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

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_D x_D \tag{3.1}$$

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）**：

$$\phi_j(x) = x^j \tag{3.4}$$

- **高斯基函数（Gaussian basis function）**：

$$\phi_j(x) = \exp\{-\frac{(x-\mu_j)^2}{2s^2}\} \tag{3.5}$$

- **S基函数（Sigmoidal basis function）**：

$$\phi_j(x) = \sigma(\frac{x-\mu_j}{s}) \tag{3.6}$$

$$\sigma(a) = \frac{1}{1+\exp(-a)} \tag{3.7}$$

3.1.1 最大似然和最小二乘（Maximum likelihood and least squares）

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

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \tag{3.8}$$

其中 ϵ 服从均值为0，精度为 β 的的高斯分布

$$p(t|\mathbf{x},\mathbf{w},\beta)=\mathcal{N}(t|y(\mathbf{x},\mathbf{w}),\beta^{-1}) \tag{3.9}$$

t的条件期望为

$$\mathbb{E}[t|\mathbf{x}]=\int tp(t|\mathbf{x})dt=y(\mathbf{x},\mathbf{w}) \tag{3.10}$$

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

$$p(\mathsf{t}|\mathbf{X},\mathbf{w},\beta)=\prod_{n=1}^{N}\mathcal{N}(t_n|\mathbf{w}^\top\phi(\mathbf{x}_n),\beta^{-1}) \tag{3.11}$$

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

$$\begin{aligned}\ln p(\mathsf{t}|\mathbf{w},\beta)&=\sum_{n=1}^{N}\ln\mathcal{N}(t_n|\mathbf{w}^\top\phi(\mathbf{x}_n),\beta^{-1})\\&=\sum_{n=1}^{N}\ln\sqrt{\frac{\beta}{2\pi}}\exp(\frac{-\beta(t_n-\mathbf{w}^\top\phi(\mathbf{x}_n))^2}{2})\\&=\frac{N}{2}\ln\beta-\frac{N}{2}\ln(2\pi)-\beta\sum_{n=1}^{N}\frac{(t_n-\mathbf{w}^\top\phi(\mathbf{x}_n))^2}{2}\\&=\frac{N}{2}\ln\beta-\frac{N}{2}\ln(2\pi)-\beta E_D(\mathbf{w})\end{aligned} \tag{3.12}$$

其中$E_D(\mathbf{w})$即为最小二乘优化的目标

$$E_D(\mathbf{w})=\frac{1}{2}\sum_{n=1}^{N}\{t_n-\mathbf{w}^\top\phi(\mathbf{x}_n)\}^2 \tag{3.13}$$

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

$$\nabla_{\mathbf{w}}\ln p(\mathsf{t}|\mathbf{w},\beta)=\beta\sum_{n=1}^{N}(t_n-\mathbf{w}^\top\phi(\mathbf{x}_n))\phi(\mathbf{x}_n)^\top \tag{3.14}$$

设置梯度为0

$$\mathbf{0}=\sum_{n=1}^{N}t_n\phi(\mathbf{x_n})-\mathbf{w}^\top(\sum_{n=1}^{N}\phi(\mathbf{x}_n)\phi(\mathbf{x}_n)^\top) \tag{3.15}$$

解出最大似然点 $\mathbf{w}$ 的值

$$\mathbf{w}_{ML}=(\Phi^\top\Phi)^{-1}\Phi^\top\mathsf{t} \tag{3.16}$$

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)**

$$\Phi=\begin{pmatrix}\phi_0(\mathbf{x}_1)&\phi_1(\mathbf{x}_1)&\cdots&\phi_{M-1}(\mathbf{x}_1)\\\phi_0(\mathbf{x}_2)&\phi_1(\mathbf{x}_2)&\cdots&\phi_{M-1}(\mathbf{x}_2)\\\vdots&\vdots&\ddots&\vdots\\\phi_0(\mathbf{x}_N)&\phi_1(\mathbf{x}_N)&\dots&\phi_{M-1}(\mathbf{x}_N)\end{pmatrix} \tag{3.17}$$

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

$$\Phi^{\dagger}=(\Phi^{\top}\Phi)^{-1}\Phi^{\top} \tag{3.18}$$

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

$$E_D(\mathbf{w})=\frac{1}{2}\sum_{n=1}^{N}\{t_n-w_0-\sum_{j=1}^{M-1}w_j\phi_j(\mathbf{x}_n)\}^2 \tag{3.19}$$

