Machine Learning - Foundations

Tutorial - Week 7

IIT Madras Online Degree

Outline

1. Gradient Descent

2. Gradient Descent vs Newton's Method

3. Comparative Examples

Gradient Descent

https://www.geogebra.org/m/uyvajpsy

Gradient Descent vs Newton's Method

Gradient Descent Update Rule

Update Rule

$$\boxed{x_{n+1} = x_n - \eta f'(x_n)}$$

Newton’s Method’s Update Rule

Update Rule

$$\boxed{x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}}$$

Comparative Examples

Example 1

Find the value of x that minimizes the following function:

$$f(x) = (3x - 9)^2$$

Traditional approach

$$f(x) = (3x - 9)^2$$
$$f'(x) = 2(3x - 9)3$$

$$f(x) = (3x - 9)^2$$
$$f'(x) = 2(3x - 9)3$$

Equating $f'(x)$ to zero, we get: $x = 3$

Traditional approach

$$f(x) = (3x - 9)^2$$
$$f'(x) = 2(3x - 9)3$$

Equating $f'(x)$ to zero, we get: $x = 3$

Putting $x = 3$ in $f(x)$, we get $f(x) = 0$.

Newton's method

$$f(x) \quad = \quad (3x-9)^2$$

Newton's method

$$
\begin{aligned}
f(x) &= (3x-9)^2 \\
f'(x) &= 2(3x-9)3
\end{aligned}
$$

Newton’s method

$$
\begin{aligned}
f(x) &= (3x-9)^2 \\
f'(x) &= 2(3x-9)3 \\
f''(x) &= 18
\end{aligned}
$$

Newton method when $x_0 = 2$

Iteration 1:

$$\begin{aligned} x_1 &= x_0 - \frac{f'(x_0)}{f''(x_0)} \\ x_1 &= 2 - \frac{18(2) - 54}{18} \\ x_1 &= 3 \end{aligned}$$

Newton method when $x_0 = 2$

Iteration 1:

$$\begin{aligned} x_1 &= x_0 - \frac{f'(x_0)}{f''(x_0)} \\ x_1 &= 2 - \frac{18(2) - 54}{18} \\ x_1 &= 3 \end{aligned}$$

Only one iteration!

Newton method when $x_0 = 2$

Iteration 2:

$$x_2 = x_1 - \frac{f'(x_1)}{f''(x_1)}$$

$$x_2 = 3 - \frac{18(3) - 54}{18}$$

$$x_2 = 3$$

Newton method when $x_0 = 1$

$$x_1 = x_0 - \frac{f'(x_0)}{f''(x_0)}$$

$$x_1 = 1 - \frac{18(1) - 54}{18}$$

$$x_1 = 3$$

Newton method when $x_0 = 0$

$$x_1 = x_0 - \frac{f'(x_0)}{f''(x_0)}$$

$$x_1 = 0 - \frac{18(0) - 54}{18}$$

$$x_1 = 3$$

Newton method when $x_0 = -100$

$$x_1 = x_0 - \frac{f'(x_0)}{f''(x_0)}$$

$$x_1 = -100 - \frac{18(-100) - 54}{18}$$

$$x_1 = 3$$

Newton method when $x_0 = 100$

$$x_1 = x_0 - \frac{f'(x_0)}{f''(x_0)}$$

$$x_1 = 100 - \frac{18(100) - 54}{18}$$

$$x_1 = 3$$

Gradient Descent method when $x_0 = 2$ and $\eta = 1$

$$x_1 \quad = \quad x_0 - \eta f'(x_0)$$

Gradient Descent method when $x_0 = 2$ and $\eta = 1$

$$\begin{aligned} x_1 &= x_0 - \eta f'(x_0) \\ x_1 &= 2 - 1(18(2) - 54) \end{aligned}$$

Gradient Descent method when $x_0 = 2$ and $\eta = 1$

$$
\begin{aligned}
x_1 &= x_0 - \eta f'(x_0) \\
x_1 &= 2 - 1(18(2) - 54) \\
x_1 &= 20
\end{aligned}
$$

Gradient Descent method when $x_0 = 2$ and $\eta = 1$

$$
\begin{aligned}
x_1 &= x_0 - \eta f'(x_0) \\
x_1 &= 2 - 1(18(2) - 54) \\
x_1 &= 20
\end{aligned}
$$

$$
\begin{aligned}
x_2 &= x_1 - \eta f'(x_1) \\
x_2 &= 20 - 1(18(20) - 54) \\
x_2 &= -286
\end{aligned}
$$

