Math Basics for Machine Learning
Graded Assignment 4

Faran Taimoor Butt

Fall 2024

Instructions

This is the fourth graded assignment for the Math Basics for Machine Learning course. It contains two tasks. The instructions, as well as links to supplementary material, are given in the task descriptions.

Provide **detailed solutions** to the tasks in this assignment. Then, save your solution document as a .pdf file and submit it by filling in the corresponding Google form.

In total, you can earn 10 points for this assignment. This score will contribute to you final score for this course.

You must submit your answers by **Monday, October 21, 23:59 Moscow Time**.

Solutions must be typed in LaTeX. Hand-written solutions, as well as late submissions, will not be accepted.

It is the idea that you complete this assignment individually. Do not collaborate or copy answers of somebody else.

Have fun!

1. (4 points) Find and classify all the critical points of the following function

$$f(x,y) = 7x - 8y + 2xy - x^2 + y^3$$

Solution: We will first find the partial derivatives of the function with respect to x and y

$$f(x,y) = 7x - 8y + 2xy - x^2 + y^3$$
$$f_x(x,y) = 7 - 2x + 2y$$
$$f_y(x,y) = -8 + 2x + 3y^2 \tag{1}$$

Now for critical points take

$$f_x(x,y) = 0$$
$$f_y(x,y) = 0$$

Solving for $f_x(x,y)$ gives

$$7 - 2x + 2y = 0$$
$$x = \frac{2y+7}{2} \tag{2}$$

Solving for $f_y(x,y)$ by putting (2) in (1)

$$3y^2 + 2\left(\frac{2y+7}{2}\right) - 8 = 0$$

$$3y^2 + 2y + 7 - 8 = 0$$
$$3y^2 + 2y + 1 = 0$$

Factorizing gives

$$3y^2 + +3y - y - 8 = 0$$
$$3y(y+1) - 1(y+1) = 0$$
$$(3y+1)(y+1) = 0$$
$$y = \frac{1}{3} \tag{3}$$
$$y = -1 \tag{4}$$

Solving for x by putting (3) in (2)

$$x = \frac{2\left(\frac{1}{3}\right) + 7}{2} = \frac{\frac{2+21}{3}}{2} = \frac{2+21}{6}$$

$$x = \frac{23}{6} \tag{5}$$

Solving for x by putting (4) in (2)

$$x = \frac{2(-1) + 7}{2}$$
$$x = \frac{5}{2} \tag{6}$$

So we have critical points

$$C_1 = \left(\frac{23}{6}, \frac{1}{3}\right) \tag{7}$$

$$C_2 = \left(\frac{5}{2}, -1\right) \tag{8}$$

For second order partial derivative test, we need to calculate the second derivatives of the following values $f_{xx}(x,y), f_{yy}(x,y), f_{xy}(x,y)$

$$f_{xx}(x,y) = \frac{d^2 f}{dx^2}(7 - 2x + 2y^2)$$

$$f_{xx}(x,y) = -2 \tag{9}$$

$$f_{yy}(x,y) = \frac{d^2 f}{dx^2}(3y^2 + 2x - 8)$$

$$f_{xx}(x,y) = 6y \tag{10}$$

$$f_{xy}(x,y) = \frac{d^2 f}{dx^2}(3y^2 + 2x - 8)$$

$$f_{xx}(x,y) = 2 \tag{11}$$

Putting (9),(10),(11) in the equation of derivative

$$D = f_{xx}(x,y) f_{yy}(x,y) - (f_{xy}(x,y))^2$$

$$D = (-2)(6y) - 2^2$$

$$D = -12y - 4 \tag{12}$$

Now to perform the second-order derivative test

Putting C_1 in (12) gives

$$D_1 = -12\left(\frac{1}{3}\right) - 4$$

$$D_1 = -8 < 0$$

As $f_{xx}(x,y) = \frac{1}{3} > 0$ and $D_1 = -8 < 0$ so critical point C_1 is a **saddle point**.

Putting C_2 in (12) gives

$$D_2 = -12(-1) - 4 = 12 - 4 = 8$$

$$D_2 = 8 > 0$$

As $f_{xx}(x,y) = -1 < 0$ and $D_2 = 8 > 0$ so critical point C_2 is a **local maximum point**.

