线性回归

最小二乘法

赵海涛

haitaozhao@ecust.edu.cn

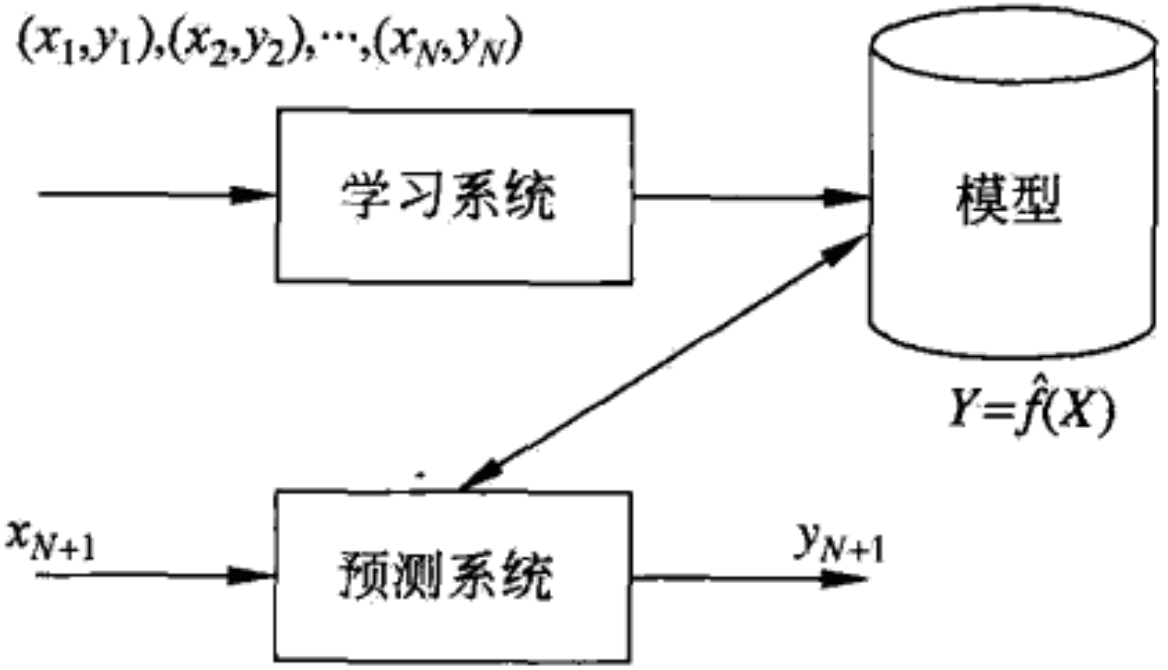
大纲

- 回归模型
- 回归的目标函数（最小二乘准则）
- 线性回归的求解
- 线性回归的扩展

回归问题

- 回归模型是表示从输入变量到输出变量之间映射的函数。回归问题的学习等价于函数拟合。
- 学习和预测两个阶段
- 训练集:

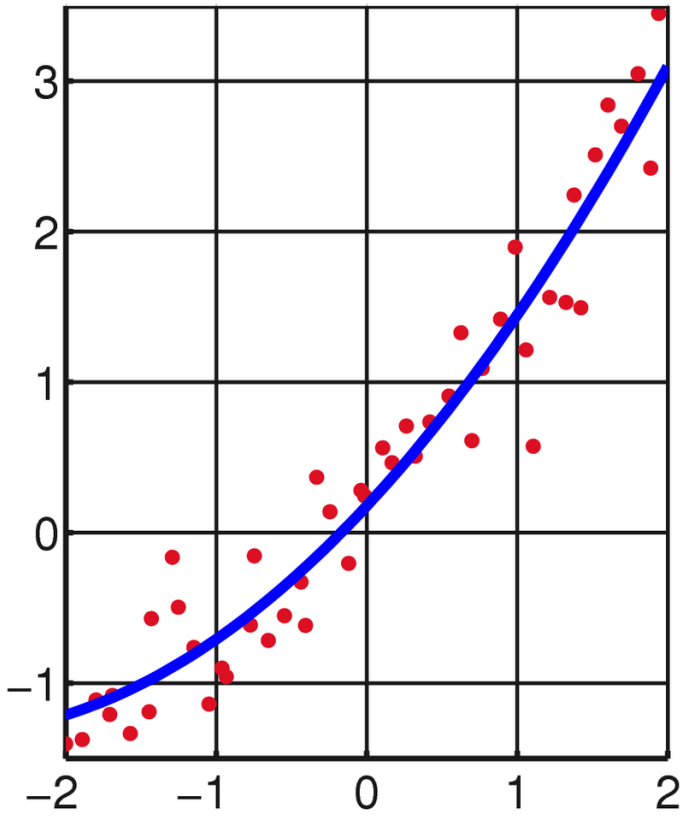
$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$$

