Machine learning II
Master Data Science
Winter Semester 2023/24

Prof. Tim Downie

Workshop 10
Steepest descent algorithm

with solutions

Exercise 1 One-dimensional optimisation

The function minimised the in the lecture demonstration was $f(x) = 3x^4 + 5x^3 - 20x^2 + 8x^1 + 10$. The code is given in the file `ML2_Workshop10.R`

(a) Work through this this example yourself.

(b) Define a new function called `sinf` which corresponds to $f(x){=}\sin(x)$.
Copy the 1-dim minimisation code and adapt it for this problem. You will need to define $f'(x)$

(c) Try different starting values to see how this affects the found minimum.

(d) Try adapting the step length `s` so that the convergence is quicker. For example try increasing the step size slightly with iteration number.

(e) Use $x = 7.8$ as the starting point. Run a couple of interations one by on inspecting the value of x afterwards each iteration. Then run 10 iterations in one go. What do you notice about the values of x? Can you explain why? Continue until you have reached the local minimum.

(f) Try finding a local minimum for $f(x) = e^{x^2+2x-4}$. You will need to use the chain rule to calculate the derivative. $\frac{dy}{dx} = (2x+2)e^{x^2+2x-4}$

Exercise 2 Two-dimensional optimisation

The code for the second demonstration is also given in `Workshop10.R`

(a) Work through this code, trying different starting points.

(b) The outline code has been given for this part. The function to minimise is
$f(x) = x_1^2 + x_2^2 + 2x_1 - 4x_2 - 1$. Calculate the partial derivatives and complete the code to find the approximate minimum.

(c) Use calculus to obtain the exact location of the function minimum. $(-1, 2)$

Exercise 3 Linear model parameter optimisation

The code for Exercise 3 defines a quadratic regression model of the form

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + e_i$$

Where e_i is the error term with a $N(0, 0.2^2)$ distribution. The n=25 observations are simulated in the code. Usually the parameters β_1, β_2, and β_3 are calculated using exact solutions to the so called normal equations, giving the least squares estimates. In this exercise you will obtain the results using Newton-Raphson minimisation.

The loss function for a linear model is the residual sum of squares

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{25} (y_i - \beta_1 - \beta_2 x_i - \beta_3 x_i^2)^2$$

Derive the partial derivatives. As a hint, the third partial derivative is given for you, and is also given in the source file.

$$\frac{\partial L}{\partial \beta_3} = -2 \sum_{i=1}^{25} x_i^2 (y_i - \beta_1 - \beta_2 x_i - \beta_3 x_i^2)$$

$$\frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^{25} (y_i - \beta_1 - \beta_2 x_i - \beta_3 x_i^2)$$

$$\frac{\partial L}{\partial \beta_2} = -2 \sum_{i=1}^{25} x_i (y_i - \beta_1 - \beta_2 x_i - \beta_3 x_i^2)$$

Complete the code and confirm that the parameters converge to values close to those those specified in the simulation code.

Homework Exercises

Exercise 4 Newton-Raphson method

Finding a root of f.

A function can be approximated around a given point x_0 using the first order Taylor series

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0)$$

This is a linear polynomial, which uses the function, and the derivative value evaluated at the point x_0

Rearrange the expression so that it is in the form

$$x \approx \tag{1}$$

$$x \approx x_0 + \frac{f(x)}{f'(x_0)} - \frac{f(x_0)}{f'(x_0)}$$

If x_1 is a root of $f(x)$, then $f(x_1) = 0$. Using this and Equation (1), show that x_1 approximately satisfies the following equation: $x_1 \approx x_0 - \dfrac{f(x_0)}{f'(x_0)}$.
x_1 will not be an exact root, because of the approximation, but multiple iterations of the following formula usually give a better an better approximation for a root of $f(x)$.

This leads directly to the Newton-Raphson method for finding a root of f:

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$$

Let $f(x) = 3x^2 - 4x - 5$ and $x_0 = -2$. Carry out 2 iterations of the N-R method to obtain x_2.

$f'(x) = 6x - 4$

i	x_i	$f(x_i)$	$f'(x_i)$
0	*-2*	*15*	*-16*
1	*-1.0625*	*2.636719*	*-10.375*
2	*-0.8083584*		

After 2 iterations the current approximation of the root is -0.8083584, the quadratic formula (German: abc-Formel) gives the exact root as -0.7862996.

Finding a minimum of f

A very similar method can be used to minimise a function using the second order Taylor series.

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2$$

This is a quadratic polynomial, which uses the function, first and second derivative value evaluated at the point x_0

Differentiate both sides of the above equation to obtain

$$f'(x) \approx \tag{2}$$

$$f'(x) \approx f'(x_0) + f''(x_0)(x - x_0)$$

Use an analogous argument to the above, to obtain an iterative procedure that finds a point where $f'(x_i) = 0$

$$\begin{aligned} f'(x_1) \approx 0 \approx f'(x_0) + x_1 f''(x_0) - x_0 f''(x_0) \\ x_1 \approx \frac{1}{f''(x_0)}\left(x_0 f''(x_0) - f'(x_0)\right) \\ x_1 \approx x_0 - \frac{f'(x_0)}{f''(x_0)} \end{aligned}$$

$$x_{i+1} = x_i - \frac{f'(x_i)}{f''(x_i)}$$

Show that when $f(x) = 3x^2 - 4x - 5$ and $x_0 = -2$, the true minimum is found after just one iteration. The reason for this is that f is quadratic and the second order Taylor series is exact for quadratic functions.

$f''(x) = 6$
$x_1 = -2 - \dfrac{(-16)}{6} = \frac{2}{3}$

Using exact methods $f'(x) = 6x - 4 \stackrel{!}{=} 0$ *gives* $x = \frac{2}{3}$

Exercise 5 Chain rule in back propagation

(a) Using the NN from Lecture 8, obtain the following partial derivatives

$$\frac{\partial R}{\partial w_{12}^{(2)}} \quad \text{and} \quad \frac{\partial R}{\partial b_1^{(2)}}$$

$$\begin{aligned} \frac{\partial R}{\partial w_{12}^{(2)}} &= \frac{\partial R}{\partial a_1^{(2)}} \frac{\partial a_1^{(2)}}{\partial w_{12}^{(2)}} \\ \frac{\partial R}{\partial w_{11}^{(2)}} &= -2\left(y - a_1^{(2)}\right) a_2^{(1)} \end{aligned}$$

$$\frac{\partial R}{\partial b_1^{(2)}} = \frac{\partial R}{\partial a_1^{(2)}} \frac{\partial a_1^{(2)}}{\partial b_1^{(2)}}$$

$$\frac{\partial a_1^{(2)}}{\partial b_1^{(2)}} = 1$$

$$\frac{\partial R}{\partial b_1^{(2)}} = -2\left(y - a_1^{(2)}\right)$$

(b) Show that the derivative of the sigmoid function $\sigma(v) = (1 + e^{-v})^{-1}$ is $\sigma'(v) = e^{-v}(1 + e^{-v})^{-2}$ and that $\sigma'(v) = \sigma(v)(1 - \sigma(v))$.

Put $w = 1 + e^{-v}$ and apply the chain rule. $\sigma = w^{-1}$

$$\begin{aligned}\sigma'(v) = \frac{d\sigma}{dv} &= \frac{d\sigma}{dw}\frac{dw}{dv} \\ &= -w^{-2}(-e^{-v}) \\ &= (1 + e^{-v})^{-2}e^{-v}\end{aligned}$$