Gradient Descent method when $x_0 = 2$ and $\eta = 0.05$

$$x_1 = x_0 - \eta f'(x_0)$$

$$x_1 = 2 - 0.05(18(2) - 54)$$

$$x_1 = 2.9$$

Gradient Descent method when $x_0 = 2$ and $\eta = 0.05$

$$x_1 = x_0 - \eta f'(x_0)$$
$$x_1 = 2 - 0.05(18(2) - 54)$$
$$x_1 = 2.9$$

$$x_2 = x_1 - \eta f'(x_1)$$
$$x_2 = 2.9 - 0.05(18(2.9) - 54)$$
$$x_2 = 2.99$$

Gradient Descent method when $x_0 = 2$ and $\eta = 0.05$

$$x_1 = x_0 - \eta f'(x_0)$$
$$x_1 = 2 - 0.05(18(2) - 54)$$
$$x_1 = 2.9$$

$$x_2 = x_1 - \eta f'(x_1)$$
$$x_2 = 2.9 - 0.05(18(2.9) - 54)$$
$$x_2 = 2.99$$

$$x_3 = x_2 - \eta f'(x_2)$$
$$x_3 = 2.99 - 0.05(18(2.99) - 54)$$
$$x_3 = 2.999$$

Example 2

$$f(x) = x^3 - x^2 - x + 5$$

Example 2

$f(x) = x^3 - x^2 - x + 5$

Traditional: $f'(x) = 0$ gives $x = 1, -1/3$

Example 2

$f(x) = x^3 - x^2 - x + 5$

Traditional: $f'(x) = 0$ gives $x = 1, -1/3$

Gradient Descent with $x_0 = 0.75$ and $\eta = 0.25$

Example 2

$f(x) = x^3 - x^2 - x + 5$

Traditional: $f'(x) = 0$ gives $x = 1, -1/3$

Gradient Descent with $x_0 = 0.75$ and $\eta = 0.25$

GD I1: $x_1 = 0.953$

Example 2

$f(x) = x^3 - x^2 - x + 5$

Traditional: $f'(x) = 0$ gives $x = 1, -1/3$

Gradient Descent with $x_0 = 0.75$ and $\eta = 0.25$

GD I1: $x_1 = 0.953$

GD I2: $x_2 = 0.998$

Example 2

$f(x) = x^3 - x^2 - x + 5$

Traditional: $f'(x) = 0$ gives $x = 1, -1/3$

Gradient Descent with $x_0 = 0.75$ and $\eta = 0.25$

GD I1: $x_1 = 0.953$

GD I2: $x_2 = 0.998$

Newton's Method with $x_0 = 0.75$

Example 2

$f(x) = x^3 - x^2 - x + 5$

Traditional: $f'(x) = 0$ gives $x = 1, -1/3$

Gradient Descent with $x_0 = 0.75$ and $\eta = 0.25$

GD I1: $x_1 = 0.953$

GD I2: $x_2 = 0.998$

Newton's Method with $x_0 = 0.75$

NM I1: $x_1 = 1.075$

Example 2

$f(x) = x^3 - x^2 - x + 5$

Traditional: $f'(x) = 0$ gives $x = 1, -1/3$

Gradient Descent with $x_0 = 0.75$ and $\eta = 0.25$

GD I1: $x_1 = 0.953$

GD I2: $x_2 = 0.998$

Newton's Method with $x_0 = 0.75$

NM I1: $x_1 = 1.075$

NM I2: $x_2 = 1.004$

Example 2

$f(x) = x^3 - x^2 - x + 5$

Traditional: $f'(x) = 0$ gives $x = 1, -1/3$

Gradient Descent with $x_0 = 0.75$ and $\eta = 0.25$

GD I1: $x_1 = 0.953$

GD I2: $x_2 = 0.998$

Newton's Method with $x_0 = 0.75$

NM I1: $x_1 = 1.075$

NM I2: $x_2 = 1.004$

NM I3: $x_3 = 1.001$

Newton Method: caveats

- For higher dimensional functions, it requires computing inverse of Hessian matrix. Newton method does not work if the Hessian is not invertible.

Newton Method: caveats

- For higher dimensional functions, it requires computing inverse of Hessian matrix. Newton method does not work if the Hessian is not invertible.
- It may not converge. For some functions and some starting points, it may enter an infinite cycle.

Newton Method: caveats

- For higher dimensional functions, it requires computing inverse of Hessian matrix. Newton method does not work if the Hessian is not invertible.
- It may not converge. For some functions and some starting points, it may enter an infinite cycle.
- It may converge to a saddle point instead of a local minimum.

Newton Method: caveats

- For higher dimensional functions, it requires computing inverse of Hessian matrix. Newton method does not work if the Hessian is not invertible.
- It may not converge. For some functions and some starting points, it may enter an infinite cycle.
- It may converge to a saddle point instead of a local minimum.
- It takes more time per iteration and is more computation and memory intensive.

Thank you.