机器学习

最小二乘学习法

以d维实向量$\boldsymbol{x}$作为输入、以实数值y作为输出的函数$y = f(\boldsymbol{x})$的学习问题进行说明。

这里的真实关系f是未知的，通过学习过程中作为训练集而输入输出的训练样本$\{\boldsymbol{x}_i, y_i\}_{i=1}^n$来对其进行学习。但是 在一般情况下，输出样本$y_i$的直实值$f(\boldsymbol{x}_i)$中经常会观测到噪声。

均方误差损失

最小二乘法习法是 对模型 的输出$f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$和训练集输出$\{y_i\}_{i=1}^n$的平方误差

$$J_{\mathrm{LS}}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^{n} (f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i)^2$$

为最小时的参数$\boldsymbol{\theta}$进行学习。

$$\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = \arg\min_{\boldsymbol{\theta}} J_{\mathrm{LS}}(\boldsymbol{\theta})$$

平方误差$(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i)^2$是残差

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i$$

的ℓ_2范数，因此最小二乘学习法有时也称为ℓ_2损失最小化学习法。

如果使用线性模型

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{b} \theta_i \phi_i(\boldsymbol{x}) = \boldsymbol{\theta}^T \phi(\boldsymbol{x})$$

则训练样本的平方差J_{LS}就能够表示为下述形式。

$$J_{\mathrm{LS}}(\boldsymbol{\theta}) = \frac{1}{2} \|\boldsymbol{\Phi}\boldsymbol{\theta} - \boldsymbol{y}\|^2$$

在这里，$\boldsymbol{y} = (y_1, \cdots, y_n)^T$是训练输出的$n$维向量,$\boldsymbol{\Phi}$是$n \times b$矩阵，也称为设计矩阵，如下所示：

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_1(\boldsymbol{x}_1) & \cdots & \phi_b(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\boldsymbol{x}_n) & \cdots & \phi_b(\boldsymbol{x}_n) \end{pmatrix}$$

求解

训练样本的平方差J_{LS}的参数向量$\boldsymbol{\theta}$的偏微分$\nabla_{\boldsymbol{\theta}} J_{\mathrm{LS}}$为

$$\nabla_{\boldsymbol{\theta}} J_{\mathrm{LS}} = \left(\frac{\partial J_{\mathrm{LS}}}{\partial \theta_1}, \cdots, \frac{\partial J_{\mathrm{LS}}}{\partial \theta_b} \right)^T = \boldsymbol{\Phi}^T \boldsymbol{\Phi} \boldsymbol{\theta} - \boldsymbol{\Phi}^T \boldsymbol{y}$$

令此微分为0,得最小二乘解满足

$$\boldsymbol{\Phi}^T \boldsymbol{\Phi} \boldsymbol{\theta} = \boldsymbol{\Phi}^T \boldsymbol{y}$$

此方程的解$\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$使用设计矩阵$\boldsymbol{\Phi}$的广义逆矩阵$\boldsymbol{\Phi}^{\dagger}$来进行计算，得

$$\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = \boldsymbol{\Phi}^{\dagger} \boldsymbol{y}$$

加权最小二乘

对顺序为i的训练样本的平方差通过权重$w_i \geqslant 0$进行加权，然后再采用最小二乘法学习，称为加权最小二乘学习法。

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \sum_{i=1}^{n} w_i (f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i)^2$$

加权最小二乘法，与没有权重时相同，

$$(\boldsymbol{\Phi}^T \boldsymbol{W} \boldsymbol{\Phi})^{\dagger} \boldsymbol{\Phi}^T \boldsymbol{W} \boldsymbol{y}$$

上式中的$\boldsymbol{W}$是以$w_1, \cdots, w_n$为对角元素的对角矩阵。

核模型

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{n} \theta_j K(\boldsymbol{x}, \boldsymbol{x}_j)$$