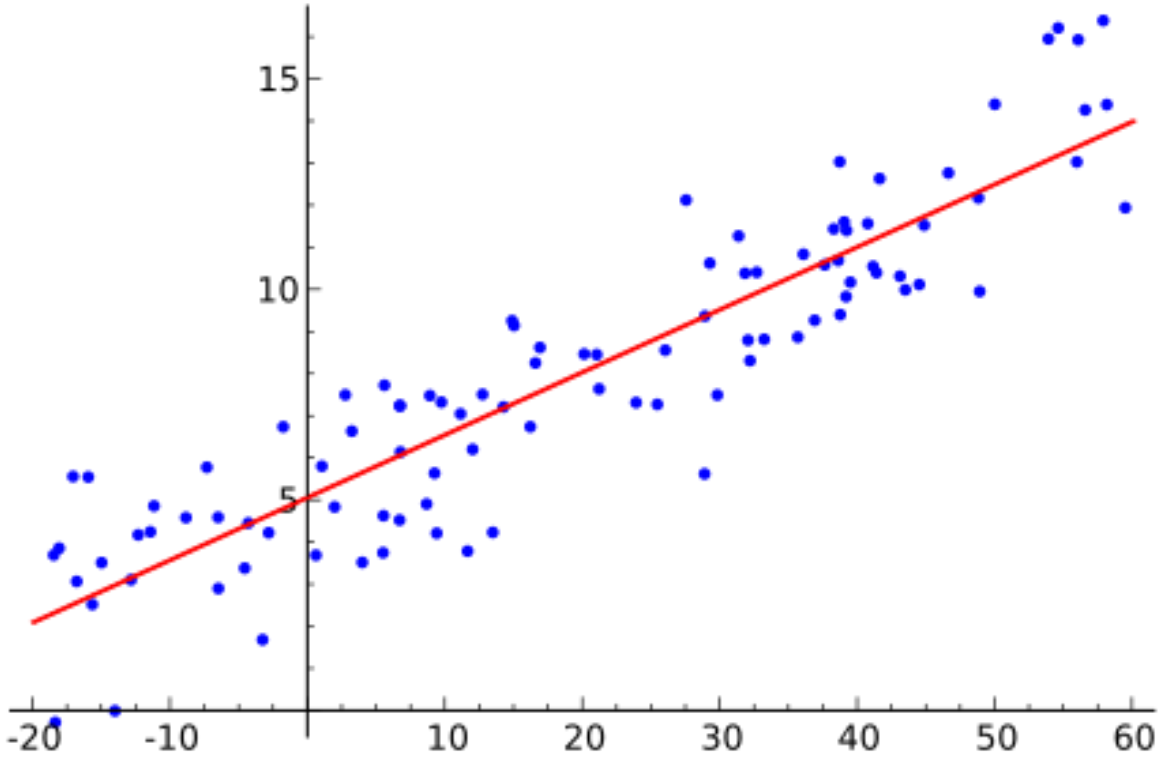


回归问题

- 回归学习最常用的损失函数是平方损失函数，在此情况下，回归问题可以由 著名的最小二乘法(least squares)求解。

- 股价预测
- 房价预测





回归问题

- 假设我们有一个数据集包含n个数据对$(\boldsymbol{x}_i, y_i), i = 1,2,\ldots,n$，其中$\boldsymbol{x}_i$是自变量，$\boldsymbol{x}_i \in R^{1\times d}$，$y_i$是因变量，$y_i \in R$。$\boldsymbol{x}_i$可以认为是我们在生产过程中获得的一些传感器和仪表的数据构成的向量，比如$\boldsymbol{x}_i$的各个维度分别对应压力，温度，浓度等等，i表示采样的时刻，y_i是对应$\boldsymbol{x}_i$的某个指标，比如产品的质量指标等，它是连续的一个数值。
- 我们希望有一个模型$f(\boldsymbol{x}_i, \boldsymbol{\beta})$可以预测未来的$y_i$，即$y_i = f(\boldsymbol{x}_i, \boldsymbol{\beta})$（其中$\boldsymbol{x}_i \in R^{1\times d}$，$\boldsymbol{\beta} \in R^{d\times 1}$）。简单地说，这个模型应该“非常好”地适合于已有的数据，我们令误差$\varepsilon_i = y_i - f(\boldsymbol{x}_i, \boldsymbol{\beta})$。

