function, the Hessian matrix is constant (it does not depend on x or y), indicating that the curvature is the same at every point.

By analyzing the eigenvalues of the Hessian, we can determine the nature of the critical points. In this case, both eigenvalues are positive, meaning the function has a local minimum.

7 Jacobians

The Jacobian matrix generalizes the concept of the derivative for vector-valued functions. If we have a vector-valued function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$, where each component of $\mathbf{f}$ is a function of multiple variables, the Jacobian is the matrix of all first-order partial derivatives:

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

Example:

Consider the function $\mathbf{f}(x, y) = \begin{pmatrix} x^2 \\ xy \end{pmatrix}$. The Jacobian matrix is:

$$J(\mathbf{f}) = \begin{pmatrix} 2x & 0 \\ y & x \end{pmatrix}$$

8 Simple Linear Regression

Simple linear regression is a method used to model the relationship between a single input feature x and a single output y. The model is represented by the equation:

$$y = w_0 + w_1 \cdot x + \epsilon$$

where w_0 is the y-intercept, w_1 is the slope or gradient of the function, and ϵ is the error term.

We are given a set of N data points $(x_i, y_i)_{i=1}^N$. We let $\hat{y}(x) = w_0 + w_1 \cdot x$ denote the prediction of the linear model (without the noise term). Then for w_0, w_1 the least squares estimate looks like the following function,

$$\mathcal{L}(w_0, w_1) = \frac{1}{2N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 = \frac{1}{2N} \sum_{i=1}^{N} (w_0 + x_i \cdot w_1 - y_i)^2$$

The objective function $\mathcal{L}$ is called the loss function; sometimes called the cost function or energy function. Geometrically this corresponds to looking for a straight line such that when the actual data points are projected on the line the sum of squares of the error terms ϵ or the 'residual errors' are minimised. For these reason the objective function is

usually called the *least squares* or *residual sum of squares*. The estimate for (w_0, w_1) is called the *least squares estimate*.

8.1 Deriving the Least Squares Estimate

(Simple) linear regression is one of those rare cases where we can actually find the solution to the loss function analytically, using multivariate calculus. First we take the partial derivatives of the loss function $\mathcal{L}(w_0, w_1)$ with respect to w_0 and w_1,

$$\frac{\partial \mathcal{L}}{\partial w_0} = \sum_{i=1}^{N} (w_0 + w_1 \cdot x_i - y_i)$$

$$\frac{\partial \mathcal{L}}{\partial w_1} = \sum_{i=1}^{N} (w_0 + w_1 \cdot x_i - y_i)\, x_i$$

To obtain the solution for (w_0, w_1) we set the partial derivatives to 0 and solve the resulting system of equations.

$$w_0 + w_1 \cdot \frac{\sum_i x_i}{N} = \frac{\sum_i y_i}{N}$$

$$w_0 \cdot \frac{\sum_i x_i}{N} + w_1 \cdot \frac{\sum_i x_i^2}{N} = \frac{\sum_i x_i y_i}{N}$$

This system of equations is not too difficult to solve but to make it even easier for us to see we define the following quantities,

$$\bar{x} = \frac{\sum_i x_i}{N}$$

$$\bar{y} = \frac{\sum_i y_i}{N}$$

$$\widehat{\text{var}}(x) = \frac{\sum_i x_i^2}{N} - \bar{x}^2$$

$$\widehat{\text{cov}}(x, y) = \frac{\sum_i x_i y_i}{N} - \bar{x} \cdot \bar{y}$$

Full derivation:

