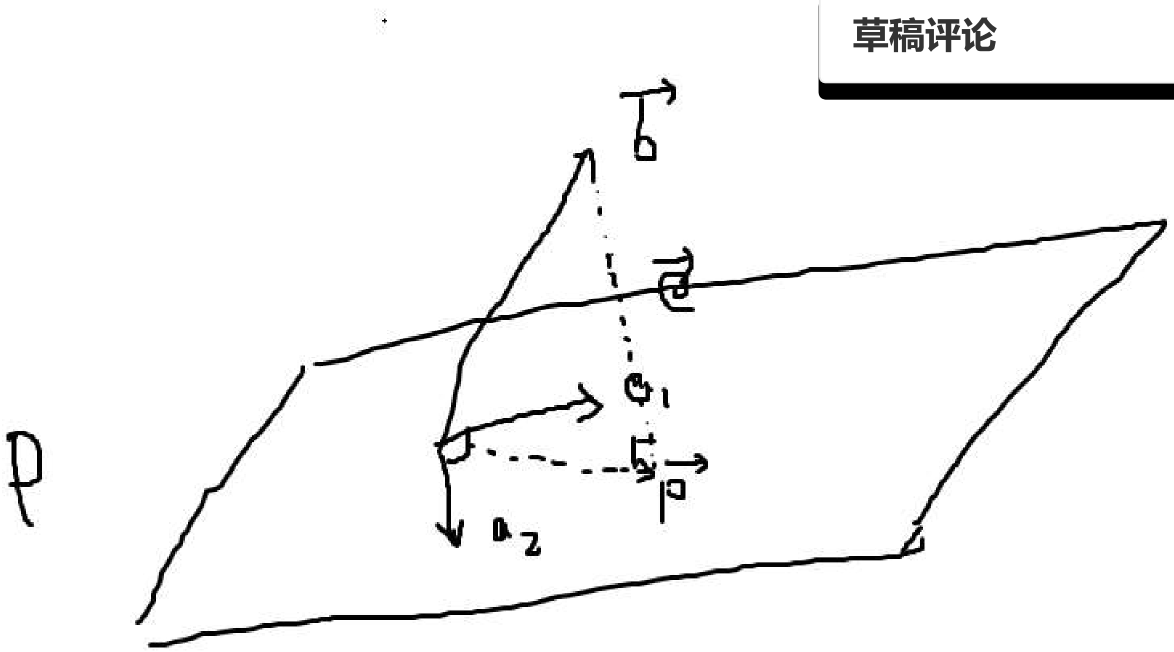


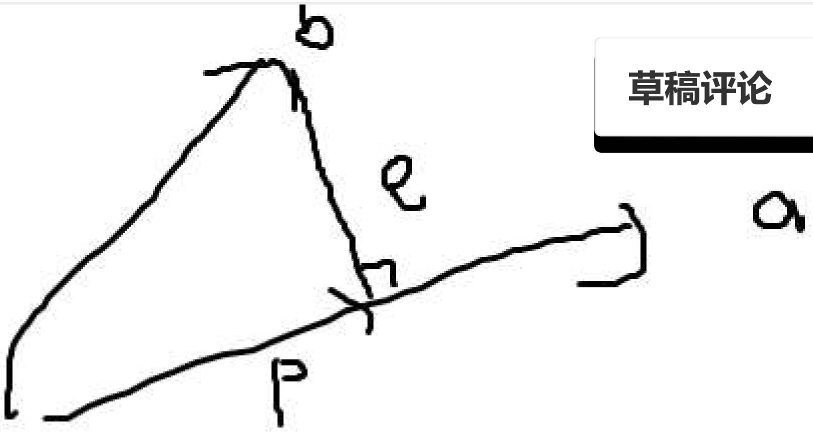
掰开揉碎推导Normal Equation

Normal Equation是一种基础的最小二乘方法，本文将从线性代数的角度来分析Normal Equation（而不是从矩阵求导 matrix derivative 的角度）。