回归问题的目标函数

- 回归的目标就是要求取$\boldsymbol{\beta}$来极小化ε_i的平方和

$$J(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} {\varepsilon_i}^2 = \frac{1}{n}\sum_{i=1}^{n}[y_i - f(\boldsymbol{x_i}, \boldsymbol{\beta})]^2$$

上式即为回归问题中的目标函数。

- 通常，将$f(\boldsymbol{x}_i, \boldsymbol{\beta})$进行简化，最简单的，$f(\boldsymbol{x}_i, \boldsymbol{\beta})$可以看作是一个超平面，即$f(\boldsymbol{x}_i, \boldsymbol{\beta}) = \boldsymbol{x}_i\boldsymbol{\beta} + \beta_0$，其中$\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^T$。则有方程

$$\begin{cases} y_1 = \beta_0 + \boldsymbol{x}_1\boldsymbol{\beta} + \varepsilon_1 \\ y_2 = \beta_0 + \boldsymbol{x}_2\boldsymbol{\beta} + \varepsilon_2 \\ \vdots \\ y_n = \beta_0 + \boldsymbol{x}_n\boldsymbol{\beta} + \varepsilon_n \end{cases}$$

线性回归

- 令$Y=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix}$，$X=\begin{bmatrix}1&x_{11}&x_{12}&\cdots&x_{1d}\\1&x_{21}&x_{22}&\ddots&x_{2d}\\\vdots&\vdots&\vdots&\ddots&\vdots\\1&x_{n1}&x_{n2}&\cdots&x_{nd}\end{bmatrix}$，$E=\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{bmatrix}$，$B=\begin{bmatrix}\beta_0\\\boldsymbol{\beta}\end{bmatrix}$,则有

$$Y=XB+E$$

$$J(B)=\frac{1}{n}\sum_{i=1}^{n}\left[y_i-(\sum x_{ij}\beta_j+\beta_0)\right]^2=\frac{1}{n}(Y-XB)^T(Y-XB)$$

- β_0的作用主要是产生一个偏置，如$\sum \boldsymbol{x}_i=\mathbf{0}$，而$\sum y_{\mathrm{i}}\neq 0$时，那么在$y$轴上就会产生一个截距，这是线性回归问题优化时要考虑的。

线性回归问题的求解

- 求解 $J(B)$ 的最小值，可以直接求解，也可以用求导数的方式来求解，即用梯度下降法来求解。

- 梯度下降法和牛顿法求解

- 通过正交投影直接求解

梯度下降法

一元情况下的Taylor展开式：

$$f(x+\delta)=f(x)+f'(x)\delta+\frac{1}{2}f''(x)\delta^2+\ldots$$

推广到多元情况下的Taylor展开式，即$\boldsymbol{x}\in R^{d\times 1}$,$f(\boldsymbol{x})\in R$时：

$$f(\boldsymbol{x}+\Delta\boldsymbol{x})=f(\boldsymbol{x})+\nabla^T f(\boldsymbol{x})\Delta\boldsymbol{x}+\frac{1}{2}\Delta\boldsymbol{x}^T\nabla^2 f(\boldsymbol{x})\Delta\boldsymbol{x}+\cdots$$

其中，$\nabla f(x)=\begin{bmatrix}\frac{\partial f}{\partial x_1}\\ \vdots \\ \frac{\partial f}{\partial x_d}\end{bmatrix}$，$\nabla^2 f(x)=\begin{bmatrix}\frac{\partial^2 f}{\partial x_1\partial x_1} & \frac{\partial^2 f}{\partial x_2\partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d\partial x_1}\\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1\partial x_d} & \cdots & \cdots & \frac{\partial^2 f}{\partial x_d\partial x_d}\end{bmatrix}$.

