文章编号:1674-2869(2010)07-0099-08

非参数估计方法

张煜东[1,2],颜 俊[1],王水花[1],吴乐南[1]
(1.东南大学信息科学与工程学院,江苏 南京 210096;
2.哥仑比亚大学精神病学系脑成像实验室,纽约州 纽约 10032)

摘 要:为了解决函数估计问题,首先讨论了传统的参数回归方法.由于传统方法需要先验知识来决定参数模型,因此不稳健,且对模型敏感.因此,引入了基于数据驱动的非参数方法,无需任何先验知识即可对未知函数进行估计.本文主要介绍最新的8种非参数回归方法:核方法、局部多项式回归、正则化方法、正态均值模型、小波方法、超完备字典、前向神经网络、径向基函数网络.比较了不同的算法,给出算法之间的相关性与继承性.最后,将算法推广到高维情况,指出面临计算的维数诅咒与样本的维数诅咒两个问题.通过研究指出前者可以通过智能优化算法求解,而后者是问题固有的.

关键词:参数统计;非参数统计;核方法;局部多项式回归;正则化方法;正态均值模型;小波;超完备字典;前向神经网络;径向基函数网络
中图分类号:O212.7 文献标识码:A doi:10.3969/j.issn.1674-2869.2010.07.025

0 引 言

函数估计[1]是一个经典反问题,一般定义为给定输入输出样本对,求未知的系统函数[2].传统的方法为参数方法,即构建一个参数模型,再定义某个误差项,通过最小化误差项来求解模型的参数[3].

参数方法尽管较为简单,但不够灵活.例如参数模型假设有误,则会导致整个求解流程失败[4].因此学者们发展出不少新技术,非参数估计就是其中一项较好的方法.该方法无需提前假设参数模型的形式,而是基于数据结构推测回归曲面[5].

本文首先研究了经典的2种参数回归方法:最小二乘法与内插函数法,分析了它们的不足,然后主要讨论8种非参数回归方法:核方法、局部多项式回归、正则化方法(样条估计)、正态均值模型、小波方法、过完全字典、前向神经网络、径向基函数网络,尤其详细介绍了其间的相关性与继承性.最后,研究了高维情况下面临的计算维数诅咒与样本维数诅咒.

1 回归模型

考虑模型

$$y_i = r(x_i) + \varepsilon_i \tag{1}$$

式(1)中(x_i, y_i)为观测样本,假定误差ε具有方差齐性,则$r = E(y|x)$称为y对x的回归函数,简称回归.一般地,可以假设x取值在$[0,1]$区间内.定义"规则设计"为$x_i = i/n\ (i = 1, 2, \cdots, n)$.并定义风险函数为

$$R = \sum_{i=1}^{n} [r(x_i) - \hat{r}(x_i)]^2 = \sum_{i=1}^{n} [y_i - \hat{r}(x_i)]^2 \tag{2}$$

式(2)中$\hat{r}$为系统函数r的估计.

回归一词源于高尔顿(Galton),他和学生皮尔逊(Pearson)在研究父母身高和子女身高的关系时,以每对夫妇的平均身高为x,取其一个成年儿子的身高为y,并用直线$y = 33.73 + 0.512x$来描述y与x的关系.研究发现:如果双亲属于高个,则子女比他们还高的概率较小;反之,若双亲较矮,则子女以较大概率比双亲高.所以,个子偏高或偏矮的夫妇,其子女的身高有"向中心回归"的现象,因此高尔顿称描述子女与双亲身高关系的直线为"回归直线"[6].

然而,并非所有的$x-y$函数均有回归性,但历史沿用了这个术语.更为精确的表达是"函数估计".

收稿日期:2010-04-02
基金项目:国家自然科学基金(60872075);国家高技术发展计划(2008AA01Z227);高等学校科技创新工程重大项目培育资金项目(706028)
作者简介:张煜东(1985-),男,江苏苏州人,哥伦比亚大学博士后.研究方向:人工智能、数据挖掘、脑图像处理.

2 传统方法

理论上描述一个函数需要无穷维数据，因此函数估计本身也可称为“无穷维估计”[7]. 传统的估计方法有下列两种极端情形.