For w_0

$$w_0 = \frac{\sum_i y_i}{N} - w_1 \cdot \frac{\sum_i x_i}{N}$$

$$w_0 = \hat{y} - w_1 \cdot \bar{x}$$

For w_1,

$$\frac{\sum_i y_i}{N} \cdot \frac{\sum_i x_i}{N} - w_1 \cdot \frac{\sum_i x_i}{N} \cdot \frac{\sum_i x_i}{N} + w_1 \cdot \frac{\sum_i x_i^2}{N} = \frac{\sum_i x_i y_i}{N}$$

$$\bar{x} \cdot \bar{y} - w_1 \cdot \bar{x}^2 + w_1 \cdot \frac{\sum_i x_i^2}{N} = \frac{\sum_i x_i y_i}{N}$$

$$w_1 \cdot \left(\frac{\sum_i x_i^2}{N} - \bar{x}^2 \right) = \frac{\sum_i x_i y_i}{N} - \bar{x} \cdot \bar{y}$$

$$w_1 \cdot \widehat{\text{var}}(x) = \widehat{\text{cov}}(x, y)$$

$$w_1 = \frac{\widehat{\text{cov}}(x, y)}{\widehat{\text{var}}(x)}$$

Then the least squares estimate for (w_0, w_1) is given by:

$$w_1 = \frac{\widehat{\text{cov}}(x, y)}{\widehat{\text{var}}(x)}$$

$$w_0 = \hat{y} - w_1 \cdot \bar{x}$$

Example

Distance (km)	**Commute Time (min)**
2.7	25
4.1	33
1.0	15
5.2	45
2.8	22

Table 1: Distances and Commute Times Dataset

Try compute the least squares estimate for (w_0, w_1) by hand using this simple dataset.

8.2 The Hessian Matrix of the Least Squares Estimate

The Hessian matrix is a square matrix of second-order partial derivatives. In the context of least squares regression, the Hessian helps us understand the curvature of the error surface, which is crucial for optimization methods such as gradient descent. For our sum of squared errors function $\mathcal{L}(w_0, w_1)$, the Hessian matrix provides information about the second-order behavior of the objective function.

The Hessian matrix H for $\mathcal{L}(w_0, w_1)$ is defined as:

$$H = \begin{bmatrix} \frac{\partial^2 \mathcal{L}}{\partial w_0^2} & \frac{\partial^2 \mathcal{L}}{\partial w_0 \partial w_1} \\ \frac{\partial^2 \mathcal{L}}{\partial w_1 \partial w_0} & \frac{\partial^2 \mathcal{L}}{\partial w_1^2} \end{bmatrix}$$

Let us compute each element of the Hessian matrix for the least squares problem:

- The second-order partial derivative with respect to w_0 is:

$$\frac{\partial^2 \mathcal{L}}{\partial w_0^2} = N$$

- The mixed partial derivative with respect to w_0 and w_1 is:

$$\frac{\partial^2 \mathcal{L}}{\partial w_0 \partial w_1} = \sum_{i=1}^{N} x_i$$

- The second-order partial derivative with respect to w_1 is:

$$\frac{\partial^2 \mathcal{L}}{\partial w_1^2} = \sum_{i=1}^{N} x_i^2$$

Thus, the Hessian matrix for the least squares function $\mathcal{L}(w_0, w_1)$ is:

$$H = \begin{bmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{bmatrix}$$

This Hessian matrix tells us about the curvature of the sum of squared errors function. Since the Hessian is positive definite (as long as the input data points x_i are not all identical), the function $\mathcal{L}(w_0, w_1)$ is convex. This guarantees that any critical point we find by setting the gradient to zero will be a global minimum, ensuring that the least squares method finds the best possible fit for the data.

Hessian of the least squares plotted

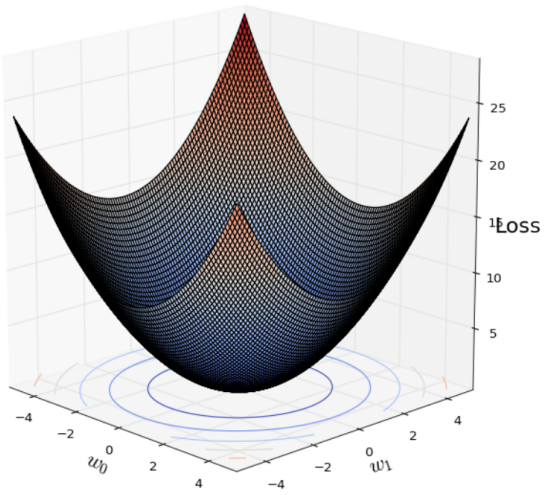

Figure 1: Example Hessian of the least squares