梯度下降法

- 梯度下降法的目的就是求取$f(\boldsymbol{x})$的极值。我们知道当$\nabla f(\boldsymbol{x}) = \mathbf{0}$，且$\nabla^2 f(\boldsymbol{x}) \succ 0$时(这里表示$\nabla^2 f(\boldsymbol{x})$应为正定矩阵，正定矩阵的充要条件为它的所有特征值都大于零或者各阶主子式大于零),存在唯一极值点。

- 对于一阶Taylor展开式，当$\Delta\boldsymbol{x}$足够小的时候有以下的近似式：

$$f(\boldsymbol{x} + \Delta\boldsymbol{x}) \approx f(\boldsymbol{x}) + \nabla^T f(\boldsymbol{x})\Delta\boldsymbol{x}$$

从而

$$f(\boldsymbol{x} + \Delta\boldsymbol{x}) - f(\boldsymbol{x}) \approx \nabla^T f(\boldsymbol{x})\Delta\boldsymbol{x}$$

梯度下降法

如果已经有一个$\boldsymbol{x}_t$，希望$\boldsymbol{x}_t + \Delta\boldsymbol{x}_t$使$f(\boldsymbol{x}_t + \Delta\boldsymbol{x}_t)$小于$f(\boldsymbol{x}_t)$，那么可以令

$$\Delta\boldsymbol{x}_t = -\nabla f(\boldsymbol{x}_t)。$$

此时就有

$$f(\boldsymbol{x}_t + \Delta\boldsymbol{x}_t) - f(\boldsymbol{x}_t) = -\nabla^T f(\boldsymbol{x}_t)\nabla f(\boldsymbol{x}_t) = -\sum_{i=1}^{d}\left(\frac{\partial f(\boldsymbol{x}_t)}{\partial x_i}\right)^2 < 0$$

在Taylor展开中，通常需要对$\Delta\boldsymbol{x}_t$有一定的限制，在梯度下降法中一般取

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \alpha\Delta\boldsymbol{x}_t,$$

其中α为学习率或步长。

梯度下降法的步骤

下面给出进行梯度下降法求解的一般步骤：

①给定初值x_1，$i=1$，学习率α，给定阈值ε；

②求$\nabla f\left(x_i\right)$，令$\Delta \boldsymbol{x}_i=-\nabla f\left(\boldsymbol{x}_i\right)$；

③求$\boldsymbol{x}_{i+1}=\boldsymbol{x}_i+\alpha \Delta \boldsymbol{x}_i$；

④若$\| f\left(\boldsymbol{x}_{i+1}\right)-f\left(\boldsymbol{x}_i\right) \|_2^2 \leq \varepsilon$，转向⑤，否则，$i=i+1$，转向②

⑤输出$\boldsymbol{x}_{i+1}$.

牛顿法

- 如果考虑二阶Taylor展开，则：

$$f(\boldsymbol{x}+\Delta\boldsymbol{x}) \approx f(\boldsymbol{x}) + \nabla^T f(\boldsymbol{x})\Delta\boldsymbol{x} + \frac{1}{2}\Delta\boldsymbol{x}^T \nabla^2 f(\boldsymbol{x})\Delta\boldsymbol{x}$$

注：有很多书和论文用$\boldsymbol{g} = \nabla f(\boldsymbol{x})$，$H = \nabla^2 f(\boldsymbol{x})$ 来代替书写。

- 要求解$f(\boldsymbol{x}+\Delta\boldsymbol{x})$的最小值，可以看作求$\Delta\boldsymbol{x}$使得

$$f(\boldsymbol{x}) + \nabla^T f(\boldsymbol{x})\Delta\boldsymbol{x} + \frac{1}{2}\Delta\boldsymbol{x}^T \nabla^2 f(\boldsymbol{x})\Delta\boldsymbol{x}$$

最小化。对$\Delta\boldsymbol{x}$求导可得

$$\Delta\boldsymbol{x} = -\left(\nabla^2 f(\boldsymbol{x})\right)^{-1} \nabla f(\boldsymbol{x}) = -H^{-1} g$$

