

# 18s1: COMP9417 Machine Learning and Data Mining

---

**Lectures:** Linear Models for Regression

**Topic:** Questions from lecture topics

**Version:** review answers

## Review

How to solve the partial derivatives with respect to one value?

when

$$f(x, y) = a_1x^2y^2 + a_4xy + a_5x + a_7$$

$$\frac{\partial f(x, y)}{\partial x} = 2x(a_1y^2) + a_4y + a_5$$

$$\frac{\partial f(x, y)}{\partial y} = 2y(a_1x^2) + a_4x$$

when

$$f(x, y) = a_1x^2y^2 + a_2x^2y + a_3xy^2 + a_4xy + a_5x + a_6y + a_7$$

what  $\frac{\partial f(x, y)}{\partial x}$  will be?

$$\frac{\partial f(x, y)}{\partial x} = 2a_1xy^2 + 2a_2xy + a_3y^2 + a_4x + a_5$$

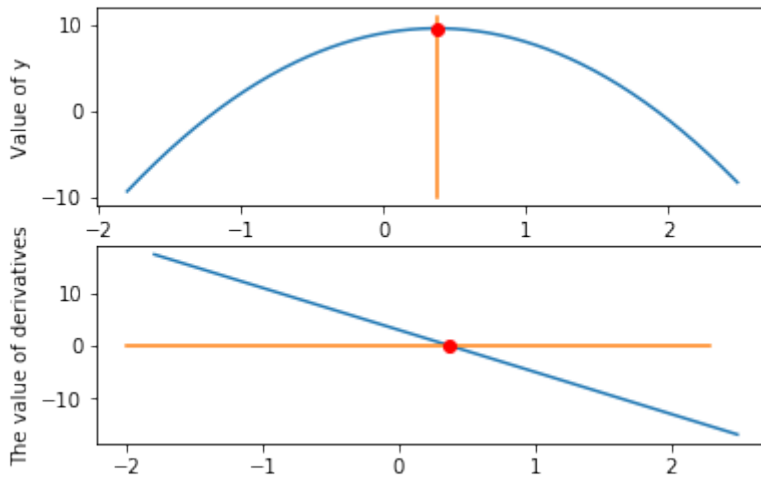
what  $\frac{\partial f(x, y)}{\partial y}$  will be?

$$\frac{\partial f(x, y)}{\partial y} = 2a_1x^2y + a_2x^2 + 2a_3xy + a_4x + a_6$$

How to solve optimization problems for Quadratic function? For example,  $y = -4x^2 + 3x + 3$ . The solution is minimum or maximum?

$$\frac{\partial y}{\partial x} = -8x + 3$$

when  $x = \frac{8}{3}, \frac{\partial y}{\partial x} = 0$  and  $y$  is minimum.



When  $f(x,y)$  is Quadratic function, how to solve optimization problems?

Solving optimization problems is not easy. There are some cases without solution and some have many solution. But in our cases, it is easy as the derivatives of the loss function is linear.

**Definition 22 (Hessian)**

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be continuously differentiable. The **Hessian**  $\nabla^2 f : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$  of  $f$  at  $\mathbf{x}$  is

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

- The Hessian  $\nabla^2 f(\mathbf{x})$  is an  $n$  by  $n$  matrix
- If  $f$  is **twice continuously differentiable** at  $\mathbf{x}$  then

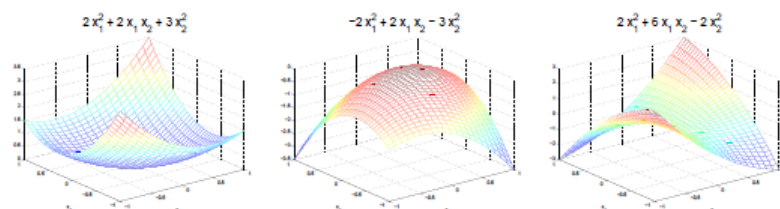
$$\frac{\partial^2 f(\mathbf{x})}{\partial x_i x_j} = \frac{\partial^2 f(\mathbf{x})}{\partial x_j x_i} \quad \text{for all } i \neq j$$