很多作者（特别是智商比较高的）在推导公式的时候有意无意的忽略了思考过程，只留下漂亮的步骤。这让很多读者（比如说我）跟不上节奏，最后一头雾水。本文将从求解“貌似无解”的方程组入手，再讲讲投影（Projection）的使用，最后进入到Normal Equation的应用。我的目的是让和我一样蠢的孩子对 $\vec{\theta} = (A^T A)^{-1} A^T \vec{y}$ 这个重要公式有一个Big Picture——即使忘记了也可以重头推出。

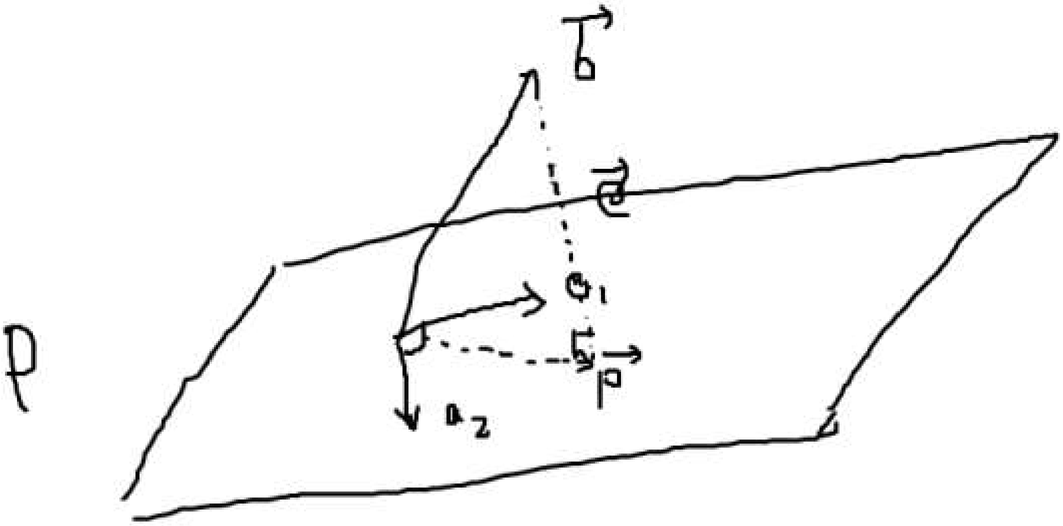
1. 求解不可解的方程组

先看一个在 R^2 中的例子：如图，求一个常数 θ 使 $\theta \vec{a} = \vec{b}$



这个方程明显不可解，因为 $\vec{b}$ 与 $\vec{a}$ 不共线，无法通过对 $\vec{a}$ 数乘得到 $\vec{b}$。

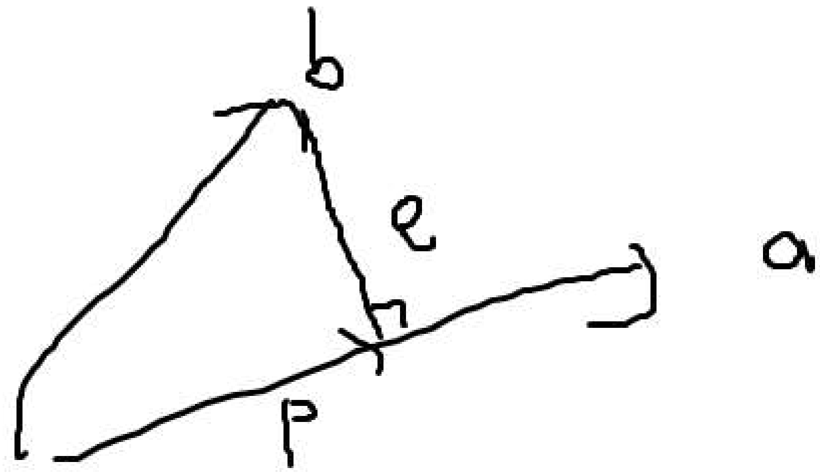
再看一个在 R^3 中的例子：$\vec{a_1}$ 和 $\vec{a_2}$ 是平面 P 的一组基，求出 θ_1 与 θ_2 使 $\theta_1\vec{a_1}+\theta_2\vec{a_2}=\vec{b}$