梯度下降法的收敛性

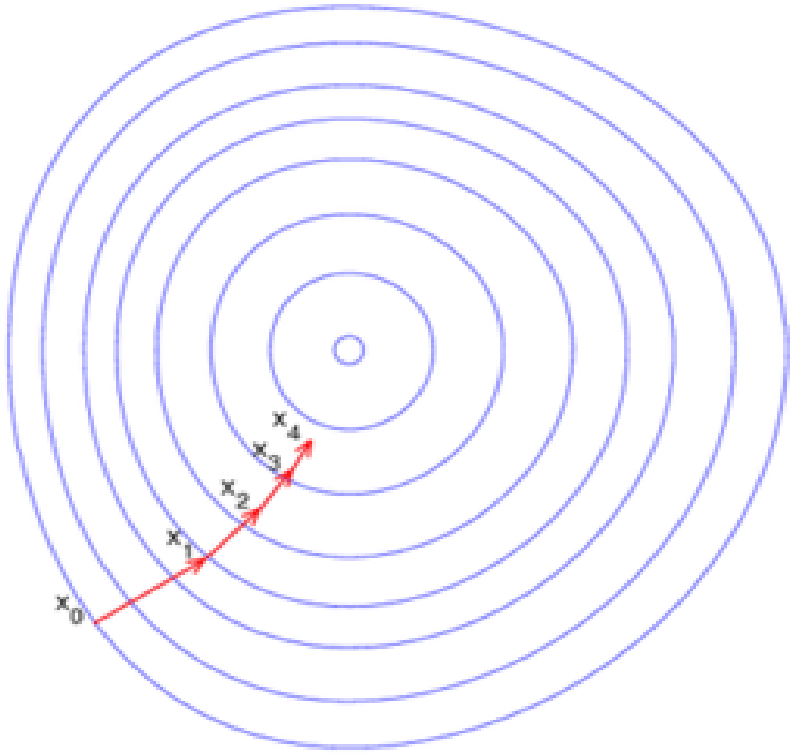
- 梯度下降法收敛速度慢与关于x_k的Hessian矩阵的“条件”有关。梯度下降法第k次的迭代过程与之前迭代过程之间的关系可以通过下式来表示：

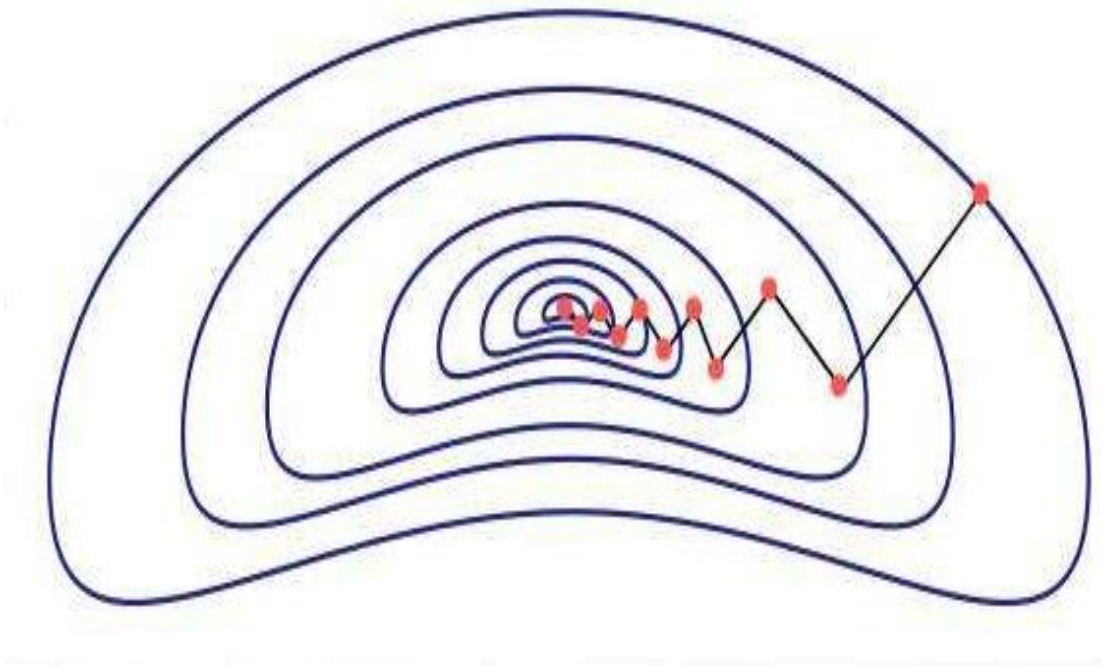
$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*) \leq (\frac{1-r}{1+r})^2 (f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*))$$

其中，r是关于$\boldsymbol{x}_k$的Hessian矩阵（用H_k表示）最小特征值与最大特征值的比值。从上式我们可以看出当r较小或者H_k的条件数较大的时候，梯度下降法的收敛速度就可能会较慢。经典的梯度下降法是线性收敛的，而牛顿法是二次收敛的。

- 另外，梯度下降法容易出现锯齿效应。

梯度下降法的收敛性





梯度下降法的收敛性

- 梯度下降法收敛速度慢与关于x_k的Hessian矩阵的“条件”有关。梯度下降法第k次的迭代过程与之前迭代过程之间的关系可以通过下式来表示：

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}^*) \leq (\frac{1-r}{1+r})^2 (f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*))$$