That is the Hessian matrix  $G = \nabla^2 f(\mathbf{x})$  is **symmetric** ( $G^T = G$ )

#### Example 4 (Unconstrained stationary points)

Find the unconstrained stationary points of

- 1  $f_1(\mathbf{x}) = 2x_1^2 + 2x_1x_2 + 3x_2^2$
- 2  $f_2(\mathbf{x}) = -2x_1^2 + 2x_1x_2 - 3x_2^2$
- 3  $f_3(\mathbf{x}) = 2x_1^2 + 6x_1x_2 - 2x_2^2$



### Second order necessary conditions continued

#### Corollary 6 (Local maximizer)

$\bar{\mathbf{x}}$  a local maximizer  $\implies \nabla f(\bar{\mathbf{x}}) = \mathbf{0}$  and  $\nabla^2 f(\bar{\mathbf{x}})$  negative semi-definite

Proof.

- $\bar{\mathbf{x}}$  a local maximizer of  $f \iff \bar{\mathbf{x}}$  a local minimizer of  $-f$
- Second order necessary conditions for  $-f \implies -\nabla f(\bar{\mathbf{x}}) = \mathbf{0}$  and  $-\nabla^2 f(\bar{\mathbf{x}})$  positive semi-definite
- $-\nabla^2 f(\bar{\mathbf{x}})$  positive semi-definite  $\iff \nabla^2 f(\bar{\mathbf{x}})$  negative semi-definite  $\square$

#### Corollary 7 (Contrapositive of second order necessary conditions)

Let  $\mathbf{x}^*$  be a stationary point of  $f \in C^2(\mathbb{R}^n)$ , so  $\nabla f(\mathbf{x}^*) = \mathbf{0}$ . Then

- 1  $\nabla^2 f(\mathbf{x}^*)$  has a negative eigenvalue  $\implies \mathbf{x}^*$  is **not** a local minimizer
- 2  $\nabla^2 f(\mathbf{x}^*)$  has a positive eigenvalue  $\implies \mathbf{x}^*$  is **not** a local maximizer
- 3  $\nabla^2 f(\mathbf{x}^*)$  indefinite (both a positive eigenvalue and a negative eigenvalue)  $\implies \mathbf{x}^*$  is **neither** a local minimizer **nor** a local maximizer

What is the loss function of linear regression?

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^n (y_i - (w_0 + w_1 x_i))^2$$

### Q1

A *univariate linear regression model* is a linear equation  $y = a + bx$ . Learning such a model requires fitting it to a sample of training data so as to minimize the error function  $\mathcal{L} = \sum_{i=1}^n (y_i - (w_0 + w_1 x_i))^2$ . To find the best parameters  $a$  and  $b$  that minimize this error function we need to find the error

gradients  $\frac{\partial \mathcal{L}}{\partial w_0}$  and  $\frac{\partial \mathcal{L}}{\partial w_1}$ . So we need to derive these expressions by taking partial derivatives, set them to zero, and solve for  $w_0$  and  $w_1$ .

First we write the loss function for the univariate linear regression  $y = w_0 + w_1x$  as

$$\begin{aligned}\mathcal{L} &= \frac{1}{N} \sum_{n=1}^N (y_n - (w_0 + w_1x_n))^2 \\ &= \frac{1}{N} \sum_{n=1}^N (y_n - (w_0 + w_1x_n))(y_n - (w_0 + w_1x_n)) \\ &= \dots \\ &= \frac{1}{N} \sum_{n=1}^N [w_1^2 x_n^2 + 2w_1x_n(w_0 - y_n) + w_0^2 - 2w_0y_n + y_n^2]\end{aligned}$$

