用于回归的线性模型

- 回归问题的目标：在给定D维输入变量x的情况下，预测一个或者多个连续目标变量t的值。
- 线性回归模型的最简单形式：是输入变量的线性函数。
- 基函数：将一组输入变量的非线性函数进行线性组合。这样的模型是参数的线性函数，同时关于输入变量是非线性的。

3.1 线性基函数模型

线性回归：

$$y(x,w)=w_0+w_1x_1+\cdots w_Dx_D$$

关键在于这个函数既是参数的线性函数，也是输入变量的线性函数，所以给模型带来了极大的局限性。所以进行以下拓展：

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

其中$\phi_j(x)$被称为基函数，参数的总数为M.
而w_0使得数据可以存在任意固定的偏置，这个值通常被称为偏置参数。此时，如果我们定义一个额外的基函数$\phi_0(x)=1$,则可以得到：

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

其中$w=(w_0,w_1,\cdots,w_{M-1})^T$,$\phi=(\phi_0,\phi_1,\cdots,\phi_{M-1})^T$.
在实际应用中，我们首先要进行特征提取，那么提取出的特征可以用$\phi_j(x)$来表示。
对于多项式进行拟合的例子是一个特例，其中每一个基函数都是x的幂指数的形式。但是存在局限性：多项式基函数是输入变量的全局函数，因此对于输入空间一个区域的改变将会影响所有其他区域。所以我们可以将输入空间划分为若干个区域，然后对于每个区域使用不同的多项式函数进行拟合，这样的函数叫做样条函数。
其他选择：

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

此外还可以选择sigmod函数

$$\phi_j(x)=\sigma(\frac{1}{1+\exp(-a)})$$

与sigmod函数等价，我们可以使用tanh函数，因为

$$tanh(x)=2\sigma(2x)-1$$

所以tanh函数的一般线性组合等价于sigmod函数的一般线性组合。
此外还可以使用傅利叶基函数。

3.1.1 最大似然与最小平方

假设t由确定的函数$y(x,w)$给出，这个函数被附加了高斯噪声：

$$t=y(x,w)+\epsilon$$

所以

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

所以我们对于假设的平方损失函数，对于x的一个新值，最优预测值由目标变量的条件均值得到：

$$\mathbb{E}[t|x] = \int tp(t|x)\mathrm{d}t = y(x,w)$$

但是由于我们假设高斯噪声，在给定x的条件下，t的条件分布是单峰的，对于实际任务可能并不合适，因此在之后我们会将其拓展到条件高斯分布的混合。

- 似然函数：
  对于输入数据集$X = \{x_1, \cdots, x_N\}$,对应的目标值为$t_1, \cdots, t_N$.将目标向量$\{t_n\}$组成列向量$\mathbf{t}$,假设这些数据点是从分布

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

  中独立抽取的，那么我们可以得到以下似然函数的表达式

  $$p(\mathbf{t}|X,w,\beta) = \prod_{n=1}^{N}\mathcal{N}(t_n|w^T\phi(x_n),\beta^{-1})$$

  注：由于x总会出现在条件变量的位置上，所以为了保持记号的简洁性，在表达式中不再显示写出x
  我们可以得到：

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

  其中

  $$E_D(w) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}^2$$

  然后我们就可以使用最大似然的方法求出w和β.首先对于w，我们求导可以得到：

  $$\nabla lnp(\mathbf{t}|w,\beta) = \beta\sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}\phi(x_n)^T$$

  令梯度等于零，得到：

  $$\sum_{n=1}^{N}t_nw^T\phi(x_n)^T - w^T(\sum_{n=1}^{N}\phi(x_n)\phi(x_n)^T)$$

  求解w，可以得到

  $$w = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{t}$$

  我们令$\Phi = \begin{bmatrix}\phi_0(x_1) & \phi_1(x_1) & \cdots & \phi_{M-1}(x_1) \\ \phi_0(x_2) & \phi_1(x_2) & \cdots & \phi_{M-1}(x_2) \\ \cdots & \cdots & \cdots & \cdots \\ \phi_0(x_N) & \phi_1(x_N) & \cdots & \phi_{M-1}(x_N)\end{bmatrix}$
  令$\Phi^\dagger = (\Phi^T\Phi)^{-1}\Phi^T$称为矩阵$\Phi$的伪逆矩阵。
  对于w_0:对于$E_D(w)$，如果我们显式的写出偏置参数，那么误差函数可以写为

  $$E_D(w) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - w_0 - \sum_{j=1}^{M-1}w_j\phi_j(x_n)\}^2$$

