CS5014 Machine Learning

$x^T Ax$ Quadratic form

Lei Fang

01/02/2021

1 Introduction

A quadratic form is defined as

$$x^T Ax,$$

where $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ is a $n \times 1$ (column) vector, A is a $n \times n$ square matrix (the corresponding i-th row and j-th column entry is a_{ij}. Therefore, $x^T Ax$ is a function that maps a vector input x to a scalar: $f : R^n \to R$.

According to matrix multiplication rule, the quadratic form can be expanded as

$$x^T Ax = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

essentially the weighted sum of all the second order products between x_i, x_j; and the weights are given by a_{ij}.

Example 1. $f(x) = x_1^2 + x_2^2$ is a quadratic form, as

$$f(x) = x^T \begin{bmatrix} 1, 0 \\ 0, 1 \end{bmatrix} x = x^T x.$$

Its surface plot in R^3 and contour plot is show below in Fig 1. Note that a contour plot shows all the levels sets: *i.e.* all the $x \in R^n$ in the input space such at $f(x) = c$ is a constant.

Example 2. $f(x) = 4x_1^2 + x_2^2$ is a quadratic form as well:

$$f(x) = [x_1, x_2] \begin{bmatrix} 4, 0 \\ 0, 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

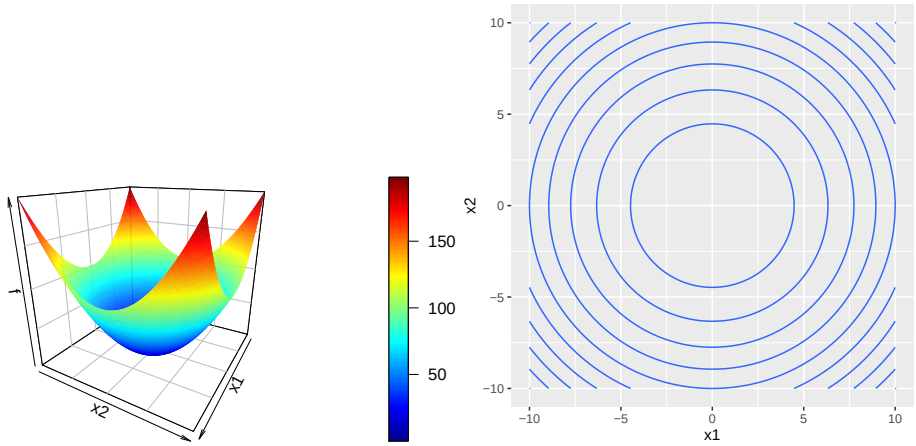

Figure 1: surface plot of a circular paraboloid

The plots shown in Fig 2 and 3 suggest the contours become ellipses. So scaling a_{11} (or other diagonal entries) has the effect of compressing that direction. If $A = \begin{bmatrix} 1, 0 \\ 0, 4 \end{bmatrix}$, the effect is compressing x_2 direction.

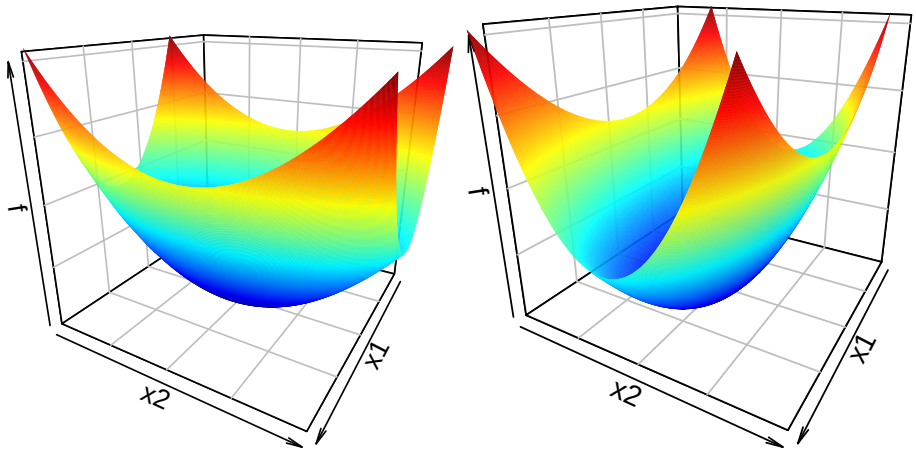

Figure 2: surface plot of two axis-aligned elliptic paraboloid: scaling diagonal entry of A has the effect of compressing the bowl

