最小二乘法推导

数据

$$\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_N}]^T = \begin{bmatrix} \mathbf{x_1^T} \\ \mathbf{x_2^T} \\ \vdots \\ \mathbf{x_N^T} \end{bmatrix}$$

标签

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

损失函数方程如下

$$L(\mathbf{W}) = \sum_{i=1}^{N} ||\mathbf{W}^T \mathbf{x}_i - y_i||$$

也即是求

$$W_{LSE} = \arg\min_{\mathbf{W}} \sum_{i=1}^{N} ||\mathbf{W}^T \mathbf{x}_i - y_i||$$

矩阵表达如下

$$L(\mathbf{w})=[\mathbf{X}\mathbf{W}-\mathbf{Y}]^T[\mathbf{X}\mathbf{W}-\mathbf{Y}]$$

令

$$\mathbf{Z}=\mathbf{X}\mathbf{W}-\mathbf{Y}$$

$$L(\mathbf{w})=\mathbf{Z}^T\mathbf{Z}$$

$$\frac{dL(\mathbf{W})}{d\mathbf{W}}=\frac{dL(\mathbf{W})}{d\mathbf{Z}}\frac{d\mathbf{Z}}{d\mathbf{W}}$$

由矩阵求导法则

$$\frac{dL(\mathbf{W})}{d\mathbf{Z}}=2\mathbf{Z}^T$$

$$\frac{d\mathbf{Z}}{d\mathbf{W}}=\mathbf{X}$$

所以

$$\frac{dL(\mathbf{W})}{d\mathbf{W}}=\frac{dL(\mathbf{W})}{d\mathbf{Z}}\frac{d\mathbf{Z}}{d\mathbf{W}}=2\mathbf{Z}^T\mathbf{X}=2[\mathbf{X}\mathbf{W}-\mathbf{Y}]^T\mathbf{X}=0$$

$$[\mathbf{W}^T\mathbf{X}^T-\mathbf{Y}^T]\mathbf{X}=0$$

$$\mathbf{W}^T\mathbf{X}^T\mathbf{X}-\mathbf{Y}^T\mathbf{X}=0$$

$$\mathbf{W}^T\mathbf{X}^T\mathbf{X} = \mathbf{Y}^T\mathbf{X}$$

$$\mathbf{W}^T = \mathbf{Y}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

$$\mathbf{W} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

最小二乘法-几何解释

$$L(\mathbf{w}) = [\mathbf{XW} - \mathbf{Y}]^T\,[\mathbf{XW} - \mathbf{Y}]$$

对于

$$\mathbf{X} = [\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_N}]^T = \begin{bmatrix} \mathbf{X_1^T} \\ \mathbf{X_2^T} \\ \vdots \\ \mathbf{X_N^T} \end{bmatrix} = \begin{bmatrix} x_{11}, & x_{12}, & \cdots, & x_{1p} \\ x_{21}, & x_{22}, & \cdots, & x_{2p} \\ \vdots & & \ddots & \\ x_{N1}, & x_{N2}, & \cdots, & x_{NP} \end{bmatrix}$$

所以

$$\mathbf{XW} = w_1\begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{N1} \end{bmatrix} + w_2\begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{N2} \end{bmatrix} + \cdots + w_p\begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{Np} \end{bmatrix}$$

$$= w_1\mathbf{v_1} + w_2\mathbf{v_2} + \cdots + w_p\mathbf{v_p}$$

求解Y找出向量$\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_p}$线性组合中最接近$Y$的向量。也即是$Y$在$\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_p}$向量空间中的投影：

$$\mathbf{X}^T(\mathbf{Y} - \mathbf{XW}) = \mathbf{0}$$

也即是

$$\mathbf{X}^T\mathbf{Y} - \mathbf{X}^T\mathbf{X}\mathbf{W} = \mathbf{0}$$

$$\mathbf{X}^T\mathbf{X}\mathbf{W} = \mathbf{X}^T\mathbf{Y}$$

$$\mathbf{W} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

最小二乘法-高斯噪声-最大似然估计

最小二乘法求解

由前节推导可知，问题描述
损失函数方程如下

$$L(\mathbf{W}) = \sum_{i=1}^{N} ||\mathbf{W}^T\mathbf{x}_i - y_i||$$

也即是求

$$W_{LSE} = \arg\min_{\mathbf{W}} \sum_{i=1}^{N} ||\mathbf{W}^T\mathbf{x}_i - y_i||$$

$$\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \cdots, \mathbf{x_N}]^T = \begin{bmatrix} \mathbf{x_1^T} \\ \mathbf{x_2^T} \\ \vdots \\ \mathbf{x_N^T} \end{bmatrix} = \begin{bmatrix} x_{11}, & x_{12}, & \cdots, & x_{1p} \\ x_{21}, & x_{22}, & \cdots, & x_{2p} \\ \vdots & & \ddots & \\ x_{N1}, & x_{N2}, & \cdots, & x_{NP} \end{bmatrix}$$