也可以认为是线性模型的一种， 把设计矩阵$\boldsymbol{\Phi}$置换为核矩阵K， 就可以使用和线性模型相同的方法来求得核模型的最小二乘解

$$K = \begin{pmatrix} K(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & K(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ K(\boldsymbol{x}_n, \boldsymbol{x}_1) & \cdots & K(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{pmatrix}$$

最小二乘解的性质

设计矩阵$\boldsymbol{\Phi}$的奇异值分解

$$\boldsymbol{\Phi} = \sum_{k=1}^{\min(n,b)} \kappa_k \boldsymbol{\psi}_k \boldsymbol{\varphi}_k^T$$

其中$\kappa_k, \boldsymbol{\psi}_k, \boldsymbol{\varphi}_k$分别称为奇异值，左奇异向量，右奇异向量。奇异值全部是非负的，奇异向量满足正交性

$$\boldsymbol{\psi}_i^T \boldsymbol{\psi}_{i'} = \begin{cases} 1 & (i = i') \\ 0 & (i \neq i') \end{cases} \qquad \boldsymbol{\varphi}_j^T \boldsymbol{\varphi}_{j'} = \begin{cases} 1 & (j = j') \\ 0 & (j \neq j') \end{cases}$$

进行奇异值分解后，$\boldsymbol{\Phi}$的广义逆矩阵$\boldsymbol{\Phi}^{\dagger}$就可以表示为

$$\boldsymbol{\Phi}^{\dagger} = \sum_{k=1}^{\min(n,b)} \kappa_k^{\dagger} \boldsymbol{\varphi}_k \boldsymbol{\psi}_k^T$$

其中$\kappa^{\dagger}$是标量κ的广义逆矩阵

$$\kappa^{\dagger} = \begin{cases} \frac{1}{\kappa} & (\kappa \neq 0) \\ 0 & (\kappa = 0) \end{cases}$$

最小二乘解$\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$可以表示为

$$\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = \sum_{k=1}^{\min(n,b)} \kappa_k^{\dagger} (\boldsymbol{\psi}_k^T \boldsymbol{y}) \boldsymbol{\varphi}_k$$

无偏性

将最小二乘学习法中得到的函数以训练输入$\{\boldsymbol{x}_i\}_{i=1}^n$得到输出值$\{f_{\hat{\boldsymbol{\theta}}_{\mathrm{LS}}(\boldsymbol{x}_i)}\}_{i=1}^n$，变换为列向量表示，得

$$(f_{\hat{\boldsymbol{\theta}}_{\mathrm{LS}}}(\boldsymbol{x}_1), \cdots f_{\hat{\boldsymbol{\theta}}_{\mathrm{LS}}}(\boldsymbol{x}_n))^T = \boldsymbol{\Phi}\hat{\boldsymbol{\theta}}_{\mathrm{LS}} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\dagger}\boldsymbol{y}$$

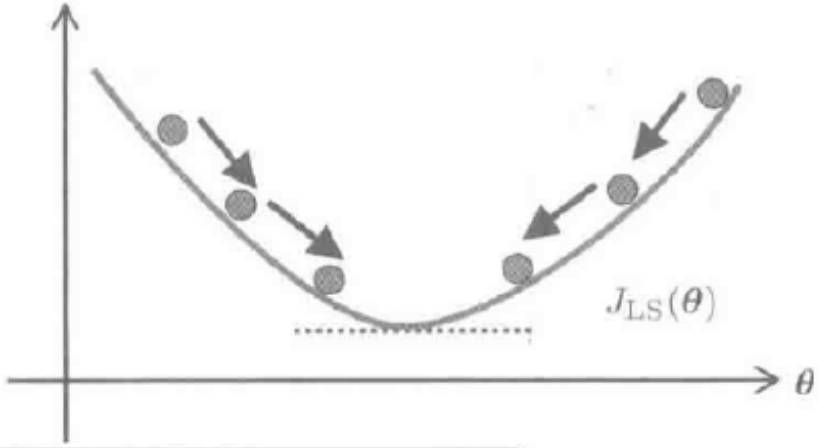
其中$\boldsymbol{\Phi}\boldsymbol{\Phi}^{\dagger}$是$\boldsymbol{\Phi}$的值域$\mathcal{R}(\boldsymbol{\Phi})$的正交投影矩阵， 最小二乘学习法的输出向量是由$\mathcal{R}(\boldsymbol{\Phi})$的正交投影得到的。