Example 3.

$$A = \begin{bmatrix} 3, 4 \\ 4, 3 \end{bmatrix}$$

The plots shown in Fig 4 suggest the contours become rotated ellipses. Therefore,

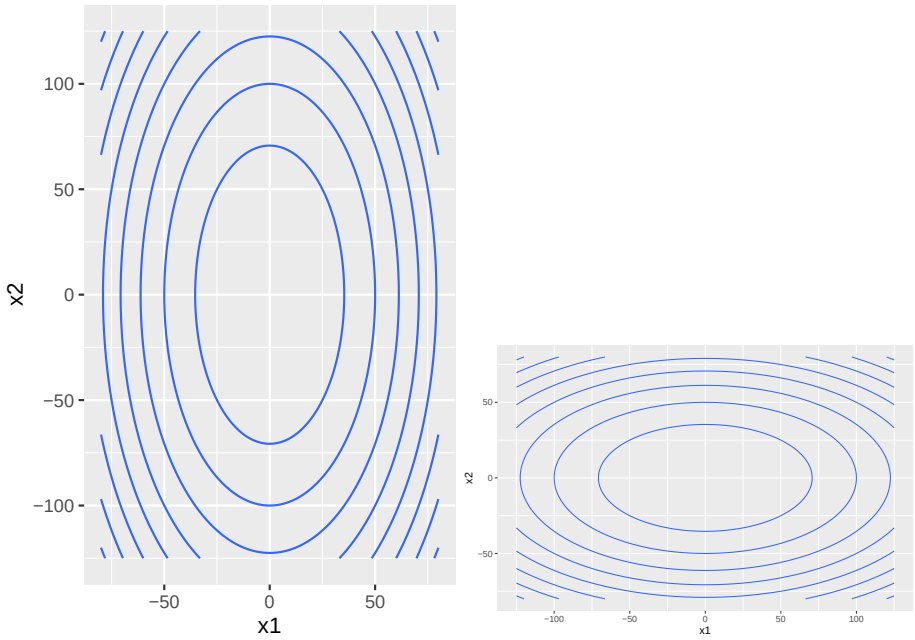

Figure 3: contour plots of two axis aligned elliptic paraboloid

non-diagnonal entries of A rotate the elliptic paraboloid (A has to be either positive definite or negative definite).

1.1 Definite quadratic forms are distance measures

All three examples above are cases where the quadratic forms have a minimum. It turns out that all definite quadratic forms (either positive or negative definite, see Section 3.3), the quadratic forms can be viewed as some forms of distance measure: it measures the distance between x to 0. Example 1 with $A = I$, it measures the Eclidean distance between x and 0. For others, the distances are either discounted or expanded at one direction (axis aligned or other off-axis direction). For example, Example 2 with a compressed x_1 direction will shorten the distance in that direction by a factor: in other words, you have to travel shorter distance in x_1 direction to reach the same height or value of f (comparing with the Eclidean distance). The contour plots show the "adjusted distances".

For positive definite forms, f is positively correlated with the distance measure: the further away x is from the center 0, the larger value f takes (so it has a minimum at the centre). For negative definite forms, f is negatively correlated with the distance measure: the further away x is from 0, the smaller value f takes (f is facing down; so it has a maximum at the centre).

