A Geometric View to Least Squares

Xiaocan Li
lixiaocan2017@ia.ac.cn
Institute of Automation, Chinese Academy of Sciences

1 Abstract

Least squares are the most important and widely used methods to do regression analysis. But the intuition behinds them are not clear. In this short paper, column picture and projection perspective is introduced to give an more explicit explanation.

Keywords: Least squares, column space, projection

2 Introduction

Least squares are learnt from high school to university, but with different depth of explanation. In high school, only formulas of linear form coefficient, slope α and intercept term β are introduced. However, the derivation is ignored during high school. Although during college time, the derivation may have been introduced in statistics, the meaning is quite ambiguous and not intuitive at all. Thus, in this short paper, we will introduce the space projection with minimal error to explain the least squares.

3 High School Version of Least Squares

Think of the problem: we have 3 points in 2D, $(1, 1)$, $(2, 2)$, $(3, 2)$, we want to find a line that fits these 3 points, and we want to minimize the sum of the square y-axis difference, i.e. the least squares. We set the line

$$y = \theta_0 + \theta_1 x = X\theta$$

where $\mathrm{X} = \begin{bmatrix}1 & x\end{bmatrix}$, $\theta = \begin{bmatrix}\theta_0 & \theta_1\end{bmatrix}^T$.

y is the true value, $\mathrm{X}\theta$ is the predicted value, we want to minimize the sum of the square y-axis difference, i.e.

$$min_\theta \|X\theta - y\|_2^2$$

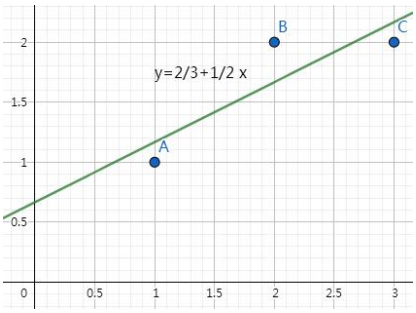

Figure 1: Plot of points and best fit line

In our problem, the cost

$$\begin{aligned} J &= \|e_1\|^2 + \|e_2\|^2 + \|e_3\|^2 \\ &= (\theta_0 + \theta_1 - 1)^2 + (\theta_0 + 2\theta_1 - 2)^2 + (\theta_0 + 3\theta_1 - 2)^2 \end{aligned}$$

We take derivative of J w.r.t. θ_0 and θ_1

$$\begin{aligned} \frac{\partial J}{\partial \theta_0} &= 2(3\theta_0 + 6\theta_1 - 5) = 0 \\ \frac{\partial J}{\partial \theta_1} &= 2(6\theta_0 + 14\theta_1 - 11) = 0 \end{aligned}$$

Then we get

$$\theta = \begin{bmatrix} \frac{2}{3} & \frac{1}{2} \end{bmatrix}^T$$

Thus, $\hat{y} = \frac{2}{3} + \frac{1}{2}\hat{x}$.

Figure 1 shows points and best fit line.

In high school, basically this is the end of the least squares teaching. But we are not satisfied with this non-intuitive method.

Let's see the error $e_i = y - \hat{y} = trueValue - predictedValue$ corresponding to each point

$$\begin{aligned} e_1 &= 1 - (\frac{2}{3} + \frac{1}{2} \times 1) = -\frac{1}{6} \\ e_2 &= 2 - (\frac{2}{3} + \frac{1}{2} \times 2) = \frac{2}{6} \\ e_3 &= 2 - (\frac{2}{3} + \frac{1}{2} \times 3) = -\frac{1}{6} \end{aligned}$$

We arrange the error into a column vector

$$e = \begin{bmatrix} dfrac16 & \frac{2}{6} & -\frac{1}{6} \end{bmatrix}^T$$

Remember we have

$$X = \begin{bmatrix} 1 & x \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 2 \end{bmatrix}$$

Interestingly, we are surprised to find that the 2 column vectors of X are orthogonal to the error vector e:

$$e^T \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = 0$$

$$e^T \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} = 0$$

In brief notation:

$$e^T X = 0$$

Here, the 0 in the right hand side is a vector 0 with size 1×2.

It means the error vector e lies in the left nullspace of X, or the nullspace of X^T. In mathematical words:

$$e \in N(X^T)$$

This is quite an interesting observation, which seems there are more profound explanation waiting for us to discover. Hence, the column picture is introduced.

4 Advanced view: Column picture

In Section 3 we talked about high school version of least squares, which is non-intuitive. But we found that the error vector is orthogonal to the column vector of X.

