



原创 置顶 胤风 已于 2022-05-27 15:57:32 修改 阅读量10w+ 收藏 3.2k 点赞数 986

分类专栏: 数学知识 文章标签: 线性代数 矩阵 机器学习 深度学习 概率论

 华为云开发者联盟 该内容已被华为云开发者联盟社区收录

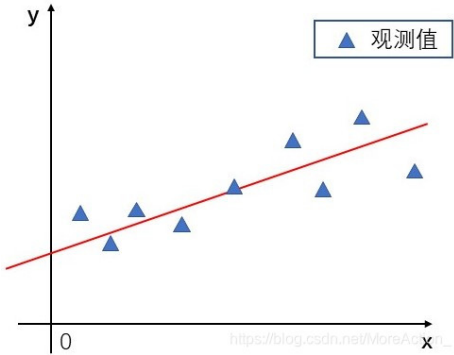
 数学知识 专栏收录该内容 188 订阅 13 篇文章

要解决的问题

在工程应用中，我们经常会用一组观测数据去估计模型的参数，模型是我们根据先验知识定下的。比如我们有一组观测数据 (x_1, y_1) （一维），通过分析我们猜测 y 和 x 之间存在线性关系，那么我们的模型就可以定为： $f(x) = kx + b$

这个模型只有两个参数，所以理论上，我们只需要观测两组数据建立两个方程，即可解出两个未知数。类似的，假如模型有 n 个参数，我们只需要观测 n 组数据即可求出参数，换句话说，在这种情况下，模型的参数是唯一确定解。

但是在实际应用中，由于我们的观测会存在误差（偶然误差、系统误差等），所以我们总会做多余观测。比如在上述例子中，尽管只有两个参数，但我们会观测 n 组数据 $(x_1, y_1) \dots (x_n, y_n)$ ，这会导致我们无法找到一条直线经过所有的点，也就是说，方程无确定解。



于是这就是我们要解决的问题：虽然没有确定解，但是我们能不能求出近似解，使得模型能在各个观测点上达到“最佳”拟合。那么“最佳”的准则是什么？所有观测点到直线的距离和最小，也可以是所有观测点到直线的误差（真实值-理论值）绝对值和最小，也可以是其它，如果你面临这个问题你会怎么做？

早在19世纪，勒让德就认为让“误差的平方和最小”估计出来的模型是最接近真实情形的。

为什么就是误差平方而不是其它的，这个问题连欧拉、拉普拉斯都未能成功回答，后来是高斯建立了一套误差分析理论，从而证明了确实是使误差平方和最小的情况下系统是最优的。理论的证明也并不难，我写在了另外一篇博客 [最小二乘法的原理解](#)，相信你了解后会对最小二乘法有更深刻的认识。

按照勒让德的最佳原则，于是就是求：

$$L = \sum_{i=1}^n (y_i - f(x))^2$$

这个目标函数取得最小值时的函数参数，这就是最小二乘法的思想，所谓“二乘”就是平方的意思。从这里我们可以看到，**最小二乘法其实就是要来做最小二乘法的一种思想。**

至于怎么求出具体的参数那就是另外一个问题了，理论上可以用导数法、几何法，工程上可以用梯度下降法。下面以最常用的线性回归为例进行推导：

线性回归

线性回归因为比较简单，可以直接推导出解析解，而且许多非线性的问题也可以转化为线性问题来解决，所以得到了广泛的应用。甚至许多人认为最好的就是线性回归，其实并不是，最小二乘法就是一种思想，它可以拟合任意函数，线性回归只是其中一个比较简单而且也很常用的函数，所以讲最小二乘法都会以它为例。

下面我会先用矩阵法进行推导，然后再用几何法来

型中得：

$$\begin{aligned}h_1 &= \theta_0 + \theta_1 x_{1,1} + \theta_2 x_{1,2} + \dots + \theta_{n-1} x_{1,n-1} \\h_2 &= \theta_0 + \theta_1 x_{2,1} + \theta_2 x_{2,2} + \dots + \theta_{n-1} x_{2,n-1} \\&\vdots \\h_m &= \theta_0 + \theta_1 x_{m,1} + \theta_2 x_{m,2} + \dots + \theta_{n-1} x_{m,n-1}\end{aligned}$$

为方便用矩阵表示，我们令 $x_0 = 1$ ，于是上述方程可以用矩阵表示为：

$$\mathbf{h} = \mathbf{X}\theta$$

其中， \mathbf{h} 为 $m \times 1$ 的向量，代表模型的理论值， θ 为 $n \times 1$ 的向量， \mathbf{X} 为 $m \times n$ 维的矩阵， m 代表样本的个数， n 代表样本的特征数，于是目标损失函数用矩阵表示

$$J(\theta) = \|\mathbf{h} - \mathbf{Y}\|^2 = \|\mathbf{X}\theta - \mathbf{Y}\|^2 = (\mathbf{X}\theta - \mathbf{Y})^T (\mathbf{X}\theta - \mathbf{Y})$$