最大似然估计求解

假设 $\varepsilon \sim N(0,\sigma^2)$ 为随机噪声，$Y_i = \mathbf{W}^T\mathbf{x}_i + \varepsilon$
所以 $Y_i|\mathbf{x}_i,\mathbf{W} \sim N(\mathbf{W}^T\mathbf{x}_i,\sigma^2)$
即

$$p(y_i|\mathbf{x}_i,\mathbf{W}) = \frac{1}{\sqrt{2\pi\sigma}}\exp\{-\frac{(y_i-\mathbf{W}^T\mathbf{x}_i)^2}{2\sigma^2}\}$$

似然函数如下

$$\begin{aligned}
\mathcal{L}(\mathbf{W}) &= \log \mathbf{P}(\mathbf{Y}|\mathbf{X},\mathbf{W}) \\
&= \log \prod_{i=1}^{N} p(y_i|\mathbf{x}_i,\mathbf{W}) \\
&= \sum_{i=1}^{N}\left[\log\frac{1}{\sqrt{2\pi\sigma}} - \frac{(y_i-\mathbf{W}^T\mathbf{x}_i)^2}{2\sigma^2}\right]
\end{aligned}$$

$$\begin{aligned}
\mathbf{W}_{MLE} &= \arg\max_{\mathbf{W}} \mathcal{L}(\mathbf{W}) \\
&= \arg\max_{\mathbf{W}} \sum_{i=1}^{N}\left[\log\frac{1}{\sqrt{2\pi\sigma}} - \frac{(y_i-\mathbf{W}^T\mathbf{x}_i)^2}{2\sigma^2}\right] \\
&= \arg\max_{\mathbf{W}} \sum_{i=1}^{N} -\frac{(y_i-\mathbf{W}^T\mathbf{x}_i)^2}{2\sigma^2} \\
&= \arg\max_{\mathbf{W}} \sum_{i=1}^{N} -(y_i-\mathbf{W}^T\mathbf{x}_i)^2 \\
&= \arg\min_{\mathbf{W}} \sum_{i=1}^{N} (y_i-\mathbf{W}^T\mathbf{x}_i)^2
\end{aligned}$$

由此可知，若假设噪声为 ε 服从正态分布，则最小二乘法和最大似然估计求解效果一致，即:
若 $Y = \mathbf{W}^T\mathbf{X} + \varepsilon$,其中 $\varepsilon \sim N(0,\sigma)$,则$\mathbf{W}_{LSE} = \mathbf{W}_{MLE}$

正则化-岭回归

对于最小二乘法

$$L(\mathbf{W}) = \sum_{i=1}^{N} ||\mathbf{W}^T\mathbf{x}_i - y_i||$$

$$\mathbf{W}_{LSE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

其中 $\mathbf{X}_{N\times P}$, 样本数为N, 特征数量为P, 一般$P \ll N$.

- 若 $N < p$，则 $\mathbf{X}^T\mathbf{X}$ 存在不可逆的情况
- 若 $N < p$，会发生过拟合

过拟合一般解决办法如下

- 增加数据
- 降维(特征选择/特征提取(PCA))
- 正则化(参数空间的约束)

对于线性回归，正则化框架如下

$$\mathbf{W}_{RidgeRegression} = \arg\min_{\mathbf{W}} \sum_{i=1}^{N} \left[(y_i - \mathbf{W}^T\mathbf{x}_i)^2 + \lambda\mathbf{W}^T\mathbf{W}\right]$$

矩阵表达如下

$$L(\mathbf{W}) = [\mathbf{X}\mathbf{W} - \mathbf{Y}]^T[\mathbf{X}\mathbf{W} - \mathbf{Y}] + \lambda\mathbf{W}^T\mathbf{W}$$

$$\begin{aligned}
\frac{dL(\mathbf{W})}{d\mathbf{W}} &= 2(\mathbf{X}\mathbf{W} - \mathbf{Y})^T\mathbf{X} + 2\lambda\mathbf{W}^T = 0 \\
&\Rightarrow (\mathbf{W}^T\mathbf{X}^T - \mathbf{Y}^T)\mathbf{X} + \lambda\mathbf{W}^T = 0 \\
&\Rightarrow \mathbf{W}^T\mathbf{X}^T\mathbf{X} - \mathbf{Y}^T\mathbf{X} + \lambda\mathbf{W}^T = 0 \\
&\Rightarrow \mathbf{W}^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}) = \mathbf{Y}^T\mathbf{X} \\
&\Rightarrow \mathbf{W}^T = \mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1} \\
&\Rightarrow \mathbf{W} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}
\end{aligned}$$

