线性回归的通用化概率解释

【参考资料】

PRML 第三章

1. 线性回归的拓展

回归问题的最简单模型是输入变量的线性组合：

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D \tag{1.1}$$

其中$\boldsymbol{x} = (x_1, \ldots, x_D)^T$。这就是线性回归模型。这个模型的关键性质是它是参数$w_0, \ldots, w_D$的一个线性函数，但同时，它也是输入变量x_i的一个线性函数，这给模型带来很大的局限性。因此我们这样扩展模型的类别：将输入变量进行非线性映射，然后再建立它们的线性组合，形式为：

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\boldsymbol{x}) \tag{1.2}$$

其中，$\phi_j(\boldsymbol{x})$被称为基函数（basis function），w_0是偏置。通过把下标j的最大值记作$M-1$，这个模型的参数总数为M。

通常，定义一个额外的"虚基函数"$\phi_0(\boldsymbol{x}) = 1$，有：

$$y(\boldsymbol{x}, \boldsymbol{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) \tag{1.3}$$

其中$w = (w_0, \ldots, w_{M-1})^T$且$\phi = (\phi_0, \ldots, \phi_{M-1})^T$。

当我们使用基函数时，实际上我们相当于对原始的输入进行了特征变换，新生成的特征就是各基函数的值。

通过使用非线性基函数，我们能够让函数$y(\boldsymbol{x}, \boldsymbol{w})$成为输入向量$\boldsymbol{x}$的一个非线性函数。但是，形如式（1.2）的模型仍被称为线性模型，因为这个函数是$\boldsymbol{w}$的线性函数。

多项式回归就是用基函数拓展的线性回归中的一种。除此以外，还会使用高斯基函数：

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$

或者sigmoid基函数：

$$\phi_j(x) = \sigma\left(\frac{x-\mu_j}{s}\right)$$

其中$\sigma(a)$是sigmoid函数，定义为：

$$\sigma_a = \frac{1}{1+\exp(-a)}$$

2. 最大似然与MSE

假设目标变量t由确定的函数$y(\boldsymbol{x},\boldsymbol{w})$给出，这个函数被附加了高斯噪声，即：

$$t = y(\boldsymbol{x},\boldsymbol{w}) + \epsilon \tag{2.1}$$

其中ϵ是一个零均值的高斯随机变量，精度（方差的倒数）为$\beta = 1/\sigma^2$。因此我们有：

$$p(t|\boldsymbol{x},\boldsymbol{w},\beta) = \mathcal{N}\left(t|y(\boldsymbol{x},\boldsymbol{w}),\beta^{-1}\right) \tag{2.2}$$

现在考虑一个输入数据集$\boldsymbol{X} = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$，对应的目标值为$\boldsymbol{T} = \{t_1,\ldots,t_N\}$。假设这些数据点是独立地从分布（2.2）中抽取的，那么我们可以得到下面的似然函数的表达式：

$$p(\mathbf{T}|\boldsymbol{X},\boldsymbol{w},\beta) = \prod_{n=1}^{N}\mathcal{N}\left(t_n|\boldsymbol{w}^T\boldsymbol{\phi}\left(\boldsymbol{x}_n\right),\beta^{-1}\right) \tag{2.3}$$

其中我们使用了式（1.3）。注意在有监督学习问题中，我们不是在寻找模型来对输入变量的概率分布建模。因此x总会出现在条件变量的位置上。从现在开始，为了保持记号的简洁性，我们在诸如$p(\mathbf{t}|\boldsymbol{x},\boldsymbol{w},\beta)$这类的表达式中不显式地写出$x$。

取对数似然函数，并且使用一元高斯分布的标准形式，我们有

$$\begin{aligned}\ln p(\mathbf{t}|\boldsymbol{w},\beta) &= \sum_{n=1}^{N}\ln\mathcal{N}\left(t_n|\boldsymbol{w}^T\boldsymbol{\phi}\left(\boldsymbol{x}_n\right),\beta^{-1}\right) \\ &= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\boldsymbol{w})\end{aligned} \tag{2.4}$$

其中平方和误差（MSE）函数的定义为

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}\left(\boldsymbol{x}_n\right)\right\}^2 \tag{2.5}$$