  求导令导数等于0，可以得到

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

其中

$$\bar{t} = \frac{1}{N} \sum_{n=1}^{N} t_n$$
$$\bar{\phi_j} = \frac{1}{N} \sum_{n=1}^{N} \phi_j(x_n)$$

所以偏置参数w_0补偿了目标值的平均值与基函数值的平均值加权求和之间的差。
另外对于β进行极大似然估计，可以得到

$$\frac{1}{\beta_{ML}} = \sum_{n=1}^{N} \{t_n - w_{ML}^T \Phi(x_n)\}^2$$

可以看到噪声精度的倒数等于目标值在回归函数周围的残留方差。

3.1.2 最小平方的几何描述

一个由t_n给出坐标轴的N维空间中，每个在N个数据点处估计的基函数$\phi_j(x_n)$也可以表示为该空间中的一个向量，记为φ_j ，如果基函数的数目M小于N，则M个向量φ可以张成一个M维子空间S,由于y是φ_j的任意组合，所以y在M为子空间的任意位置，所以w的最优解对应的是位于S的与t最小距离的y，所以该解对应于t在子空间上的正交投影。

3.1.3 顺序学习

随机梯度下降：在观测到模式n之后，我们可以立即更新参数：

$$w^{\tau+1} = w^{\tau} - \eta \nabla_w E_n$$

对于平方和误差：

$$w^{\tau+1} = w^{\tau} + \eta(t_n - w^{(\tau)T} \phi_n)\phi_n$$

其中 $\phi_n = \phi(x_n)$,这被称为最小均方或者LMS算法。

3.1.4 正则化最小平方

对于平方和误差函数：

$$E_D(w) = \frac{1}{2} \sum_{n=1}^{N} \{t_n - w^T \phi(x_n)\}^2$$

正则化项

$$E_w(w) = \frac{1}{2} w^T w$$

则总误差函数

$$\frac{1}{2} \sum_{n=1}^{N} \{t_n - w^T \phi(x_n)\}^2 + \frac{\lambda}{2} w^T w$$

这样的正则化选择被称为权重衰减，因为在顺序学习过程中，它倾向于让权值向0的方向衰减。我们可以得到解析解：

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

这是最小平方解的一个拓展。对于更一般的正则化项：

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}^2 + \frac{\lambda}{2}\sum_{n=1}^{N}|w_j|^q$$

对于q=1的情况，我们可以得到Lasso,当λ足够大的时候，我们可以得到某些系数w_j变为0，从而产生一个稀疏模型。Lasso等价于在满足下面限制的条件下最小化平方误差得到的结果：

$$\sum_{j=1}^{M}|w_j|^q \leq \eta$$

正则化方法就等同于限制模型复杂度，使之不会出现严重的过拟合。

3.1.5 多个变量

3.2 偏差方差分解

当我们知道了条件概率分布$p(t|x)$,每一种损失函数都可以给出最优的预测结果。当选择平方损失函数的时候，此时最优预测由条件期望h(x)给出：

$$h(x) = \mathbb{E}[t|x] = \int tp(t|x)\mathrm{d}t$$

而平方损失函数的期望可以写成：

$$\mathbb{E}[L] = \int\{y(x) - h(x)\}^2 p(x)\mathrm{d}x + \iint\{h(x) - t\}^2 p(x,t)\mathrm{d}x\mathrm{d}t$$

这里面的第二项与y无关，可以看作是数据本身的噪声造成的。而对于第一项，我们可以根据对不同的数据集D进行建模得到不同的y分别计算第一项，然后取均值去评估y的好坏。
对于一个特定的数据集D,我们可以将被积函数的第一项写为：

$$\{y(x;D) - h(x)\}^2$$

可以得到：

$$\begin{aligned}\{y(x;D) - h(x)\}^2 &= \{y(x;D) - \mathbb{E}_D[y(x;D)] + \mathbb{E}_D[y(x;D)] - h(x)\}^2 \\ &= \{y(x;D) - \mathbb{E}_D[y(x;D)]\}^2 + \{\mathbb{E}_D[y(x;D)] - h(x)\}^2 + 2\{y(x;D) - \mathbb{E}_D[y(x;D)]\}\{\mathbb{E}_D[y(x;D)] - h(x)\}\end{aligned}$$