2. (6 points) Fitting a machine learning model means finding the optimal values of its parameters, which comes down to optimizing some loss function $\mathcal{L}$. Suppose you want to fit a linear regression model of the form

$$y = w_0 + w_1 \cdot x_1 + ... + w_n \cdot x_n$$

You want to do so by minimizing the following loss function:

$$\mathcal{L}(w) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \cdot ||w||^2$$

Here, $\hat{y}_i$ is model's prediction for example i and $||w||^2$ is the l_2 norm of the unknown weights vector $w = (w_0, w_1, ..., w_n)^T$. The effect of adding this extra term to the loss function is that it forces us to choose small values for the unknown coefficients. The larger the value of the hyperparameter λ, the larger is the effect of regularization.

(a) (4 points) Find the gradient of $\mathcal{L}$. *Hint: it might be useful to re-write $\mathcal{L}$ using matrix notation and use matrix calculus.*

Solution:
As the Loss function is given by

$$\mathcal{L}(w) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \cdot ||w||^2$$

We need to take the first derivative to calculate the gradient of $\nabla\mathcal{L}(w)$

$$\hat{y} = Xw$$

So

$$\mathcal{L}(w) = \sum_{i=1}^{n} (y_i - Xw_i)^2 + \lambda \cdot ||w||^2$$

$$\mathcal{L}(w) = \sum_{i=1}^{n} (y_i - Xw_i)^2 + \lambda \cdot ||w||^2$$

As X, y are data points and w is unknown parameters writing the above equation in matrix form

$$d\mathcal{L}(w) = d\left((y - Xw)^T (y - Xw) + \lambda w^T w\right)$$

$$d\mathcal{L}(w) = d\left((y - Xw)^T (y - Xw)\right) + d\left(\lambda w^T w\right)$$

Let's divide our equation into $d\mathcal{L}(w)_1$ and $d\mathcal{L}(w)_2$

$$d\mathcal{L}(w)_1 = d\left[(y - Xw)^T (y - Xw)\right]$$

Using Mulitplication rule

$$d(AB) = (dA)B + A(dB)$$

$$= \left(d[y - Xw]^T\right)(y - Xw) + (y - Xw)^T d[y - Xw]$$

$$= \left(-d[Xw]^T\right)(y - Xw) - (y - Xw)^T (d[Xw])$$

As

$$(AB)^T = B^T A^T$$

$$= \left(-d[w^T X^T]\right)(y - Xw) - (y - Xw)^T X\,(d[w])$$
$$= [-(dw^T)X^T(y - Xw)]^T - (y - Xw)^T X dw$$
$$= -(y - Xw)^T X dw - (y - Xw)^T X dw$$
$$\nabla\mathcal{L}(w)_1^T \cdot dw = -2(y - Xw)^T X \cdot dw$$

So taking transpose of $\nabla\mathcal{L}(w)_1^T$ gives

$$\nabla\mathcal{L}(w)_1 = -2X^T(y - Xw) \tag{13}$$

Now for $d\mathcal{L}(w)_2$

$$d\mathcal{L}(w)_2 = d(\lambda w^T w)$$

As

$$w^T w = \sum_{i=1}^{n} w_i^2$$

$$dw^T w = \lambda 2w$$

$$\nabla\mathcal{L}(w)_2 \cdot dw = 2\lambda w \cdot dw$$

So

$$\nabla\mathcal{L}(w)_2 = 2\lambda w \tag{14}$$

Now adding equations (13) and (14) for $\nabla\mathcal{L}(w)$

$$\nabla\mathcal{L}(w) = \nabla\mathcal{L}(w)_1 + \nabla\mathcal{L}(w)_2$$

So gradient of $\mathcal{L}(w)$ is

$$\nabla\mathcal{L}(w) = -2X^T(y - Xw) + 2\lambda w$$

(b) (2 points) Using the gradient obtained above, derive the expression for the optimal model weights w that minimize the loss $\mathcal{L}$.

Solution: To find optimal model weights w that minimize the loss $\mathcal{L}$ we need to put the gradient of the loss function equal to zero
So,

$$\nabla\mathcal{L}(w) = 0$$
$$-2X^T(y - Xw) + 2\lambda w = 0$$
$$-2X^T y + 2X^T Xw + 2\lambda w = 0$$

Dividing both sides by 2 gives

$$-\frac{2}{2}X^T y + \frac{2}{2}X^T Xw + \frac{2}{2}\lambda w = \frac{0}{2}$$

$$-X^T y + X^T Xw + \lambda w = 0$$
$$X^T Xw + \lambda w = X^T y$$

$$w(X^T X + \lambda I) = X^T y$$

So

$$w^* = (X^T X + \lambda I)^{-1} X^T y$$

So w^* are the optimal weights that minimize the loss function $\mathcal{L}$.