3.1 线性基函数模型（Linear Basis Function Models）

下面是一种最简单的线性回归模型，该模型是由输入变量线性组合而成的

$$y(\mathbf{x},\mathbf{w}) = w_0 + w_1x_1 + w_2x_2 + \ldots + w_Dx_D \tag{3.1}$$

其中，$\mathbf{x} = (x_1,\ldots,x_D)^\top$, 该模型也被称作**线性回归(linear regression)**。

为了提高模型的拟合能力，引入非线性函数 $\phi(x)$，使得目标函数的形式为

$$y(\mathbf{x},\mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j\phi_j(\mathbf{x}) \tag{3.2}$$

其中$\phi_j(\mathbf{x})$也被叫做**基函数(basis function)**，w_0可以用来描述函数的偏置值，通常被叫做**偏置(bias)**参数

定义dummy基函数$\phi_0(\mathbf{x}) = 1$，上式简化为

$$y(\mathbf{x},\mathbf{w}) = \sum_{j=0}^{M-1} w_j\phi_j(x) = \mathbf{w}^\top\boldsymbol{\phi}(\mathbf{x}) \tag{3.3}$$

其中 $\mathbf{w} = (w_0,\ldots,w_{M-1})^\top$，$\boldsymbol{\phi} = (\phi_0,\ldots,\phi_{M-1})^\top$

几个基函数模型例子：

- **多项式基函数（Polynomial basis function）**： $\phi_j(x) = x^j$
- **高斯基函数（Gaussian basis function）**：

$$\phi_j(x) = \exp\{-\frac{(x-\mu_j)^2}{2s^2}\} \tag{3.4}$$

- **S基函数（Sigmoidal basis function）**：

$$\phi_j(x) = \sigma(\frac{x-\mu_j}{s}) \tag{3.5}$$

$$\sigma(a) = \frac{1}{1+\exp(-a)} \tag{3.6}$$

3.1.1 最大似然和最小二乘

在第一章，通过最小化平方和误差来训练多项式函数，并且证明了平方和误差函数可以在高斯噪声模型下，通过最大似然估计法导出，该节将讨论最大似然和最小二乘之间的关系。

假设目标变量等于一个确定函数和一个高斯噪声的和

$$t = y(\mathbf{x},\mathbf{w}) + \epsilon \tag{3.7}$$

其中 ϵ 服从均值为0，精度为 β 的的高斯分布

$$p(t|\mathbf{x},\mathbf{w},\beta) = \mathcal{N}(t|y(\mathbf{x},\mathbf{w}),\beta^{-1}) \tag{3.8}$$

t的条件期望为

$$\mathbb{E}[t|\mathbf{x}] = \int tp(t|\mathbf{x})dt = y(\mathbf{x}, \mathbf{w}) \tag{3.9}$$

接下来根据高斯分布独立同分布地采取样本点 $\{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N)\}$，设 $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$，$\mathsf{t} = (t_1, \ldots t_N)^\top$。联合分布函数为

$$p(\mathsf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^\top \phi(\mathbf{x}_n), \beta^{-1}) \tag{3.10}$$

通过已知的样本点$(\mathbf{X}, \mathsf{t})$，可以对参数进行极大似然估计

$$\begin{aligned}
\ln p(\mathsf{t}|\mathbf{w}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^\top \phi(\mathbf{x}_n), \beta^{-1}) \\
&= \sum_{n=1}^{N} \ln \sqrt{\frac{\beta}{2\pi}} \exp(\frac{-\beta(t_n - \mathbf{w}^\top \phi(\mathbf{x}_n))^2}{2}) \\
&= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta \sum_{n=1}^{N} \frac{(t_n - \mathbf{w}^\top \phi(\mathbf{x}_n))^2}{2} \\
&= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w})
\end{aligned} \tag{3.11}$$

其中$E_D(\mathbf{w})$即为最小二乘优化的目标

$$E_D(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)\}^2 \tag{3.12}$$

我们希望通过对 $\mathbf{w}$ 进行的取值，使得最大似然估计函数的值达到最小，通过下式求出最大似然估计对数函数的梯度

$$\nabla_{\mathbf{w}} \ln p(\mathsf{t}|\mathbf{w}, \beta) = \beta \sum_{n=1}^{N} (t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)) \phi(\mathbf{x}_n)^\top \tag{3.13}$$