然后我们对这个式子在D上进行求期望操作，可以得到第三项求期望后为0，剩下两项的期望中，第一项可看做偏置的平方，第二项则是方差。因此，由于上面提到平方损失函数的期望可以写成：

$$\mathbb{E}[L] = \int\{y(x) - h(x)\}^2 p(x)\mathrm{d}x + \iint\{h(x) - t\}^2 p(x,t)\mathrm{d}x\mathrm{d}t$$

所以平方损失函数的期望可以看作是偏差的平方+方差+噪声：其中

$$\text{偏置}^2 = \int \{\mathbb{E}_D[y(x;D)] - h(x)\}^2 p(x)\mathrm{d}x$$
$$\text{方差} = \int \mathbb{E}_D[\{y(x;D) - \mathbb{E}_D[y(x;D)]\}^2] p(x)\mathrm{d}x$$
$$\text{噪声} = \iint \{h(x) - t\}^2 p(x,t)\mathrm{d}x\mathrm{d}t$$

- 灵活的模型：偏置小，方差大
- 固定的模型：偏置大，方差小
- 所以正则化系数λ小的时候方差大，偏置小，λ大的时候方差小，偏置大。

3.3 贝叶斯线性回归

3.3.1 参数分布

首先对参数w引入先验概率分布，现将噪声β当作已知常数。由于似然函数是w的二次函数的指数形式，所以共轭先验为高斯分布

$$p(w) = \mathcal{N}(w|m_0, S_0)$$

接下来后验分布：

$$p(w|\mathbf{t}) = \mathcal{N}(w|m_N, S_N)$$

其中

$$m_N = S_N(S_0^{-1}m_0 + \beta\Phi^T t)$$
$$S_N^{-1} = S_0^{-1} + \beta\Phi^T\Phi$$

后验是高斯分布，因此众数与均值相等，因此最大后验权向量值w_{MAP}与m_N相等。
而如果我们给出无限宽的先验$S_0 = \alpha^{-1}I, \alpha \to 0$,则我们可以得到，均值与最大似然得到的结果$w_{ML}$相等。类似的，如果N=0，那么后验就等于先验，如果数据顺序到来，那么目前的后验就是后续数据的先验。
我们考虑高斯先验的一个特殊形式

$$p(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}I)$$

则后验可以得到：

$$p(w|\mathbf{t}) = \mathcal{N}(w|m_N, S_N)$$
$$m_N = \beta S_N\Phi^T\mathbf{t}$$
$$S_N^{-1} = \alpha I + \beta\Phi^T\Phi$$

取对数后得到

$$\ln p(w|\mathbf{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}^2 - \frac{\alpha}{2}w^T w + \text{const}$$

所以我们得到的结果等价于平方和误差加上一个二次正则项，其中$\lambda = \frac{\alpha}{\beta}$
此外，还有其他先验形式：

$$p(w|\alpha) = [\frac{q}{2}(\frac{\alpha}{2}^{\frac{1}{q}}\frac{1}{\Gamma(\frac{1}{q})})]^M \exp[-\frac{\alpha}{2}\sum_{j=0}^{M-1}|w_j|^q]$$

其中q=2时对应高斯先验。

3.3.2 预测分布

在进行预测时，我们可以得到预测分布（忽略输入变量）

$$p(t|\mathbf{t},\alpha,\beta)=\int p(t|w,\beta)p(w|\mathbf{t},\alpha,\beta)\mathrm{d}w$$