这个方程也明显不可解，因为 $\vec{b}$ 不在在平面P上，而 $\vec{a_1}$ 与 $\vec{a_2}$ 的线性组合只能得到平面上的向量。

以上两个问题非常的典型，因为在解决实际问题的时候，我们很难得到Perfect Solution，我们只能尽力而为的争取Best Solution。以上两个例子明显没有perfect solution (方向都错了谈何perfect)，那么best solution在哪里呢？

再回到例1.0，那么应该如何寻找 $\theta\vec{a} = \vec{b}$ 的解呢？



最好的方法就是抛弃 $\vec{b}$ 向量中垂直 $\vec{a}$ 的分量，只要计算 θ 使 $\theta\vec{a}$ 等于向量 $\vec{b}$ 在 $\vec{a}$ 方向的分量（即 $\vec{b}$ 在 $\vec{a}$ 上的投影 $\vec{p}$ ），同时把向量 $\vec{b}$ 垂直 $\vec{a}$ 方向的分量称为 $\vec{e}$。

原来的问题 $\theta\vec{a} = \vec{b}$ 变成了求解 $\theta^*\vec{a} = \vec{p}$ （ θ^* 是 θ 的估计量）

因为 $\vec{p}$ 与 $\vec{e}$ 合成了 $\vec{b}$ 向量（ $\vec{e} + \vec{p} = \vec{b}$ ），而且 $\vec{e}$ 垂直于 $\vec{a}$ ($\vec{e}\perp\vec{a}$)，所以有

$$\vec{a}^T(\vec{b} - \theta^*\vec{a}) = 0$$

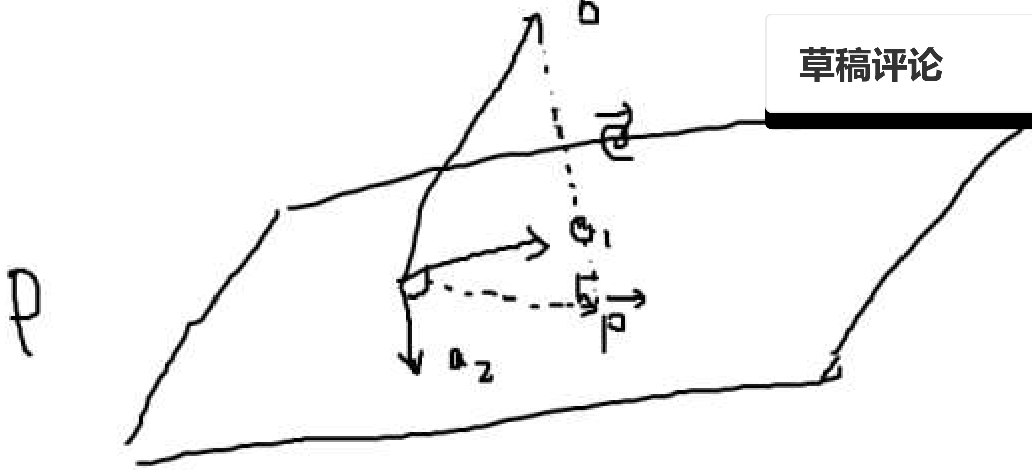
这是一个**非常重要的方程**，后面可以看到Normal Equation可以从这个里推出。

继续改写方程

$$\theta^*\vec{a}^T\vec{a} = \vec{a}^T\vec{b}$$
$$\theta^* = \frac{\vec{a}^T\vec{b}}{\vec{a}^T\vec{a}}$$

如果想求出得到投影 $\vec{p}$ 的投影矩阵P，可以从 $\vec{p} = \theta^*\vec{a}$ 开始推导，发现投影矩阵 P 在形式上就等于乘数 $\theta^* = \dfrac{\vec{a}^T\vec{b}}{\vec{a}^T\vec{a}}$ ，即 $\vec{p}$ 满足 $\vec{p} = P\vec{a}$。