设置梯度为0

$$\mathbf{0} = \sum_{n=1}^{N} t_n \phi(\mathbf{x_n}) - \mathbf{w}^\top (\sum_{n=1}^{N} \phi(\mathbf{x}_n) \phi(\mathbf{x}_n)^\top) \tag{3.14}$$

解出最大似然点 $\mathbf{w}$ 的值

$$\mathbf{w}_{ML} = (\Phi^\top \Phi)^{-1} \Phi^\top \mathsf{t} \tag{3.15}$$

这个方程也叫做**正规方程（normal equations）**，其中 Φ 为 $N \times M$ 的**设计矩阵(design matrix)**

$$\Phi = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \ldots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix} \tag{3.16}$$

其中，定义 Φ 的**伪逆矩阵（Moore-Penrose pseudo-inverse）**如下

$$\Phi^{\dagger}=(\Phi^{\top}\Phi)^{-1}\Phi^{\top} \tag{3.17}$$

其中，当 Φ 是可逆方阵的时候，$\Phi^{\dagger}=(\Phi^{\top}\Phi)^{-1}\Phi^{\top}=\Phi^{-1}$

将**偏置参数（bias）**w_0单独提取出，误差函数(3.12)变为

$$E_D(\mathbf{w})=\frac{1}{2}\sum_{n=1}^{N}\{t_n-w_0-\sum_{j=1}^{M-1}w_j\phi_j(\mathbf{x}_n)\}^2 \tag{3.18}$$

对 w_0 求偏导，求得 w_0 的最优点

$$w_0=\bar{t}-\sum_{j=1}^{M-1}w_j\overline{\phi_j} \tag{3.19}$$

其中，对$\bar{t},\overline{\phi_j}$的定义为

$$\bar{t}=\frac{1}{N}\sum_{n=1}^{N}t_n \qquad \overline{\phi_j}=\frac{1}{N}\sum_{n=1}^{N}\phi_j(\mathbf{x}_n) \tag{3.20}$$

同理，可以求出似然对数函数 $\ln p(\mathsf{t}|\mathbf{w}_{ML},\beta)$ 关于 β 的偏导数，令偏导数等于0可以得到精度 β 的最大似然估计

$$\frac{\partial(\ln p(\mathsf{t}|\mathbf{w}_{ML},\beta))}{\partial(\beta)}=\frac{N}{2\beta}-E_D(\mathbf{w}_{ML})=0 \tag{3.21}$$

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

3.1.2 最小二乘的几何解释

定义一个N维空间，其中每一维的坐标表示一次训练的输出值，$\mathsf{t}=(t_1,\ldots,t_N)^{\top}$构成该N维空间的一个目标向量。每一个基函数 ϕ_j 在N个数据点上的输出构成了该N维空间下的一个特征向量 $\boldsymbol{\varphi}_j=(\phi_j(\mathbf{x}_1),\ldots,\phi_j(\mathbf{x}_N))^{\top}$，M个基函数生成的这样的M个特征向量 $\{\boldsymbol{\varphi}_0,\boldsymbol{\varphi}_1,\ldots,\boldsymbol{\varphi}_{M-1}\}$，构成了N维空间下的一个M维子空间 $\mathcal{S}$，即 Φ 的列空间。对于任意线性回归模型而言，定义模型N次输出产生的N维输出向量$\mathbf{y}$如下所示

$$\begin{aligned}\mathsf{y}&=(y(\mathbf{x}_1,\mathbf{w}),\ldots,y(\mathbf{x}_N,\mathbf{w}))^{\top}=(\mathbf{w}^{\top}\boldsymbol{\phi}(\mathbf{x}_1),\ldots,\mathbf{w}^{\top}\boldsymbol{\phi}(\mathbf{x}_N))^{\top}\\&=(\mathbf{w}^{\top}\Phi^{\top})^{\top}=\Phi\mathbf{w}=\sum_{j=0}^{M-1}w_j\boldsymbol{\varphi}_j\end{aligned}$$

根据上述定义可以得到两个结论：

- $E_D(\mathbf{w})=\frac{1}{2}\sum_{n=1}^{N}\{t_n-\mathbf{w}^{\top}\boldsymbol{\phi}(\mathbf{x}_n)\}^2=\frac{1}{2}\sum_{n=1}^{N}\{t_n-y(\mathbf{x}_n,\mathbf{w})\}^2=\frac{1}{2}\|\mathsf{t}-\mathsf{y}\|^2$，即为 t,y 之间的欧氏距离
- $\mathbf{y}$ 在 $\mathcal{S}$ 空间里。