写出了似然函数后，我们就可以使用最大似然的方法确定w和β。首先求w，可以看到公式（2.4）给出的对数似然函数中，只有误差平方和那一项与参数w有关，这就是为什么我们选择MSE作为损失函数，其梯度为：

$$\nabla\ln p(\mathbf{t}|\boldsymbol{w},\beta) = \beta\sum_{n=1}^{N}\left\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}\left(\boldsymbol{x}_n\right)\right\}\boldsymbol{\phi}(\boldsymbol{x}_n)^T \tag{2.6}$$

令其等于0，可得

$$0 = \sum_{n=1}^{N}t_n\phi(x_n)^T - w^T\left(\sum_{n=1}^{N}\phi\left(x_n\right)\phi(x_n)^T\right)$$

求解w，我们有

$$\boldsymbol{w}_{ML} = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\mathbf{t} \tag{2.7}$$

这被称为最小二乘问题的规范方程（normal equation）。这里Φ是一个$N \times M$，其中N是样本数量，M是基函数映射后的特征维度。Φ被称为设计矩阵（design matrix）：

$$\boldsymbol{\Phi}=\left(\begin{array}{cccc}\phi_0\left(\boldsymbol{x}_1\right) & \phi_1\left(\boldsymbol{x}_1\right) & \cdots & \phi_{M-1}\left(\boldsymbol{x}_1\right) \\ \phi_0\left(\boldsymbol{x}_2\right) & \phi_1\left(\boldsymbol{x}_2\right) & \cdots & \boldsymbol{\phi}_{M-1}\left(\boldsymbol{x}_2\right) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0\left(\boldsymbol{x}_N\right) & \phi_1\left(\boldsymbol{x}_N\right) & \cdots & \phi_{M-1}\left(\boldsymbol{x}_N\right)\end{array}\right)$$

现在，我们可以更进一步地认识偏置系数w_0。如果我们显式地写出偏置系数，那么误差函数（2.5）就变为

$$E_D(\boldsymbol{w})=\frac{1}{2} \sum_{n=1}^N\left\{t_n-w_0-\sum_{j=1}^{M-1} w_j \phi_j\left(\boldsymbol{x}_n\right)\right\}^2$$

令其关于w_0的导数为0，解出w_0，可得：

$$w_0=\bar{t}-\sum_{j=1}^{M-1} w_j \overline{\phi}_j$$

其中我们定义：

$$\bar{t}=\frac{1}{N} \sum_{n=1}^N t_n, \quad \overline{\phi}_j=\frac{1}{N} \sum_{n=1}^N \phi_j\left(\boldsymbol{x}_n\right)$$

因此偏置w_0补偿了目标值的平均值（在训练集上的）与基函数的值的平均值的加权求和之间的差。

我们也可以关于噪声精度参数β最大化似然函数（2.4），结果为：

$$\sigma_{ML}^2=\frac{1}{\beta_{ML}}=\frac{1}{N} \sum_{n=1}^N\left\{t_n-\boldsymbol{w}_{ML}^T \boldsymbol{\phi}\left(\boldsymbol{x}_n\right)\right\}^2$$

因此我们可以看到噪声的方差由目标值在回归函数周围的残留方差（residual variance）给出。

3. L2正则与贝叶斯先验

现在，我们假定参数w服从一个精度为α，均值为0的高斯先验分布，即：

$$p(\boldsymbol{w} \mid \boldsymbol{\alpha})=\mathcal{N}\left(\boldsymbol{w} \mid \mathbf{0}, \alpha^{-1} \boldsymbol{I}\right)=\left(\frac{\alpha}{2 \pi}\right)^{(M+1) / 2} \exp \left\{-\frac{\alpha}{2} \boldsymbol{w}^T \boldsymbol{w}\right\}$$

于是根据贝叶斯规则，我们可以得到后验概率：

$$p(w \mid X, t, \alpha, \beta)=p(t \mid X, w, \beta) p(w, \alpha) \tag{3.1}$$

进行最大后验估计，对上面的式子做对数似然，并且去除无关项后，可以得到：

$$\ln p(\boldsymbol{w} \mid \mathbf{t})=-\frac{\beta}{2} \sum_{n=1}^N\left\{t_n-\boldsymbol{w}^T \boldsymbol{\phi}\left(\boldsymbol{x}_n\right)\right\}^2-\frac{\alpha}{2} \boldsymbol{w}^T \boldsymbol{w} \tag{3.2}$$

这告诉我们，后验分布关于w的最大化等价于对平方和误差函数加上一个二次正则项进行最小化，其中正则系数$\lambda = \frac{\alpha}{\beta}$。

所以，**当我们对参数作L2正则化时，实际上我们就是要求参数服从于一个高斯的先验分布。**