From now on, we switch the notation X to be A, θ to be x, and y to be b. Thus under the new notation, the problem becomes

$$min_x \|Ax - b\|_2^2$$

Think of a least square problem: We delete the third point (3,2), now we only have 2 points (1,1),(2,2). We will get exactly a line $y = x$ passes through these 2 points. The matrix A and vector b will be

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

$$b = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

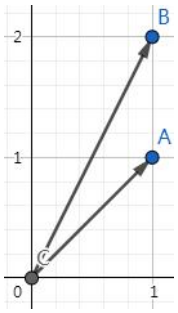

Figure 2: 2 linearly independent 2D vectors

Set A=$\begin{bmatrix} a_1 & a_2 \end{bmatrix}$, $a_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ $a_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

$$Ax = \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = a_1 x_1 + a_2 x_2 = b$$

Apparently, a_1 and a_2 are linearly independent. They are 2 2D vectors that can span the whole 2D space. No matter what 2D vector b is, we can always find a combination coefficient x_1, x_2, in this case $x_1 = 0, x_2 = 1$ to express b under the basis of a_1, a_2, for a_1 and a_2 are linearly independent and can reach everywhere in 2D space.

If a_1 and a_2 are independent, then adding extra column to matrix A does not change the rank, i.e. $rank(A|b) = rank(A)$;

If a_1 and a_2 are dependent, then only vector b that lies in the line of vector a_1(or a_2) can be expressed by a_1 or a_2. Also, adding extra dependent vector b does not change the rank of A either, i.e. $rank(A|b) = rank(A)$.

That's why $rank(A|b) = rank(A)$ is the sufficient and necessary condition for linear equation $Ax = b$ having solution.

Now we add the point (3,2) back, the matrix A will be

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 2 \end{bmatrix}$$

where $a_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$, $a_2 = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}$.

a_1 and a_2 are 2 independent 3D vectors, they can span a 2D plane subspace. Any vector b that lies in the plane subspace that spanned by a_1 and a_2, will have unique exact solution to $Ax = b$.

Column picture is a totally different perspective to treat the least squares, with more clearer and high level than before!

Question: What if vector b is not in the plane spanned by a_1 and a_2, what's the best to approximation a_1 and a_2 can do? What's your intuition?

Orthogonal projection!

5 Projection

Let's look at the projection to a line.

a is our vector that spans a 1D line subspace, and b is a vector. We want to find the best approximation of b in the 1D line subspace spanned by a. Our intuition tells us, the answer is the orthogonal projection $\hat{b}$.

$$\begin{aligned} \hat{b} &= ax \\ e &= b - \hat{b} = b - ax \\ a^T e &= a^T b - a^T ax = 0 \end{aligned}$$

Then we get

$$x = \frac{a^T b}{a^T a} \tag{1}$$

$$\hat{b} = ax = \frac{aa^T}{a^T a} b = P_a b \tag{2}$$

$$P_a = \frac{aa^T}{a^T a} \tag{3}$$

Here P_a is named projection matrix formed by vector a, whose function is to project some vector b to the subspace formed by vector a with best approximation, also means the minimal loss $\|e\|^2$.

Here's some properties of projection matrix P:

$$\begin{aligned} &(1) P^2 = P \\ &(2) P^T = P \\ &(3) rank(P) = rank(a) = 1 \end{aligned}$$

The first property can be explained through geometry: If we do projection once, vector b will become $\hat{b}$, and we project $\hat{b}$ to 1D line subspace again, we will still get $\hat{b}$. In fact, $P^k = P$ holds true for any positive integer $k \geq 1$, for no matter how many times we project, it will always remain the same result.

Alegebraic proof for $P^2 = P$:

$$P^2 = \frac{aa^T}{a^T a} \frac{aa^T}{a^T a} = \frac{a(a^T a)a^T}{(a^T a)^2} = \frac{aa^T}{a^T a} = P$$

Q.E.D.

Proof of property (2) and (3) will leave as exercises for readers.

Now let's look at the projection to a plane. In order to span a plane subspace, we need 2 independent vectors a_1 and a_2.Set $A = \begin{bmatrix} a_1 & a_2 \end{bmatrix}$. Then the best approximation of vector b can be expressed as

$$\hat{b} = a_1 x_1 + a_2 x_2 = Ax$$

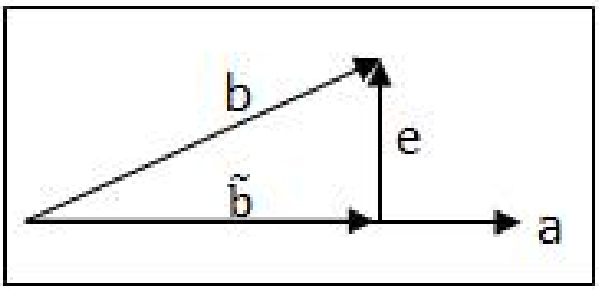

Figure 3: 1D case of projection

The error vector is

$$e = b - \hat{b} = b - Ax$$

According to the definition of orthogonal projection, we have

$$a_1^T e = 0$$
$$a_2^T e = 0$$

We can rewrite it as

$$A^T e = 0 \tag{4}$$

e is in the left nullspace of A, or the nullspace of A^T, $e \in N(A^T)$.