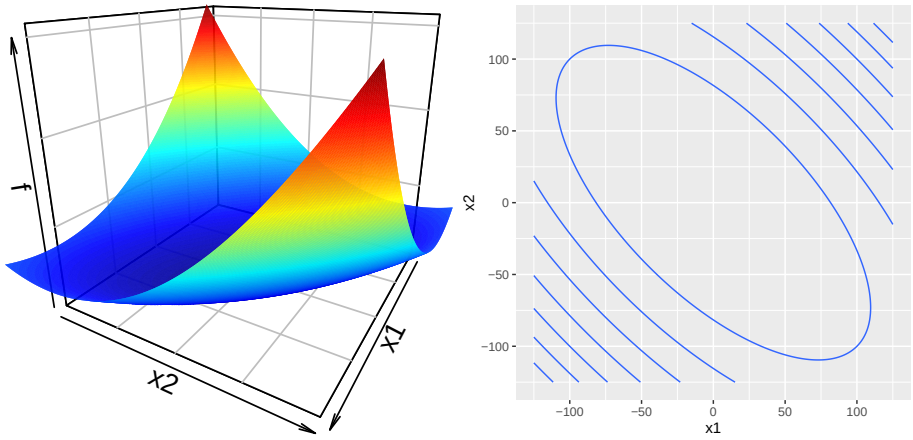

Figure 4: surface plot of a rotated elliptic paraboloid: non-diagonal entries rotate the ellipse

2 Common quadratic forms in machine learning models

Quadatric form, $x^T Ax$, is used a lot in machine learning models. The following are three cases.

Linear regression: the loss function of a linear regression is

$$L(\theta) = (y - X\theta)^T (y - X\theta),$$

which is a quadratic form of θ.

Gaussian distribution: the kernel of a multivariate Gaussian is a quadratic form, namely

$$-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu),$$

where μ, Σ are the mean and variance-covariance matrix respectively. The kernel is a quadratic form of input x.

Taylor expansion: Taylor's expansion contains quadratic forms as well:

$$T(x) = f(a) + \nabla_x f(a)(x - a) + \frac{1}{2!} \underbrace{(x - a)^T \nabla_x^2 f(a)(x - a)}_{q.f.} + \ldots,$$

where $\nabla_x f(a)$ is the gradient of f and $\nabla_x^2 f(a)$ is the hessian matrix. Note that we define gradients as row vectors, so the second term $\nabla_x f(a)(x - a)$ is an inner product. The expansion approximates a multivariate $R^n \to R$ function $f(x)$ by a polynomial function $T(x)$. The third term is a quadratic form of input x.

3 Gradient and Hessian of a quadratic form

It should be easy to remember the following results if you compare it with univariate quadratic function $f(x) = ax^2 = axa$, whose gradient is $f' = 2ax$ and second order derivative is $f'' = 2a$.

3.1 Gradient of a quadratic form

The gradient of $f(x) = x^T Ax$ is

$$\nabla_x f(x) = x^T (A + A^T)$$

Note that as we have adopted the convention that gradients are row vectors, hence the gradient is written as $2x^T A$, a $1 \times n$ vector.

If A is symmetric, then $A = A^T$, therefore

$$\nabla_x f(x) = x^T (A + A^T) = 2x^T A$$

3.2 Hessian of a quadratic form

For a general function $f(x) : R^n \to R$, the Hessian matrix of f is defined as

$$\nabla_x^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

It can be observed that a Hessian matrix is symmetric.

The Hessian or gradient of the gradient of $f(x) = x^T Ax$ is

$$\nabla_x^2 f(x) = A^T + A$$

If A is symmetric, the Hessian is

$$\nabla_x^2 f(x) = 2A.$$

3.3 Maximum and minimum of a quadratic form

- If $\nabla_x^2 f(x)$ is positive definite, $f(x) = x^T Ax$ has a minimum;
- If $\nabla_x^2 f(x)$ is negative definite, $f(x) = x^T Ax$ has a maximum;

Positive definite matrix: if A is positive definite, then $x^T Ax > 0$ for all $x \in R^n$

Negative definite matrix: if A is negative definite, then $x^T Ax < 0$ for all $x \in R^n$

You should compare the above results with univariate quadratic function $f(x) = ax^2$. For such a quadratic function,

- if $f''(x) = 2a > 0$, then f has a minimum;
- if $f''(x) = 2a < 0$, then f has a maximum.

To further understand the connection, you probably need to know

$$u^T \nabla_x^2 f(x) u$$

is the **second directional derivative** of f at direction u, if $u \in R^n$ is a unit vector, i.e. $||u||_2^2 = u^T u = 1$. Therefore, the requirement for A to be posisitve definite, it essentially means f's second derivative is possitive at all directions. In other words, the univariate test is positive at all directions, so it is going to be minimum. The logic applies to the negative definite case.