Least Squares

SML: Data Assimilation (Spring 22)
Low Qiuling

August 23, 2022

1 Why is the least squares mentioned so much?

This is based on an easy exercise from the professor's notes, page 11. All we need is multivariate calculus and linear algebra so even an undergraduate freshman can do this. ☺

Before we jump to the question, let's discuss the least squares method as a big picture. Least squares problem is a special case of minimization. You certainly will hear about it if you are asked to plot a simple linear regression. It is usually the first algorithm one comes across when venturing into machine learning. All it entails is finding the equation of the best fit line through a bunch of data points.

The method of least squares determines $x^* \in \mathbb{R}^{(p+1)}$ such that the norm of $Ax - b$ is minimized, and

$$x^* = \arg\min ||Ax - b||^2 \tag{1}$$

As $J > p + 1$, we are dealing with *overdetermined* systems of equation, which means having more equations than variables. This implies that we could not exactly solve them as there might not exist a $x^* \in \mathbb{R}^{(p+1)}$ such that $Ax - b = 0 \in \mathbb{R}^J$.

Therefore, the best we can do is to minimize the norm of **residual** (error), $r = b - Ax \in \mathbb{R}^J$. In the least square, we minimize the norm of the residual and find equation (1). The better the line fits the data, the smaller the residuals on the average.

This is not the only method to approximate the solution to an overdetermined systems, but it is elegant for several reasons.

1. Statistically, it is a sound reason to pick this route when data vector b is contaminated by Gaussian noise.
2. Mathematically, this is alluding as $||x||^2$ is a smooth function of x. Hence, the solution to the least squares is a linear function of b.
3. Geometrically, it is pretty neat as the least square solution is the projection of b onto the span of A. Meanwhile, the residual at the least square solution is orthogonal to the span of A.

2 Solving the exercise

(a) Consider $F : \mathbb{R}^{p+1} \to \mathbb{R}$, with

$$F(x) = ||Ax - b||^2 = \left\langle Ax - b, Ax - b \right\rangle$$

Show that

$$\nabla F(x^*) = 2A^T(Ax - b) \in \mathbb{R}^{(p+1)}$$

$A^T \in \mathbb{R}^{(p+1)\times J}$ denotes the transpose of A. This function $\nabla F(x)$ is called the gradient of F. The gradient is the vector formed by the partial derivatives of a scalar function.

There are two ways to do it. The easy way is using (i) multivariate calculus.
(i) Let $F(x) = g\left(h(x)\right)$ where $h(x) = Ax - b$ and $g(u) = ||u||^2$.

Find the derivative of g and h:

$$h'(x) = A \qquad g'(u) = 2u^T$$

The Chain Rule tells us that:

$$\underbrace{F'(x)}_{1\times(p+1)} = \underbrace{g'\left(h(x)\right)}_{1\times J} \underbrace{h'(x)}_{J\times(p+1)}$$

Using Chain Rule:

$$\begin{aligned} F'(x) &= g'\left(h(x)\right) \cdot h‘(x) \\ &= 2\left(Ax - b\right)^T A \end{aligned}$$

If we use the convention that the gradient is a column vector, then we have: (Recall $\left((Ax - b)^T\right)^T = (Ax - b)$

$$\nabla F(x^*) = F'(x)^T = 2A^T(Ax - b) \in \mathbb{R}^{(p+1)}$$

The second way is to a bit more mechanical but still doable.

(ii) Recall $F(x) = ||Ax - b||^2 = \left\langle Ax - b, Ax - b \right\rangle = (Ax - b)^T(Ax - b)$ and this is a scalar product in $\mathbb{R}^J$.

$$\begin{aligned} F(x^*) &= (Ax - b)^T(Ax - b) \\ &= (x^T A^T - b^T)(Ax - b) \quad (\text{Recall}(AB)^T = B^T A^T) \\ &= x^T A^T Ax - x^T A^T b - b^T Ax + b^T b \quad (\text{Just expand}) \end{aligned}$$

We have three following gradients:

$$\nabla(x^T A^T A x) = 2A^T A x, \qquad \nabla(x^T A^T b) = A^T b, \qquad \nabla(b^T A x) = A^T b$$

$$\begin{aligned} \nabla F(x) &= 2A^T A x - A^T b - A^T b \\ &= 2A^T A x - 2A^T b \\ &= 2A^T (Ax - b) \end{aligned}$$

(b) Show

$$\mathcal{H}_F(x^*) = 2A^T A \in \mathbb{R}^{(p+1)\times(p+1)}$$

where $\mathcal{H}$ is the Hessian matrix, the second partial derivatives.

This is easy. Simply expand $\nabla F(x)$ and do the derivative:

$$\begin{aligned} \nabla F(x) &= 2A^T(Ax - b) = 2A^T A x - 2A^T b \\ \therefore \mathcal{H}_F(x^*) &= 2A^T A \end{aligned}$$

(c) Show that $A^T A$ is positive definite if A has the maximum column $rank(A) = p + 1$.

A matrix is positive definite if it is symmetric. Recall the definition of the Hessian. Once we can obtain it, we can instantly tell the matrix, which in this case is, $A^T A$ is symmetric as the Hessian matrix itself is symmetric.

$$\mathcal{H}_F(x) = 2 \underbrace{A^T}_{(p+1)\times J} \underbrace{A}_{J\times(p+1)} \in \mathbb{R}^{(p+1)\times(p+1)}$$

A symmetric matrix is a square matrix that is equal to its transpose.
B is symmetric $\iff B = B^T$.

The fastest method to test for positive semi-definiteness is to check if the quadratic form is semi-positive. Every quadratic form $q \in \mathbb{R}^n$ can be uniquely written in the form, $q(x) = x^T B x \in \mathbb{R}$ which will yield a scalar. This is natural as the second-order behavior of every smooth multi-variable function is described by the quadratic form belonging to the function's Hessian. And this is the result from Taylor's theorem. Check G30 Cal II's notes, page 24-32.

Another thing we need to show is a positive definite matrix is full-rank, which in this case, $rank(A) = p + 1$.

Proposition: Let A be a matrix $\in \mathbb{R}^{n \times n}$. If A is is full-rank, it is positive definite.

Proof. Let's prove by contradiction. Suppose that A is not full-rank. Then its columns are not linearly independent. As a consequence, there is a vector,$x \in \mathbb{R}^{n \times 1}$, $x \neq 0$ such that

$$Ax = 0$$

Multiply both sides by x^T and obtain

$$x^T Ax = 0$$

Since A is positive definite, this is possible only if $x = 0$, which is a contradiction.
Thus, A must be full-rank. □