看完了在 R^2 中的例子，再看看投影怎么在 R^3 中解决不可解方程。



草稿评论

平面 P 有基向量 $\vec{a_1}$ 和 $\vec{a_2}$ ，所以平面 P 可以表示成 $\vec{a_1}$ 和 $\vec{a_2}$ 的所有线性组合 $\theta_1\vec{a_1}+\theta_2\vec{a_2}$

，即 $[a_1a_2]\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}$

令向量组成的矩阵 $A=[a_1a_2]$，参数组成的向量 $\vec{\theta}=\theta_{1\ldots n}\ (n=2)$，与平面垂直的误差向量 $\vec{e}=\vec{b}-A\vec{\theta^*}$ ，则在 R^2 中的问题 $\theta\vec{a}=\vec{b}$ 在这里拓展成为了 $A\vec{\theta}=\vec{b}$。相应的，$\theta^*\vec{a}=\vec{p}$ 问题在这里拓展成了 $A\vec{\theta^*}=\vec{p}$ ，其中 $\vec{p}=\vec{\theta_1}\vec{a_1}+\vec{\theta_2}\vec{a_1}$。

因为 $P\perp\vec{e}$ ，而且 $P\in\vec{\theta_1}\vec{a_1}+\vec{\theta_1}\vec{a_2}$ ，所以有以下方程组——

$$\begin{cases}\vec{a_1}^T(\vec{b}-A\vec{\theta^*})=0\\ \vec{a_2}^T(\vec{b}-A\vec{\theta^*})=0\end{cases}$$

整理成矩阵的形式——

$$\begin{bmatrix}\vec{a_1}^T\\ \vec{a_2}^T\end{bmatrix}(\vec{b}-A\vec{\theta^*})=0$$
$$A^T(\vec{b}-A\vec{\theta^*})=0$$

写到这里回头看看 R^2 情景下的核心公式 $\vec{a}^T(\vec{b}-\theta^*\vec{a})=0$ ，可以这家伙换一套马甲又出现了，看来方程 $A^T(\vec{b}-A\vec{\theta^*})=0$ 是一种高维的拓展。

我们继续整理这个公式——

如果你有读过Andrew Ng著名的公开课CS229的Lecture Notes

草稿评论

的Normal Equation——

11

Hence,

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \frac{1}{2}(X\theta - \vec{y})^T(X\theta - \vec{y}) \\
&= \frac{1}{2}\nabla_\theta \left(\theta^T X^T X\theta - \theta^T X^T \vec{y} - \vec{y}^T X\theta + \vec{y}^T\vec{y}\right) \\
&= \frac{1}{2}\nabla_\theta \operatorname{tr}\left(\theta^T X^T X\theta - \theta^T X^T \vec{y} - \vec{y}^T X\theta + \vec{y}^T\vec{y}\right) \\
&= \frac{1}{2}\nabla_\theta \left(\operatorname{tr}\theta^T X^T X\theta - 2\operatorname{tr}\vec{y}^T X\theta\right) \\
&= \frac{1}{2}\left(X^T X\theta + X^T X\theta - 2X^T\vec{y}\right) \\
&= X^T X\theta - X^T\vec{y}
\end{aligned}$$

In the third step, we used the fact that the trace of a real number is just the real number; the fourth step used the fact that $\operatorname{tr}A = \operatorname{tr}A^T$, and the fifth step used Equation (5) with $A^T = \theta$, $B = B^T = X^T X$, and $C = I$, and Equation (1). To minimize J, we set its derivatives to zero, and obtain the **normal equations**:

$$X^T X\theta = X^T\vec{y}$$

Thus, the value of θ that minimizes $J(\theta)$ is given in closed form by the equation

$$\theta = (X^T X)^{-1}X^T\vec{y}.$$

你会发现除了 $\vec{y}$ 和 $\vec{b}$ 不一样以外，我已经把Normal Equation（ $\vec{\theta} = (A^T A)^{-1} A^T \vec{b}$ ）推出来了……我居然在下一部分还没有开始讲就把内容说完了，场面一度非常尴尬啊。可见从投影推出Normal Equation是一件多么自然的事情啊~~~我都不知道哪里切开。