如果噪声的期望值为0,则最小二乘解$\hat{\boldsymbol{\theta}}_{\mathrm{LS}}$就是真实参数$\boldsymbol{\theta}^*$的无偏估计量。

$$\mathbb{E}[\hat{\boldsymbol{\theta}}_{\mathrm{LS}}] = \boldsymbol{\theta}^*$$

上式中$\mathbb{E}$为对噪声的期望值。 另外， 即使真实函数没有包含在模型中（即无论对于什么样的$\boldsymbol{\theta}$，都存在$f \neq f_{\boldsymbol{\theta}}$，如果增加训练样本数$n$，$\mathbb{E}[\hat{\boldsymbol{\theta}}_{\mathrm{LS}}]$也会向着模型中的最优参数方向收敛。这种性质称为无偏性。

随机梯度法



当梯度 $\nabla j_{\mathrm{LS}}^{(i)}$ 为负时，根据式(3.3)，参数会朝着正的方向(图中右方向)加以修正；当梯度 $\nabla j_{\mathrm{LS}}^{(i)}$ 为正时，根据式(3.3)，参数会朝着负的方向(图中左方向)加以修正。

图3.4 梯度法

随机梯度法是指，沿着训练平方误差J_{LS}的梯度下降，对参数依次进行学习的算法。

一般而言，与线性模型相对应的训练平方误差J_{LS}为凸函数。$J(\boldsymbol{\theta})$函数为凸函数是指对任意的两点$\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$和任意的$t \in [0,1]$，

$$J(t\boldsymbol{\theta}_1 + (1-t)\boldsymbol{\theta}_2) \leqslant \mathrm{tJ}(\boldsymbol{\theta}_1) + (1-t)J(\boldsymbol{\theta}_2)$$

凸函数只有一个峰值，可以通过梯度法得到全局最优解。

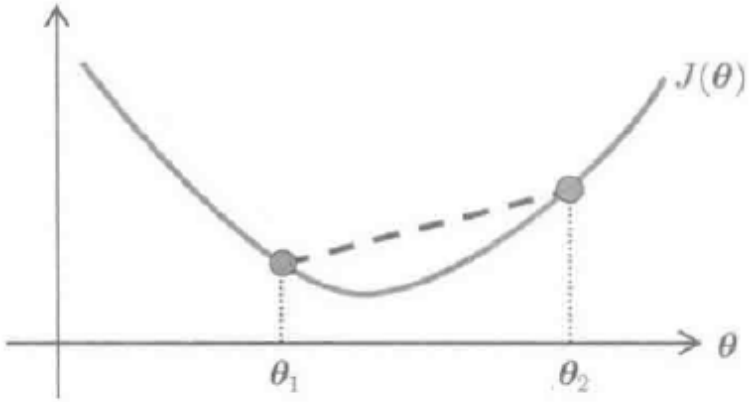

$J(\theta)$
θ
θ_1
θ_2

图3.5 凹函数


连接任意两点θ_1、θ_2的线段一定在函数的上部。

随机梯度最小二乘

随机梯度算法对线性模型进行最小二乘学习的算法流程

1. 给$\boldsymbol{\theta}$以适当的初值
2. 随机先择一个训练样本,如：$(\boldsymbol{x}_i, y_i)$
3. 对于选定的训练样本，采用使其梯度下降的方式，对参数$\boldsymbol{\theta}$进行更新

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \varepsilon \nabla J_{\mathrm{LS}}^{(i)}(\boldsymbol{\theta})$$