其中 \mathbf{Y} 是样本的输出向量，维度为 $m \times 1$ 。

根据高数知识我们知道函数取得极值就是导数为0的地方，所以我们只需要对损失函数求导令其等于0就可以解出 θ 。矩阵求导属于矩阵微积分的内容，学的(....，这里先介绍两个用到的公式：

$$\begin{aligned}\frac{\partial \mathbf{x}^T \mathbf{a}}{\partial \mathbf{x}} &= \frac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a} \\ \frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} &= \mathbf{A} \mathbf{x} + \mathbf{A}^T \mathbf{x}\end{aligned}$$

如果矩阵 \mathbf{A} 是对称的：

$$\mathbf{A} \mathbf{x} + \mathbf{A}^T \mathbf{x} = 2\mathbf{A} \mathbf{x}$$

对目标函数化简：

$$J(\theta) = \theta^T \mathbf{X}^T \mathbf{X} \theta - \theta^T \mathbf{X}^T \mathbf{Y} - \mathbf{Y}^T \mathbf{X} \theta + \mathbf{Y}^T \mathbf{Y}$$

求导令其等于0：

$$\frac{\partial}{\partial \theta} J(\theta) = 2\mathbf{X}^T \mathbf{X} \theta - 2\mathbf{X}^T \mathbf{Y} = 0$$

解得 $\theta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ ，经过推导我们得到了 θ 的解析解，现在只要给了数据，我们就可以带入解析解中直接算出 θ 。

几何意义

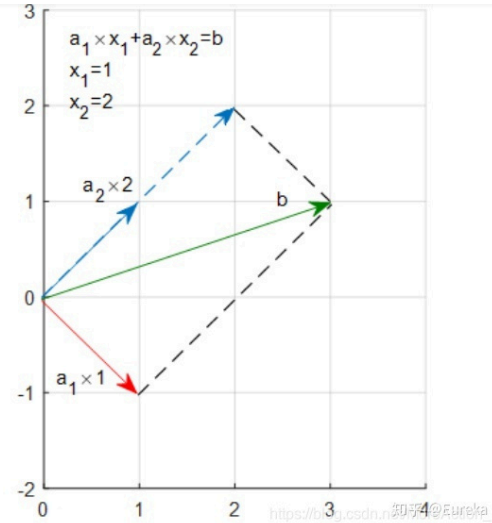
几何意义会直观的帮助你理解最小二乘法究竟在干什么。首先先来解释一下矩阵乘法的几何意义，对于一个方程组 $\mathbf{A} \mathbf{x}$ ，我们可以看做是 \mathbf{x} 对矩阵 \mathbf{A} 的线性组合，比如：

$$\begin{cases} 1 \times x_1 + x_2 = 3 \\ -1 \times x_1 + x_2 = 1 \end{cases} \Leftrightarrow \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \Leftrightarrow \mathbf{A} \times \mathbf{x} = \mathbf{b}$$

可以看作：

$$\begin{bmatrix} 1 \\ -1 \end{bmatrix} \times x_1 + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \times x_2 = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \Leftrightarrow \mathbf{a}_1 \times x_1 + \mathbf{a}_2 \times x_2 = \mathbf{b}$$

画在坐标轴上可以看到，向量 \mathbf{b} 其实就是向量 \mathbf{a}_1 与 \mathbf{a}_2 的线性组合，因为他们都在一个平面上，显然是有解的。



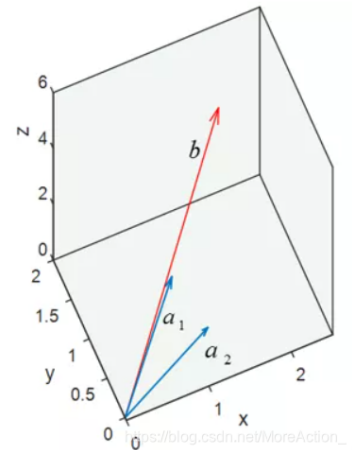
但是如文章开头所说，由于存在观测误差，我们往往会做多余观测，比如要拟合一次方程 $y = kx + b$ ，我们可能观测了三个点 (0,2)，(1,2)，(2,2)，矩阵形式如下(为表述方便，用 x_1 代替 k ， x_2 代替 b)：

$$\begin{cases} 1 \times x_1 + x_2 = 2 \\ 0 \times x_1 + x_2 = 2 \\ 2 \times x_1 + x_2 = 3 \end{cases} \Leftrightarrow \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix} \Leftrightarrow \mathbf{A} \times \mathbf{x} = \mathbf{b}$$

表示成线性组合的方式：

$$\begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} \times x_1 + \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \times x_2 = \begin{bmatrix} 2 \\ 2 \\ 3 \end{bmatrix} \Leftrightarrow \mathbf{a}_1 \times x_1 + \mathbf{a}_2 \times x_2 = \mathbf{b}$$

画在图中如下：



从图中我们可以看到，无论 \mathbf{a}_1 和 \mathbf{a}_2 怎么线性组