At a minimum of  $\mathcal{L}$  the partial derivatives with respect to  $w_0$ ,  $w_1$  should be zero. We will start with  $w_0$ , so first we remove from the above expression all terms not including  $w_0$ .

$$\frac{1}{N} \sum_{n=1}^N [w_0^2 + 2w_1x_nw_0 - 2w_0y_n]$$

Rearrange, taking terms not indexed by  $n$  outside:

$$w_0^2 + 2w_0w_1 \frac{1}{N} \left( \sum_{n=1}^N x_n \right) - 2w_0 \frac{1}{N} \left( \sum_{n=1}^N y_n \right)$$

Taking the partial derivative with respect to  $w_0$  we get:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial w_0} &= 2w_0 + 2w_1 \frac{1}{N} \left( \sum_{n=1}^N x_n \right) - \frac{2}{N} \left( \sum_{n=1}^N y_n \right) \\ E(y) &= \frac{\sum_{n=1}^N y_n}{N} \\ E(x) &= \frac{\sum_{n=1}^N x_n}{N} \\ \widehat{w_0} &= E(y) - w_1 E(x)\end{aligned}$$

Now we do the same for  $w_1$ , first removing all terms not including  $w_1$ :

$$\frac{1}{N} \sum_{n=1}^N [w_1^2 x_n^2 + 2w_1x_nw_0 - 2w_1x_ny_n]$$

Rearrange, taking terms not indexed by  $n$  outside:

$$w_1^2 \frac{1}{N} \left( \sum_{n=1}^N x_n^2 \right) + 2w_1 \frac{1}{N} \left( \sum_{n=1}^N x_n (w_0 - y_n) \right)$$

Taking the partial derivative with respect to  $w_1$  we get:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial w_1} &= 2w_1 \frac{1}{N} \left( \sum_{n=1}^N x_n^2 \right) + \frac{2}{N} \left( \sum_{n=1}^N x_n (w_0 - y_n) \right) \\ &= w_1 \frac{2}{N} \left( \sum_{n=1}^N x_n^2 \right) + \frac{2}{N} \left( \sum_{n=1}^N x_n (\widehat{w}_0 - y_n) \right) \\ &= w_1 \frac{2}{N} \left( \sum_{n=1}^N x_n^2 \right) + \frac{2}{N} \left( \sum_{n=1}^N x_n (E(y) - w_1 E(x) - y_n) \right) \\ &= \dots \\ &= 2w_1 \left[ \frac{1}{N} \left( \sum_{n=1}^N x_n^2 \right) - E(x)E(x) \right] + 2E(y)E(x) - 2 \frac{1}{N} \left( \sum_{n=1}^N x_n y_n \right) \\ &= 2w_1 E(x^2) - 2w_1 (E(x))^2 + 2E(y)E(x) - 2E(xy) \end{aligned}$$

$$\frac{\partial \mathcal{L}}{\partial w_1} = 0$$

$$w_1 (E(x^2) - (E(x))^2) = E(xy) - E(y)E(x)$$

$$\begin{aligned} \widehat{w}_1 &= \frac{E(xy) - E(x)E(y)}{E(x^2) - (E(x))^2} \\ &= \frac{E(xy) - E(x)E(y)}{Var(x)} \\ &= \frac{COV(x, y)}{Var(x)} \\ \widehat{w}_0 &= E(y) - w_1 E(x) \end{aligned}$$

To make sure you know the process, you could try to solve the following loss function for linear

regression with L-2 normalization. The result is similar.

$$\begin{aligned}\mathcal{L} &= \frac{1}{N} \sum_n (y_n - (w_0 + w_1 x_n))^2 + \frac{\lambda(w_1)^2}{2} \\ \widehat{w}_0 &= E(y) - w_1 E(x) \\ \widehat{w}_1 &= \frac{E(xy) - E(x)E(y)}{E(x^2) - (E(x))^2 + \lambda}\end{aligned}$$