这里ε是名为学习系数的小标量， 表示梯度下降的步幅， $\nabla J_{\mathrm{LS}}^{(i)}$是顺度为$i$的训练样本相对应的训练平方误差的梯度

$$\nabla J_{\mathrm{LS}}^{(i)}(\boldsymbol{\theta}) = \phi(\boldsymbol{x}_i)(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i) - y_i)$$

4. 直到解$\boldsymbol{\theta}$达到收敛精度为止，否则重复2,3

带有约束条件的最小二乘法

单纯的最小二乘法对于包含噪声的学习过程经常有过拟合的弱点。原因是学习模型对于训练样本而言过度复杂。利用约束条件可以控制模型复杂程序。

部分空间约束的最小二乘学习法

对于有参数的线性模型，

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{j=1}^{b} \theta_j \phi_j(\boldsymbol{x}) = \boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{\phi}_j(\boldsymbol{x})$$

参数$\{\theta_j\}_{j=1}^{b}$可以自由设置，利用部分空间约束的最小二乘法，通过把参数空间限制在一定范围内，防止过拟合现象。

$$\min_{\boldsymbol{\theta}} J_{\mathrm{LS}}(\boldsymbol{\theta}) \qquad s.t. \quad \boldsymbol{P}\boldsymbol{\theta} = \boldsymbol{\theta}$$

其中$\boldsymbol{P}$是满足$\boldsymbol{P}^2 = \boldsymbol{P}$和$\boldsymbol{P}^T = \boldsymbol{P}$的$b \times b$维矩阵，表示的是矩阵$\boldsymbol{P}$的值域$\mathcal{R}(\boldsymbol{P})$的正交投影矩阵。通过附加约束条件$\boldsymbol{P}\boldsymbol{\theta} = \boldsymbol{\theta}$,参数$\boldsymbol{\theta}$不会偏 移到值域$\mathcal{R}(\boldsymbol{P})$的范围外。 部分空间约束的最小二乘学习法的解$\hat{\boldsymbol{\theta}}$， 一般是通过将最小二乘学习的设计矩阵$\boldsymbol{\Phi}$置换为$\boldsymbol{\Phi}\boldsymbol{P}$的方式求得

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{\Phi}\boldsymbol{P})^{\dagger}\boldsymbol{y}$$

拉格朗日对偶问题：

可微分的凸函数$f\colon \mathbb{R}^d \to \mathbb{R}$和$\boldsymbol{g}\colon \mathbb{R}^d \to \mathbb{R}^p$的约束条件的最小化问题

$$\min_{\boldsymbol{t}} f(\boldsymbol{t}) \qquad s.t. \quad \boldsymbol{g}(\boldsymbol{t}) \leqslant 0$$

可定义拉格朗日对偶问题：

使用拉格朗日待定因子：

$$\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_p)^T$$

和拉格朗日函数：

$$L(\boldsymbol{t}, \boldsymbol{\lambda}) = f(\boldsymbol{t}) + \boldsymbol{\lambda}^T \boldsymbol{g}(\boldsymbol{t})$$

采用以下方式进行定义：

$$\max_{\boldsymbol{\lambda}} \inf_{\boldsymbol{t}} L(\boldsymbol{t}, \boldsymbol{\lambda}) \qquad s.t. \quad \boldsymbol{\lambda} \geqslant 0$$

拉格朗日对偶问题的$\boldsymbol{t}$的解，与原来的问题的解是一致的。

部分空间约束的最小二乘法中，只使用了参数空间的一部分，但是由于正交投影矩阵P的设置有很大的自由度，因此在实际应用中操作起来有很大难度。ℓ_2约束的最小二乘学习法相对容易操作。

$$\min_{\boldsymbol{\theta}} J_{\mathrm{LS}}(\boldsymbol{\theta}) \qquad s.t. \quad \|\boldsymbol{\theta}\|^2 \leqslant R$$