其中

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

可以得到预测分布的形式为

$$p(t|x,t,\alpha,\beta)=\mathcal{N}(m_N^T\phi(x),\sigma_N^2(x))$$

其中

$$\sigma_N^2(x)=\frac{1}{\beta}+\phi(x)^TS_N\phi(x)$$

方差中的第一项表示数据中的噪声，第二项显示了与参数w相关的不确定性。
当数据的数目不断增多的时候，后验概率会不断变窄，方差中第二项会逐渐趋于0，从而只与参数β控制的具有可加性的噪声有关。
而如果我们使用局部基函数，那么在距离基函数较远的区域，预测方差的第二项就会趋于0.因此，在对基函数之外的区域进行外差的时候，预测会相当确定，这是通常使用高斯过程。
若w和β都被当作未知数，那么预测分布将是一个学生t分布。

3.3.3 等价核

由于前面提到的后验概率分布形势为

$$p(w|\mathbf{t})=\mathcal{N}(w,S_N(S_0^{-1}m_0+\beta\Phi^T\mathbf{t}),S_N)$$

,所以我们将其均值带入$w(x,w)$,可以得到：

$$y(x,m_N)=\beta\phi(x)^TS_N\Phi^T\mathbf{t}=\sum_{n=1}^N\beta\phi(x)^TS_N\phi(x_n)t_n$$

所以预测均值由训练集目标变量t_n的线性组合给出：

$$y(x,m_N)=\sum_{n=1}^Nk(x,x_n)t_N$$

其中

$$k(x,x')=\beta\phi(x)^T\phi(x')$$

被称为平滑矩阵或者是 等价核
对于$y(x)$和$y(x')$的协方差：

$$\begin{aligned}\mathbf{cov}[y(x),y(x')]&=\mathbf{cov}[\phi(x)^Tw,w^T\phi(x')]\\&=\phi(x)^TS_N\phi(x')\\&=\beta^{-1}k(x,x')\end{aligned}$$

所以根据等价核的形式可以看到在附近的点处预测均值相关性较高，对于距离较远的点，均值相关性小。
所以我们可以不引入一组基函数，而是直接定义一个局部核函数，然后在给定观测数据集的条件下，利用核函数对新的输入进行预测。
一个等价核定义了模型的权值。可以证明

$$\sum_{n=1}^{N} k(x, x_n) = 1$$

对于所有的x都成立。
最后可以证明等价核满足一般核函数的性质即可以表示为非线性函数的向量$\varphi(x)$的内积形式

$$k(x, z) = \varphi(x)^T \varphi(z)$$

其中$\varphi(x) = \beta^{\frac{1}{2}} S_N^{\frac{1}{2}} \phi(x)$。

3.4 贝叶斯模型比较

对于L个模型$\{M1, \cdots, M_L\}$,在这里一个模型指的是观测数据D上的概率分布。我们假设数据是由这些模型中的一个生成的，这个不确定性可以根据先验$p(M_i)$表示，我们想估计后验概率

$$P(M_i|D) \propto p(M_i)P(D|M_i)$$

知道后验概率分布之后，就可以进行预测

$$p(t|x, D) = \sum_{i=1}^{L} p(t|x, M_I, D)p(M_i|D)$$

对于一个由w控制的模型，模型证据为

$$p(D|M_i) = \int p(D|w, M_i)p(w|M_i)\mathrm{d}w$$

我们假设后验分布在最大似然值w_{MAP}附近是一个尖峰，宽度为$\Delta w_{后验}$，那么可以近似这个积分，在近一步假设先验分布是平的，宽度为$\Delta w_{先验}$,就可以得到：

$$p(D|M_i) = \int p(D|w, M_i)p(w|M_i)\mathrm{d}w \simeq p(D|w_{MAP}, M_i)\frac{\Delta w_{后验}}{\Delta w_{先验}}$$

取对数可以得到

$$\ln p(D|M_i) \simeq \ln p(D|w_{MAP}, M_i) + \ln \frac{\Delta w_{后验}}{\Delta w_{先验}}$$