正则化-概率角度

这里的$\mathbf{W},\mathbf{x}_i$看作一维向量

贝叶斯角度

假设 $\mathbf{W}$ 的先验分布：

$$\mathbf{W} \sim N(\mathbf{0}, \sigma_w^2)$$

$$\mathbf{Y} = \mathbf{W}^T\mathbf{X} + \varepsilon$$

$$Y_i|\mathbf{x}_i, \mathbf{W} \sim N(\mathbf{W}^T\mathbf{x}_i, \sigma^2)$$

由此可得

$$p(y_i|\mathbf{x}_i, \mathbf{W}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(y_i - \mathbf{W}^T\mathbf{x}_i)^2}{2\sigma^2}\}$$

依据贝叶斯定理

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

似然函数推导如下

因$\mathbf{x}_i$为常量(观测量)，所以

$$P(Y_i|\mathbf{W}) = \sum_{\mathbf{x}} P(Y_i|\mathbf{x}_i, \mathbf{W}) = P(Y_i|\mathbf{x}_i, \mathbf{W})$$

所以

$$p(y_i|\mathbf{W}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(y_i - \mathbf{W}^T\mathbf{x}_i)^2}{2\sigma^2}\}$$

因为样本之间独立同分布，所以

$$P(\mathbf{Y}|\mathbf{W}) = \prod_{i=1}^{N} P(Y_i|\mathbf{W})$$

所以

$$
\begin{aligned}
P(\mathbf{Y}|\mathbf{W}) &= \prod_{i=1}^{N} p(y_i|\mathbf{W}) \\
&= \prod_{i=1}^{N} p(y_i|\mathbf{x}_i, \mathbf{W})
\end{aligned}
$$

由前面假设可知

$$
p(y_i|\mathbf{W}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(y_i - \mathbf{W}^T\mathbf{x}_i)^2}{2\sigma^2}\}
$$

$$
p(\mathbf{w}) = \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\{-\frac{||\mathbf{w}||^2}{2\sigma_w^2}\}
$$

最大后验概率如下

$$
\begin{aligned}
\mathbf{W}_{MAP} &= \arg\max_{\mathbf{W}} \prod_{i=1}^{N} p(\mathbf{W}|Y_i) \\
&\propto \arg\max_{\mathbf{W}} \prod_{i=1}^{N} p(Y_i|\mathbf{W})P(\mathbf{W}) \\
&\propto \arg\max_{\mathbf{W}} \sum_{i=1}^{N} \log\left[p(Y_i|\mathbf{W})P(\mathbf{W})\right] \\
&= \arg\max_{\mathbf{W}} \sum_{i=1}^{N} \log\left[\frac{1}{\sqrt{2\pi\sigma}}\frac{1}{\sqrt{2\pi\sigma_w}}\exp\{-\frac{(y_i-\mathbf{W}^T\mathbf{x}_i)^2}{2\sigma^2}-\frac{||\mathbf{w}||^2}{2\sigma_w^2}\}\right] \\
&= \arg\max_{\mathbf{W}} \sum_{i=1}^{N} \left[\log\frac{1}{\sqrt{2\pi\sigma}}+\log\frac{1}{\sqrt{2\pi\sigma_w}}-\frac{(y_i-\mathbf{W}^T\mathbf{x}_i)^2}{2\sigma^2}-\frac{||\mathbf{w}||^2}{2\sigma_w^2}\right] \\
&= \arg\max_{\mathbf{W}} \sum_{i=1}^{N} \left[-\frac{(y_i-\mathbf{W}^T\mathbf{x}_i)^2}{2\sigma^2}-\frac{||\mathbf{w}||^2}{2\sigma_w^2}\right] \\
&= \arg\min_{\mathbf{W}} \sum_{i=1}^{N} \left[\frac{(y_i-\mathbf{W}^T\mathbf{x}_i)^2}{2\sigma^2}+\frac{||\mathbf{w}||^2}{2\sigma_w^2}\right] \\
&= \arg\min_{\mathbf{W}} \sum_{i=1}^{N} \left[(y_i-\mathbf{W}^T\mathbf{x}_i)^2+\frac{2\sigma^2}{2\sigma_w^2}||\mathbf{w}||^2\right]
\end{aligned}
$$

总结如下

$$
\mathbf{W}_{MAP} = \arg\min_{\mathbf{W}} \sum_{i=1}^{N} \left[(y_i-\mathbf{W}^T\mathbf{x}_i)^2+\frac{2\sigma^2}{2\sigma_w^2}||\mathbf{w}||^2\right]
$$

$$
\mathbf{W}_{RidgeRegression} = \arg\min_{\mathbf{W}} \sum_{i=1}^{N} \left[(y_i-\mathbf{W}^T\mathbf{x}_i)^2+\lambda\mathbf{W}^T\mathbf{W}\right]
$$

可得出如下结论:

正则化的LSE $\Leftrightarrow$ MAP（$\mathbf{W}$先验分布为高斯分布，噪声为高斯分布)