ℓ_2约束的最小二乘学习法以参数空间的原点为圆心，在一定半径范围的圆（一般为超球）内进行求解。R表示的即是圆的半径。

利用拉格朗日对偶问题，通过求解下式的最优解，得到原最优化问题的解。

$$\max_{\boldsymbol{\lambda}} \min_{\boldsymbol{\theta}} \left[J_{\mathrm{LS}}(\boldsymbol{\theta}) + \frac{\lambda}{2}(\|\boldsymbol{\theta}\|^2 - R) \right] \qquad s.t. \lambda \geqslant 0$$

拉格朗日对偶问题的拉格朗日待定因子λ的解由圆的半径R决定， 如果不根据R来决定λ,而是直接指定λ的值， ℓ_2约束的最小二乘学习法的解$\hat{\boldsymbol{\theta}}$就可以通过下式求得：

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left[J_{\mathrm{LS}}(\boldsymbol{\theta}) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \right]$$

其中第一项$J_{\mathrm{LS}}(\boldsymbol{\theta})$表示的是对训练样本的拟合程度， 通过与第二项的$\frac{\lambda}{2}\|\boldsymbol{\theta}\|^2$相结合得到最小值，可防止对训练样本的过拟合。

令关于参数$\boldsymbol{\theta}$的偏微分等于0,可得：

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}$$

ℓ_2约束的最小二乘法通过将矩阵$\boldsymbol{\Phi}^T\boldsymbol{\Phi}$与$\lambda I$相加提高了其正则性，进而可以更稳定地进行逆矩阵的求解。 因此， ℓ_2约束的最小二乘学习法也称为ℓ_2正则化的最小二乘法。 $\|\boldsymbol{\theta}\|^2$称为正则项， λ为正则化参数。 ℓ_2正则化的最小二乘法在有些著作中也称为岭回归。

考虑设计矩阵$\boldsymbol{\Phi}$的奇异值分解：

$$\boldsymbol{\Phi} = \sum_{k=1}^{\min(n,b)} \kappa_k \boldsymbol{\psi}_k \boldsymbol{\varphi}_k^T$$

ℓ_2约束的最小二乘法的解可表示为：

$$\hat{\boldsymbol{\theta}} = \sum_{k=1}^{\min(n,b)} \frac{\kappa_k}{\kappa_k^2 + \lambda} (\boldsymbol{\psi}_k^T \boldsymbol{y}) \boldsymbol{\varphi}_k$$

当$\lambda = 0$时，ℓ_2约束的最小二乘法就与一般的最小二乘法相同，当设计矩阵$\boldsymbol{\Phi}$的计算条件很恶劣，即包含非常小的奇异值κ_k时，$\frac{\kappa_k}{\kappa_k^2} = \frac{1}{\kappa_k}$会非常大，训练输出向量$\boldsymbol{y}$包含的噪声会增大，$\ell_2$约束最小二乘法，$\frac{\kappa_k}{\kappa_k^2 + \lambda}$由于正常数$\lambda$的作用，不会变得过大，进而可以达到防止过拟合的目的。

通过使用$b \times b$的正则化矩阵G，就可以得到更普遍的表示方法

$$\min_{\boldsymbol{\theta}} J_{\mathrm{LS}}(\boldsymbol{\theta}) \qquad s.t. \quad \boldsymbol{\theta}^T \boldsymbol{G} \boldsymbol{\theta} \leqslant R$$

上述表示方式称为一般ℓ_2约束的最小二乘法。矩阵$\boldsymbol{G}$为对称正定矩阵时，$\boldsymbol{\theta}^T \boldsymbol{G} \boldsymbol{\theta} \leqslant R$可以把参数限制在椭圆形的区域内。一般$\ell_2$约束的最小二乘法的解：

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi} + \lambda \boldsymbol{G})^{-1} \boldsymbol{\Phi}^T \boldsymbol{y}$$