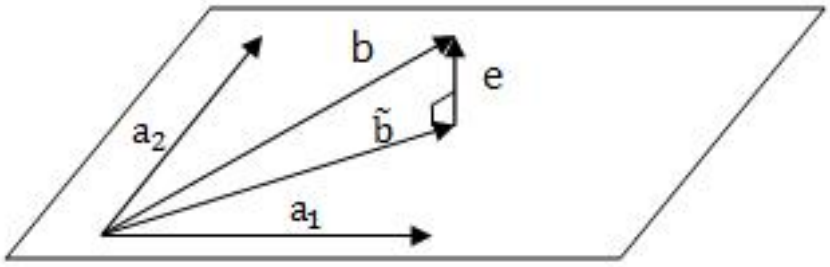

Figure 4: 2 linearly independent 3D vectors

In Figure 4, along vector e is $N(A^T)$, the plane spanned by a_1 and a_2 is R(A). e is orthogonal to a_1 and a_2, therefore, $N(A^T)$ is orthogonal to R(A). We also notice that $N(A^T) \cup R(A) = \{0\}$, dimN=1, dimR=2, so dimN+dimR=3. Thus, N and R are **orthogonal complement** with each other. which is a quite famous lemma in matrix analysis. We proved this in geometric way, which is far more intuitive and easily understood than algebraic method.

Now let's derive the projection matrix P:

$$A^T e = A^T b - A^T A x = 0 \tag{5}$$
$$x = (A^T A)^{-1} A^T b \tag{6}$$
$$\hat{b} = Ax = A(A^T A)^{-1} A^T b = P_A b \tag{7}$$
$$P_A = A(A^T A)^{-1} A^T \tag{8}$$

Equation (5) is called **normal equation** in machine learning community.
Equation (8) is our charismatic projection matrix!
Verify that $A^T e = 0$:

$$\begin{aligned} A^T e &= A^T [A(A^T A)^{-1} A^T b - b] \\ &= A^T A(A^T A)^{-1} A^T b - A^T b \\ &= A^T b - A^T b \\ &= 0 \end{aligned}$$

Q.E.D.
Also, the projection matrix P_A has the following property similar to 1D case:

$$(1) P^2 = P$$
$$(2) P^T = P$$
$$(3) rank(P) = rank(A)$$

Prove the first property:

$$\begin{aligned} P^2 &= A(A^T A)^{-1} A^T A(A^T A)^{-1} A^T \\ &= A(A^T A)^{-1} (A^T A)(A^T A)^{-1} A^T \\ &= A(A^T A)^{-1} A^T \\ &= P \end{aligned}$$

Q.E.D.

Also we can explain property (1) through geometry: If we do projection once, vector b will become $\hat{b}$, and we project $\hat{b}$ to the subspace again, we will still get $\hat{b}$. In fact, $P^k = P$ holds true for any positive integer $k \geq 1$, for no matter how many times we project, it will always remain the same result. We can give this property a fancy name You Only Project Once (YOPO).

Notice that we used $(A^T A)^{-1}$ in projection matrix, is $A^T A$ always invertible? Actually, we have the following lemma:

$$rank(A^T A) = rank(A) \tag{9}$$

Proof: Think of the linear equation $A^T Ax = 0$, we multiply both sides by x^T: $x^T A^T Ax = \|Ax\|^2 = 0$, therefore $Ax = 0$. Which means $A^T Ax = 0$ is equivalent to $Ax = 0$, both of them have the same solution space.

Hence $rank(A^TA) = rank(A)$.

Q.E.D.

Let A to be a $m \times n$(usually m is much greater than n) matrix, then A^TA is a $n \times n$ matrix. If A is full column rank, i.e. rank(A)=n, using lemma (9) we have $rank(A^TA) = n$, which means matrix A^TA is a full rank square matrix, therefore it is invertible.

Lemma: A^TA is invertible i.f.f. A is full column rank.

This lemma can be explained through geometric perspective as well. If A has full column rank, it means the n column vectors of A are linearly independent, and they can form a n-dim hypercube subspace without redundancy, therefore we can find a projection matrix into the n-dim subspace. But if A is not full column rank, which means the n column vectors are dependent, they can not form a n-dim subspace. They can only form a subspace with dimension less than n. Thus we can not project some vector b to n-dim space, since n-dim space does not exists.

6 Revisit our problem

we have 3 points in 2D, $(1,1)$, $(2,2)$, $(3,2)$, we want to find a line that fits these 3 points, and we want to minimize the sum of the square y-axis difference, i.e. the least squares.

Now we use normal equation to find our least square solution: The matrix

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$$

$$x = (A^TA)^{-1}A^Tb = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{7}{3} & -1 \\ 1 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{2}{3} \\ \frac{1}{2} \end{bmatrix}$$

We get the same answer as before!

7 Conclusion

Least squares methods are very important in data science, it can be applied to many areas. We first solved least squares by high school methods, then introduced an advanced perspective, the column picture, to treat the least squares

as a projection problem. If I may summarize this short paper in one sentence, it is:

What least squares do is to find a minimal loss projection using the basis vector under the Euclidean metric.

8 Reference

[1] Linear Algebra, MIT OpenCourseWare, Gilbert Strang.

[2] Least Squares, Wikipedia.