说到这里先总结一下**投影的几个意义（敲黑板）！！！**

$A\vec{\theta}$ 的所有可能结果都在一个固定的区域中，在线性代数中我们称这个区域为**列空间(column space)**,列空间顾名思义就是矩阵**各列的所有线性组合** $\vec{\theta_1}\vec{a_1} + \vec{\theta_2}\vec{a_2} + \cdots + \vec{\theta_n}\vec{a_n}$。在1-D的情况下列空间就是**一条线**，在2-D的情况下列空间就是**一个平面**。但是我们的数据哪里会这么恰好的落在矩阵的列空间里呢？**天底下哪有这样的好事啊！！！**

但是目标不再在空间里并不代表不能求出解，只能说**没有perfect solution**（语出Gilbert Strang），但是我们努力一下还是可以做到**最好的(best solution)**。我们用投影向量 $\vec{p}$ 来寻找最合适的 $\vec{\theta^*}$。 $\vec{\theta^*}$ 就是并不存在的完美解 $\vec{\theta}$ 的估计值。

3.Normal Equation应用

既然Normal Equation在上文都推导完了，这里我们就随便带几个数据来玩玩咯。

> 找一条直线来拟合点 (1,1)、(2,2)、(3,2)
>
> 我们如果用一条直线来拟合的话，设 $h(\theta) = \theta_0 + \theta_1 x_1$ ，我们先得到以下值——
>
> $$\vec{\theta} = \theta_{1\ldots n}\,(n=2)$$
> $$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} A^T = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix}$$
>
> $$\overrightarrow{h(\theta)} = (1, 2, 2)^T$$
>
> 我们发现 $A\vec{\theta} = \overrightarrow{h(\theta)}$ 很遗憾的没有解，于是我们左右各乘上 A^T ，祭出了投影大招—— $A^T A\vec{\theta^*} = A^T\overrightarrow{h(\theta^*)}$。
>
> 再变换成Normal Equation: $\vec{\theta^*} = (A^T A)^{-1} A^T \overrightarrow{h(\theta^*)}$
>
> 带入数值在Matlab中小跑一下就得到了结果 $\vec{\theta^*} = \begin{pmatrix} \frac{2}{3} \\ \\ \frac{1}{2} \end{pmatrix}$
>
> 即直线 $h(x) = \dfrac{2}{3} + \dfrac{1}{2}x$ 是上述三个点的拟合结果。

如果有N个 R^2 点可以供我们使用，那么矩阵A就会变成一个n*2矩阵

$$\begin{bmatrix} \vdots & \vdots \\ 1 & 3 \end{bmatrix}$$

4.其他想说的话

在前一步可以不用判断是否可解，可以直接使用 $A^TA\vec{\theta^*} = A^T\overrightarrow{h(\theta^*)}$。因为投影的性质非常美妙，如果矩阵 A 是各行线性无关的方阵(square)，说明存在 A^{-1}，则Normal Equation会变成如下形式——

$$\begin{aligned} \vec{\theta} &= (A^TA)^{-1}A^T\vec{b} \\ &= A^{-1}A^{T^{-1}}A^T\vec{b} \\ &= A^{-1}\vec{b} \end{aligned}$$

说明如果存在一个perfect solution，该解不会受到影响。

另外一点，已经在空间中的向量乘上投影矩阵 P 仍然等于本身，也就是说 $P^2 = P$。证明如下——

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

5.参考资料

> 1.Gilbert Strang *Introduction to Linear Algebra*
>
> 2.Andrew Ng *CS229 Lecture Note 1 Supervised learning/The normal equations*