其中，r是关于$\boldsymbol{x}_k$的Hessian矩阵（用H_k表示）最小特征值与最大特征值的比值。从上式我们可以看出当r较小或者H_k的条件数较大的时候，梯度下降法的收敛速度就可能会较慢。经典的梯度下降法是线性收敛的，而牛顿法是二次收敛的。

- 另外，梯度下降法容易出现锯齿效应。

梯度下降法求解线性回归

- $$J(B) = \frac{1}{n}(Y - XB)^T(Y - XB)$$
$$= \frac{1}{n}(Y^TY - Y^TXB - B^TX^TY + B^TX^TXB)$$
$$= \frac{1}{n}(Y^TY - 2B^TX^TY + B^TX^TXB)$$
- $$\nabla J(B) = \frac{2}{n}(-X^TY + X^TXB)$$

线性回归问题的一般步骤：

①给定初值B_1，$i = 1$，学习率α，给定阈值ε；；

②求$B_{i+1} = B_i - \alpha\nabla J(B_i)$ ；

③若$\| J(B_{i+1}) - J(B_i) \|_2^2 \leq \varepsilon$，转向④，否则，$i = i + 1$，转向②

④输出B_{i+1}.

梯度下降法求解线性回归

- 一种常见的方法是针对不同维度进行自适应改进学习率，对不同维度采用不同的 α，比较常用的是RMSProp(root mean square propagation)。首先对每个维度进行移动平均$\boldsymbol{\nu}_t = \mu\boldsymbol{\nu}_{t-1} + (1-\mu)[\nabla J(B_t)]^2$ （平方是每个元素的平方）。μ一般选为0.9,0.99或0.999。参数的更新变为

$$B_{t+1} = B_t - \frac{\alpha \nabla J(B_t)}{\sqrt{\boldsymbol{\nu}_t + \varepsilon}}$$

（开根号是对每个元素进行，分子除以分母也是对每个元素进行）

这里可发现和经典的梯度下降法相比，多了一个类似权重的项用于不同的维度上。

梯度下降法求解线性回归

- 另一类较常见的是Adam(adaptive moment estimation)

$$\boldsymbol{m}_t = \mu_0 \boldsymbol{m}_{t-1} + (1-\mu_0)\nabla J(B_t)$$

$$\boldsymbol{\nu}_t = \mu_1 \boldsymbol{\nu}_{t-1} + (1-\mu_1)[\nabla J(B_t)]^2$$

$$B_{t+1} = B_t - \frac{\alpha \boldsymbol{m}_t}{\sqrt{\boldsymbol{\nu}_t} + \varepsilon}$$

实际就是在梯度方向的选择上，也运用移动平均，既考虑当前的梯度方向，又考虑上一步的梯度方向，通过加权平均可减小锯齿效应。

注：其它改进的方法还有很多，主要也是针对方向和学习率。常用的这两个方法效果已经相当好。

梯度下降法求解线性回归

- 考虑牛顿法，即$\Delta B = -\left(\nabla^2 J(B)\right)^{-1}\nabla J(B) = -(X^TX)^{-1}(-X^TY + X^TXB)$

