Project "Hessian Matrices and : "
Title
IB3702 Mathematics for Machine Learning

Evertjan Karman John Stegink

15 November, 2025

1 Introduction

The basis of a machine learning algorithm that it tries to predict the right output using a certain input. First the algorithm will have to be trained using correct data. During the training process, the difference between the predicted value and the actual value must be minimized. A cost function is used to quantize the difference, this difference must be minimized. To find the minimum value, the machine learning algorithm iterates until it has found the minimum value. The methods for this iteration are numerous, ow which the most well known algorithm is gradient descent (sections 3.1 and 3.2). For the training to be as effective as possible it is necessary to find the minimum in little iterations. The method for finding the minimum that is discussed in this report is the use of Hessian Matrices (section 3.4). Both gradient descent and the Hessian matrix make use of the Newton method (sections 3.3 and 3.5).

First a description of gradient descent and Newton's will be given for functions using one variable, this is to make the principle clear. Normally for machine learning, 1 variable is not sufficient, the loss function mostly contains multiple variables. The gradient descent, Newton's method and Hessian matrices will be described from the calculus point of view. The linear algebra part will not be discussed (especially eigenvalue and eigenvectors).

Using the Hessian Matrix finds the minimum in far less iterations than using gradient descent. The problem of using Hessian Matrices is that calculating a Hessian Matrix is much more complicated and time consuming than using just the derivative as is done using gradient descent.

2 Preliminaries

2.1 Notation

2.2 Concepts

2.3 Techniques

- Calculating derivative
- Calculating second derivatie
- Computing eigen values

2.4 Problems

2.4.1 Problem 1

2.4.2 Problem 2

Text...

3 Methods

3.1 Gradient descent with one variable

Assume that we have a continuous function f defined on R (fig 1):

Assume that f is also differentiable with derivative $f'(x)$.
We also have a starting point x_0.
Then we get x_1 by subtracting $f'(x_0) \cdot \alpha$ from x, where α is called the learning rate, which we can choose before doing this procedure.
Usual values for α are 0.01 or 0.05.
We iterate this, so that we get an array which is recursively defined as:

$$x_{k+1} = x_k - f'(x_k) \cdot \alpha$$

This array will converge to the minimum of f. The pitchfalls here are, that the procedure may end in a local minimum, while f has a stronger minimum elsewhere.
Or with a less than optimal choice for the learning rate, the array could even diverge.

3.2 Gradient descent with two or more variables

With a function $f(x,y)$ of more variables, we can determine the gradient:

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

The method is the same, but where we took the derivative for one variable, we will now take the gradient, and the recursive definition of our array (x_k, y_k) becomes:

$$(x_{k+1}, y_{k+1}) = (x_k, y_k) - \nabla f(x,y) \cdot \alpha$$

3.3 Newton's method with one variable

Newton's method finds the zeroes of a function f. Because we're interested in finding a minimum of f, Newton's method will help us find the zero of it's derivative f' (fig 2).

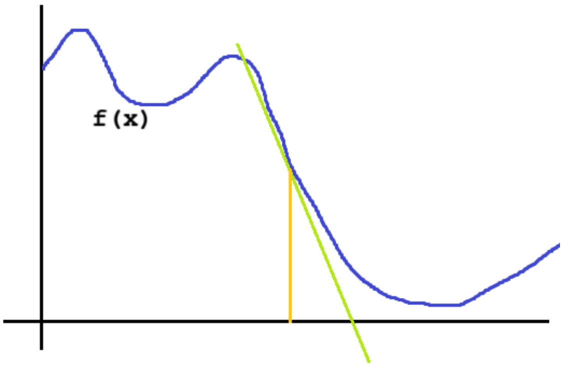

Figure 1: A continuous function f defined on R

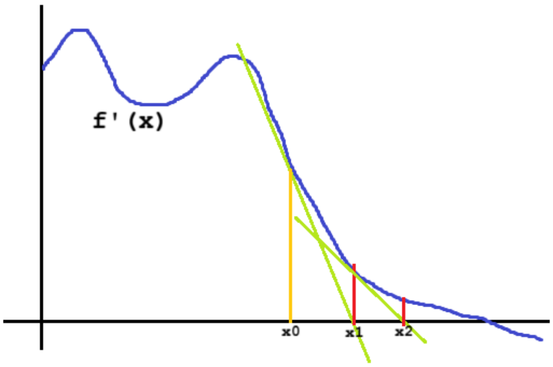

Figure 2: Newton's method

Geometrically, when you have the graph of f, and you have a starting point x_0, we start by drawing the tangent line.
Then we see where this tangent line intersects with the x-axis. That will be x_1.
By iterating this procedure we get an array $(x_k)_{k=0,1,\ldots}$.