对 w_0 求偏导，求得 w_0 的最优点

$$w_0=\bar{t}-\sum_{j=1}^{M-1}w_j\overline{\phi_j} \tag{3.20}$$

其中，对$\bar{t},\overline{\phi_j}$的定义为

$$\bar{t}=\frac{1}{N}\sum_{n=1}^{N}t_n \qquad \overline{\phi_j}=\frac{1}{N}\sum_{n=1}^{N}\phi_j(\mathbf{x}_n) \tag{3.21}$$

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

$$\begin{aligned}\frac{\partial(\ln p(\mathsf{t}|\mathbf{w}_{ML},\beta))}{\partial(\beta)}&=\frac{N}{2\beta}-E_D(\mathbf{w}_{ML})=0\\ \frac{1}{\beta_{ML}}&=\frac{1}{N}\sum_{n=1}^{N}\{t_n-\mathbf{w}_{ML}^{\top}\boldsymbol{\phi}(\mathbf{x}_n)\}^2\end{aligned} \tag{3.22}$$

3.1.2 最小二乘的几何解释（Geometry of least squares）

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，$\mathsf{t}=(t_1,\dots,t_N)^{\top}$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 $\boldsymbol{\varphi}_j=(\phi_j(\mathbf{x}_1),\dots,\phi_j(\mathbf{x}_N))^{\top}$，M个基函数生成的这样的M个特征向量 $\{\boldsymbol{\varphi}_0,\boldsymbol{\varphi}_1,\dots,\boldsymbol{\varphi}_{M-1}\}$，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y如下所示

$$\begin{aligned}\mathsf{y}&=(y(\mathbf{x}_1,\mathbf{w}),\dots,y(\mathbf{x}_N,\mathbf{w}))^{\top}=(\mathbf{w}^{\top}\boldsymbol{\phi}(\mathbf{x}_1),\dots,\mathbf{w}^{\top}\boldsymbol{\phi}(\mathbf{x}_N))^{\top}\\ &=(\mathbf{w}^{\top}\Phi^{\top})^{\top}=\Phi\mathbf{w}=\sum_{j=0}^{M-1}w_j\boldsymbol{\varphi}_j\end{aligned} \tag{3.23}$$

根据上述定义可以显然得到两个结论：

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

根据上述两个结论可知，使得$E_D(\mathbf{w})$取得最小值的y对应于t在空间$\mathcal{S}$上的投影，容易证得此时对应的参数值恰好为$\mathbf{w}_{ML}$，即$\mathsf{y}_{ML} = \Phi\mathbf{w}_{ML}$为t在$\mathcal{S}$上的投影。

下面来证明这个结论，根据(3.15)将$\mathbf{w}_{ML}$进行a代换

$$\mathsf{y}_{ML} = \Phi\mathbf{w}_{ML} = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top\mathsf{t} \tag{3.24}$$

下面只需证明$P = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top$是$\Phi$的列空间$\mathcal{S}$的一个投影矩阵，显然

$$\Phi^\top(\mathsf{t} - P\mathsf{t}) = \mathbf{0} \tag{3.25}$$

故上述的结论成立。

当任意两个φ_j近似于同一个方向时，$\Phi^\top\Phi$接近奇异，会导致$\mathbf{w}$的数值不稳定，一般的方法是给$\Phi^\top\Phi$添加**正则项（Regularization）**

3.1.3 序列学习（Sequential learning）

基于大批量的数据集进行建模是很困难的，所以考虑将数据集分成若干部分，对每一次的数据集进行建模并对参数进行更新，这样的方式称为**序列学习（sequential learning）**或者**在线学习（on-line learning）**

我们可以通过**随机梯度下降法（stochastic gradinet descent）**，也叫做**序列梯度下降法（sequential learning）**来对序列学习模型进行建模

设总的误差由每个数据点的误差之和组成的，即$E = \sum_n E_n$。因此随机梯度下降算法可以表示为

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \eta\nabla_{\mathbf{w}^{(\tau)}}E_n \tag{3.26}$$

其中τ为迭代次数，η为学习率，通过（3.12），上式也可以写成

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} + \eta(t_n - \mathbf{w}^{(\tau)\top}\boldsymbol{\phi}(\mathbf{x}_n))\boldsymbol{\phi}(\mathbf{x}_n) \tag{3.27}$$