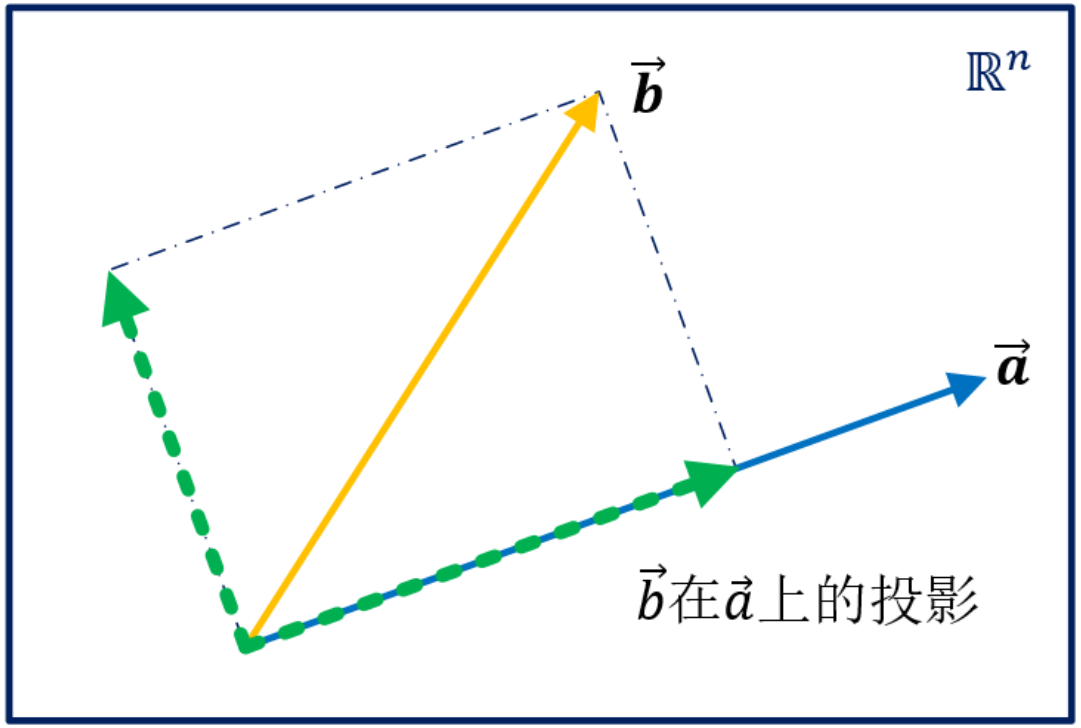
则，

$$B^* = B + \Delta B = B - B + (X^TX)^{-1}X^TY = (X^TX)^{-1}X^TY$$

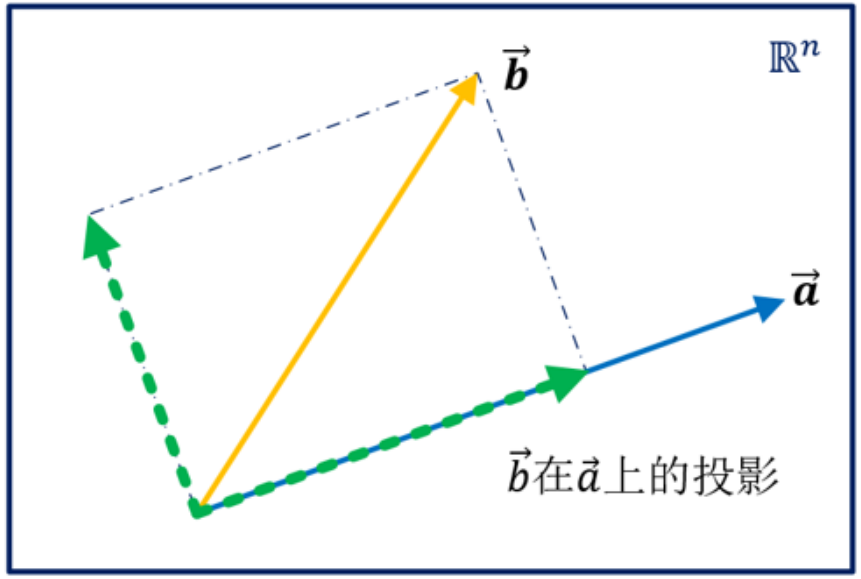
可以发现B^*即为最优解，这是因为目标函数为B的二次函数，Taylor展开到二阶即为精确展开。一次牛顿法迭代就可得到最优解。

线性回归最优解的几何解释

考虑在n维空间$\mathbb{R}^n$中的向量$\vec{a} \in \mathbb{R}^n$，对于任意给定的向量$\vec{b} \in \mathbb{R}^n$，则向量$\vec{b}$在向量$\vec{a}$方向上的长度矢量，称为$\vec{b}$在$\vec{a}$方向上的投影。



线性回归最优解的几何解释



由与$\vec{\boldsymbol{b}} - \mu\vec{\boldsymbol{a}}$部分与向量$\mu\vec{\boldsymbol{a}}$是正交的，则有

$$\mu\vec{\boldsymbol{a}}^T\left(\vec{\boldsymbol{b}} - \mu\vec{\boldsymbol{a}}\right) = 0$$

进而有

$$\vec{\boldsymbol{a}}^T\left(\vec{\boldsymbol{b}} - \mu\vec{\boldsymbol{a}}\right) = 0$$

$$\vec{\boldsymbol{a}}^T\vec{\boldsymbol{b}} - \vec{\boldsymbol{a}}^T\mu\vec{\boldsymbol{a}} = 0$$