From the geometrical aspect of the procedure, we can give a formula between x_{k+1} and x_k:

$$x_{k+1} = x_k - (f'(x_k)/f''(x_k))$$

(that is for finding the zero of f') The idea is that the array (x_k) converges to the value x where $f'(x) = 0$.

In order to know if $f'(x)$ points to a minimum of f, we need to look at the second derivative $f''(x)$:
$f''(x) > 0 \Rightarrow$ f has a minimum at x
$f''(x) < 0 \Rightarrow$ f has a maximum at x
$f''(x) = 0 \Rightarrow$ inconclusive, perhaps an inflection point

3.4 Hessian matrix

As explained in section 3.1 the derivative of a function determines the slope of the tangent line. When the slope is positive the value of the function is increasing and when it is negative the value of the function is decreasing. When the slope is 0 the value of the function is neither rising nor decreasing. This means for this point the function has reached a maximum or minimum value.

To determine whether the found extreme is a maximum or a minimum we can use the curvature of the function. This curvature can be calculated using the second order derivative (the derivative of the derivative). When the value is negative at the extreme, the extreme is a maximum when the second derivatie is negative, it is a minimum when the second derivative is positive. When it is zero no conclusion can be made about the kind of extreme.

This knwolegde can be used, together with the Newton Method, to find minima and maxima for multi variable functions. To be sure that you only find the minimum, the second order derivative must be positive. To determine this, a Hessian matrix is constructed. The Hessian matrix contains all the second order derivates for all combinations of the variables.

Say we have function $f(x, y)$ of 2 variables.

Then here we have it's Hessian matrix:

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

A concrete example for the function $f(x) = x^3 + y^3 + 2xy$. The first order derivates are:

$\frac{\partial f}{\partial x} = 3x^2 + 2y \quad \frac{\partial f}{\partial y} = 3y^2 + 2x$

The second order derivatives are:

$\frac{\partial^2 f}{\partial x^2} = 6x \quad \frac{\partial^2 f}{\partial x \partial y} = 2$
$\frac{\partial^2 f}{\partial y \partial x} = 2 \quad \frac{\partial^2 f}{\partial y^2} = 6y$

This creates the following Hessian Matrix:

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} 6x & 2 \\ 2 & 6y \end{bmatrix}$$

3.5 Newton's method with two or more variables

Say we have function $f(x, y)$ of 2 variables.

Then here we have it's Hessian matrix:

$$H_f(x, y) = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{pmatrix}.$$

Now at a given point (x,y) we'll calculate it's eigenvalues $\lambda_1, \lambda_2, \ldots$
(A 2 by 2 matrix would have at most two eigenvalues)

If the gradient has value (0,0) at point (x,y) then:

- If all the eigenvalues of H_f at (x,y) are positive, it's a minimum
- If all the eigenvalues of H_f at (x,y) are negative, it's a maximum
- In other cases, it's inconclusive

In Newton's method generalized to more than one variables, the formula for the next point is:

$$(x_{k+1}, y_{k+1}) = (x_k, y_k) - (H_f^{-1}(x_k, y_k) \cdot \nabla f(x_k, y_k))$$

3.6 Example of a function of two variables

We will look at this function (fig 3):

$$f(x, y) = 85 - \frac{1}{90} x^2 (x - 6) y^2 (y - 6)$$

Visually we see a possible minimum near point $(x, y) = (4, 4)$.

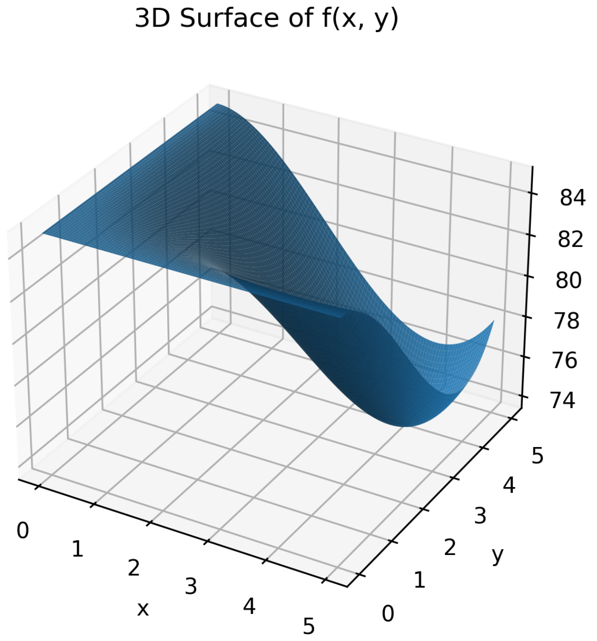

Figure 3: 3D surface of $f(x, y)$

4 Numerical Examples

Text...

5 Collaboration

Text...

6 Reflection

6.1 Student a

Text...

6.2 Student b

Text...