第一项说明拟合由最可能参数给出的数据，对于平的先验，这对应于对数似然，第二项对应于根据模型复杂度来惩罚模型。
贝叶斯模型比较框架中隐含的一个假设是，生成数据的真实的概率分布包含在考虑的模型集合当中。如果这个假设确实成立，那么我们我们可以证明，平均来看，贝叶斯模型比较会倾向于选择出正确的模型。为了证明这一点，考虑两个模型M1和M2，其中真实的概率分布对应于模型M1。对于给定的有限数据集，确实有可能出现错误的模型反而使贝叶斯因子较大的事情。但是，如果我们把贝叶斯因子在数据集分布上进行平均，那么我们可以得到期望贝叶斯因子

$$\int p(D|M_1) \ln \frac{p(D|M_1)}{p(D|M_2)} \geq 0$$

3.5 证据近似

引入对α和β的先验分布然后对超参数和w求积分进行预测，但是有时没有解析解，所以需要近似。

$$p(t|\mathbf{t}) = \iiint p(t|w,\beta)p(w,|\mathbf{t},\alpha,\beta)p(\alpha,\beta|\mathbf{t})\mathrm{d}w\mathrm{d}\alpha\mathrm{d}\beta$$

后验：

$$p(\alpha,\beta|\mathbf{t}) \propto p(\mathbf{t}|\alpha,\beta)p(\alpha,\beta)$$

如果先验相对较平，那么可以通过最大化$p(\mathbf{t}|\alpha,\beta)$得到。

3.5.1 计算证据函数

$$p(\mathbf{t}|\alpha,\beta) = \int p(\mathbf{t}|w,\beta)p(w|\alpha)\mathrm{d}w = (\frac{\beta}{2\pi})^{\frac{N}{2}}(\frac{\alpha}{2\pi})^{\frac{M}{2}}\int \exp\{-E(w)\}\mathrm{d}w$$

其中

$$\begin{aligned} E(w) &= \beta E_D(w) + \alpha E_W(w) = \frac{\beta}{2}||t_n - \Phi w||^2 + \frac{\alpha}{2}w^T w \\ &= E(m_N) + \frac{1}{2}(w-m_N)^T A(w-m_N) \end{aligned}$$

其中

$$\begin{aligned} A &= \alpha I + \beta\Phi^T\Phi = \nabla\nabla E(w) \\ E(m_N) &= \frac{\beta}{2}||t_n - \Phi m_N||^2 + \frac{\alpha}{2}m_N^T m_N \\ m_N &= \beta A^{-1}\Phi^T\mathbf{t} \end{aligned}$$

使用$S_N^{-1} = \alpha I + \eta\Phi^T\Phi$,可以得到：

$$A = S_N - 1$$

关于w的积分

$$\begin{aligned} \int \exp(-E(w))\mathrm{d}w &= \exp(-E(m_N))\int \exp(-\frac{1}{2}(w-m_N)^T A(w-m_N))\mathrm{d}w \\ &= \exp(-E(m_N))(2\pi)^{\frac{M}{2}}|A|^{-\frac{1}{2}} \end{aligned}$$

$$\ln p(\mathbf{t}|\alpha,\beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(m_N) - \frac{1}{2}\ln|A| - \frac{N}{2}\ln 2\pi$$

3.5.2 最大化证据函数

首先定义以下特征向量方程：

$$\beta\Phi^T\Phi u_i = \lambda_i u_i$$

所以可以知道，A的特征值为 $\alpha + \lambda_i$,所以

$$\frac{\partial \ln|A|}{\partial\alpha} = \sum_i \frac{1}{\lambda_i + \alpha}$$

所以$\ln p(\mathbf{t}|\alpha,\beta)$关于$\alpha$的驻点满足

$$\alpha m_N^T m_N = M - \alpha \sum_i \frac{1}{\lambda_i + \alpha} = \gamma$$

而$\alpha = \frac{\gamma}{m_N^T m_N}$,而$\gamma$和$m_N$都和$\alpha$有关，所以可通过迭代的方式求解。
类似的是，对于β

$$\frac{1}{\beta} = \frac{1}{N - \gamma} \sum_{n=1}^{N} \{t_n - m_N^T \phi(x_n)\}^2$$

3.5.3 参数的有效数目

由于$\beta \Phi^T \Phi$是正定矩阵，所以特征值为正，$\frac{\lambda_i}{\lambda_i + \alpha}$在0到1之间。所以$0 \leq \gamma \leq M$,对于$\lambda_i >> \alpha$的方向，$w_i$会与最大似然值接近。所以$\gamma$定义了参数的有效数目。剩下$M - \gamma$个参数被先验概率设为较小的值。因此修正了最大似然结果的偏差。