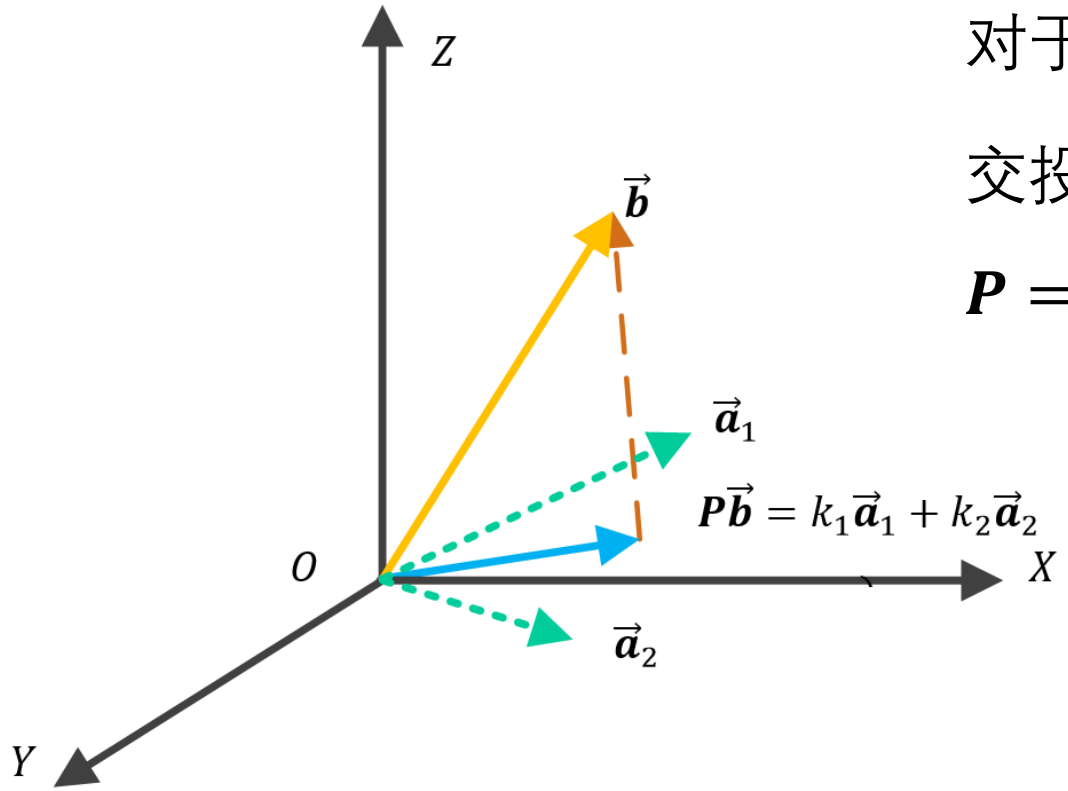
$$\vec{\boldsymbol{a}}^T\vec{\boldsymbol{b}} = \vec{\boldsymbol{a}}^T\mu\vec{\boldsymbol{a}} = \mu\vec{\boldsymbol{a}}^T\vec{\boldsymbol{a}}$$

$$\mu = (\vec{\boldsymbol{a}}^T\vec{\boldsymbol{a}})^{-1}\vec{\boldsymbol{a}}^T\vec{\boldsymbol{b}}$$

因此，从$\boldsymbol{P}\vec{\boldsymbol{b}} = \mu\vec{\boldsymbol{a}}$可得

$$\boldsymbol{P} = \boldsymbol{a}(\boldsymbol{a}^T\boldsymbol{a})^{-1}\boldsymbol{a}^T$$

线性回归最优解的几何解释



对于任意的向量$\vec{b}$，若要求出其向空间A的正交投影结果，则就是对其做变换$P\vec{b}$，其中$P = A(A^TA)^{-1}A^T$。

线性回归最优解

- 最小二乘解就是在寻找本体位置和投影位置之差的二范数最小的解。最小二乘问题的目标函数为

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2$$

- 通过正交投影的知识可以知道，对正交投影$\boldsymbol{P} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$，$\boldsymbol{y}$在$\boldsymbol{A}$的列空间中的投影为

$$\boldsymbol{P}\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}$$

- 可得

$$\hat{\boldsymbol{x}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}$$

- 如果采用伪逆的写法则记为

$$\hat{\boldsymbol{x}} = \boldsymbol{A}^+\boldsymbol{y}$$

这就是从正交投影的理解方式，导出的最小二乘解。

线性回归问题的扩展

- 加权最小二乘法
- 偏最小二乘法
- 正则项（惩罚项）
- 非线性扩展

谢谢各位同学！