该算法也叫做**最小均方误差算法（least-mean-squares, LMS）**。为了使得算法收敛，学习率这一超参数的选择是十分重要的。

3.1.4 正则化最小二乘（Regularized least squares）

为了防止**过拟合（overfitting）**，我们通常向损失函数引入一个**正则项（regularization term）**

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) \tag{3.28}$$

其中λ是**正则系数（regularization coefficient）**，控制过拟合的代价，其中λ越大，过拟合的代价越高，越不容易发生过拟合，可供选择的正则项有很多种，其中最简单的一种就是**L2正则项**

$$E_W(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\mathbf{w}^\top\mathbf{w} \tag{3.29}$$

于是引入了L2正则项的损失函数就变为了

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^\top \mathbf{w} \tag{3.30}$$

在机器学习中，正则化也叫做**权重衰减（weight decay）**。由于L2正则项保持了$\mathbf{w}$的平方性，所以依然可以通过正规方程找到最优的$\mathbf{w}$解，带L2正则项的正规方程可以表示为

$$\mathbf{w} = (\lambda \mathbf{I} + \Phi^\top \Phi)^{-1}\Phi^\top \mathbf{t} \tag{3.31}$$

不失一般性，我们一般用一种更一般的形式来表示带正则项的损失函数

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=0}^{M-1}|w_j|^q \tag{3.32}$$

L2正则项对应于$q=2$的情形，而L1正则项则对应于$q=1$的情形，带L1正则项的模型在统计学中被称为Lasso回归，当λ足够大的时候，Lasso回归可以使得权值向量$\mathbf{w}$中大多数的权值w_j都趋向于0，使得模型稀疏化。为了直观理解这个问题，我们可以将（3.32）的问题转化为一个带约束的（3.13）最优化问题

$$\begin{aligned} E_D(\mathbf{w}) &= \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n)\}^2 \\ \sum_{j=0}^{M-1}|w_j|^q &\le \eta \end{aligned} \tag{3.33}$$

带约束的最优化问题与(3.32)的问题可以通过**拉格朗日乘子法**相关联。可以看到，正则化的方法可以通过降低模型复杂度来有效的控制过拟合现象。

3.1.5 多维输出（Multiple outputs）

之前我们讨论的是单个目标变量t的情况， 这一节中将讨论K个目标变量构成的目标向量的情况，设目标向量为$\mathbf{t} = (t_1, t_2, \ldots, t_K)^\top$，用同一组基函数，不同的权重值$\mathbf{w}_k$来对目标向量进行建模，得到

$$\begin{aligned} \mathbf{y}(\mathbf{x}, \mathbf{W}) &= \mathbf{W}^\top \boldsymbol{\phi}(\mathbf{x}) \\ \mathbf{W} &= (\mathbf{w}_1, \ldots \mathbf{w}_k) = (w_{(m-1)k})_{M\times K} \end{aligned} \tag{3.34}$$

设目标向量的条件分布符合**各向同性高斯分布（Isotropic Gaussian）**

$$p(\mathbf{t}|\mathbf{x}, \mathbf{W}, \beta) = \mathcal{N}(\mathbf{t}|\mathbf{W}^\top \boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\mathbf{I}) \tag{3.35}$$

存在一组目标向量$\mathbf{t}_1, \ldots \mathbf{t}_N$，定义一个 $N\times K$ 的矩阵$\mathbf{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_N)^\top$，同理定义$\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\top$。对数似然函数可以表示为

$$\begin{aligned} \ln p(\mathbf{T}|\mathbf{X}, \mathbf{W}, \beta) &= \sum_{n=1}^{N}\ln \mathcal{N}(\mathbf{t}|\mathbf{W}^\top \boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\mathbf{I}) \\ &= \frac{NK}{2}\ln(\frac{\beta}{2\pi}) - \frac{\beta}{2}\sum_{n=1}^{N}\|\mathbf{t}_n - \mathbf{W}^\top \boldsymbol{\phi}(\mathbf{x}_n)\|^2 \end{aligned} \tag{3.36}$$

与前面的方法相似，可以得到$\mathbf{W}$的最大似然估计

$$\mathbf{W}_{ML} = (\Phi^\top \Phi)^{-1} \Phi^\top \mathbf{T} \tag{3.37}$$

对于每一个N维分量$\mathbf{t}_k$，可以得到该分量对应的权重$\mathbf{w}_k$

$$\mathbf{w}_k = (\Phi^\top \Phi)^{-1} \Phi^\top \mathbf{t}_k \tag{3.38}$$

可以发现，(3.37)可以分解为K个独立的回归问题，即只需要独立地考虑每一个单独的分量$\mathbf{t}_k$来简化问题。