2.1 最小二乘法

此时假设 $\hat{r}(x)=\beta_0+\beta_1 x$，采用最小二乘法计算权值 $\beta=(\beta_0,\beta_1)$，得到的解为最小二乘估计[8]，

$$\hat{r}(x)=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} \tag{3}$$

则对给定样本点的估计 $\boldsymbol{r}=[\hat{r}(x_1),\hat{r}(x_2),\cdots,\hat{r}(x_n)]^{\mathrm{T}}$ 可写为

$$\boldsymbol{r}=\boldsymbol{X}\hat{\beta}=\boldsymbol{L}\boldsymbol{Y} \tag{4}$$

这里 $\boldsymbol{Y}=(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$. $\boldsymbol{L}=\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$ 称为帽子矩阵[9]. 以 5 个样本点的一维规则设计矩阵为例，此时

$$\boldsymbol{X}=\begin{bmatrix}0.2\\0.4\\0.6\\0.8\\1.0\end{bmatrix}\boldsymbol{L}=\begin{bmatrix}0.018&0.036&0.054&0.072&0.090\\0.036&0.072&0.109&0.145&0.181\\0.054&0.109&0.163&0.218&0.272\\0.072&0.145&0.218&0.290&0.363\\0.090&0.181&0.272&0.363&0.454\end{bmatrix} \tag{5}$$

$\boldsymbol{L}$ 满足 $\boldsymbol{L}=\boldsymbol{L}^{\mathrm{T}}$，$\boldsymbol{L}^2=\boldsymbol{L}$. 另外，$L$ 的迹等于输入数据的维数 p，即 $\mathrm{trace}(L)=p$. 这里输入数据是一维的，所以 $\mathrm{trace}(L)=1$.

2.2 内插函数法

此时对 $\hat{r}(x)$ 不加任何限制，得到的是该数据的一个内插函数[10]. 同样以 5 个样本点的一维规则设计矩阵为例，由于样本点的估计 $\boldsymbol{r}=[\hat{r}(x_1),\hat{r}(x_2),\cdots,\hat{r}(x_n)]^{\mathrm{T}}$ 完全等于 $(y_1,y_2,\cdots,y_n)^{\mathrm{T}}$，所以帽子矩阵为

$$\boldsymbol{L}=\begin{bmatrix}1&0&0&0&0\\0&1&0&0&0\\0&0&1&0&0\\0&0&0&1&0\\0&0&0&0&1\end{bmatrix} \tag{6}$$

2.3 两种方法的缺陷

图 1 给出了这两种极端拟合的示意图，数据是被高斯噪声干扰的正弦函数，采用上述两种方法拟合，结果表明：最小二乘法过光滑，未展现数据内部的关系；而内插函数法忽略了噪声影响，显得欠光滑.

从帽子矩阵也可看出，式(5)表明最小二乘法对每个数据的估计都利用了所有样本，这显然导致过光滑，且 x 值越大的数据权重越大，这明显与经验不符；反之，式(6)表明内插函数法仅仅利用了最邻近的样本数据，这显然导致欠光滑.

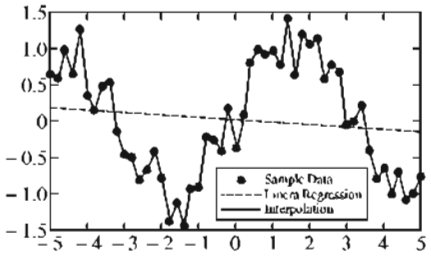

图 1 两种极端拟合

Fig. 1 Two extreme fitting

2.4 非参数回归的优势

非参数回归(non-parametric regression)作为最近兴起的一种函数估计方法，是一种分布无关(distribution free)的方法，即不依赖于数据的任何先验假设. 与此对应的是参数回归(parametric regression)，通常需要预先设置一个模型，然后求取该模型的参数. 非参方法的本质在于：模型不是通过先验知识而是通过数据决定. 需要注意的是，“非参数”并不表示没有参数，只是表示参数的数目、特征是可变的(flexible).

由于非参方法无需数据先验知识，其应用范围较参数方法更广，且性能更稳健. 其另一个优点是使用过程较参数方法更为简单. 然而，它也存在缺点，一般结构更复杂，需要更多的运算时间.

2.5 线性光滑器

需要说明的是，最小二乘法、内插函数法、核方法、正则化方法、正态均值模型均是线性光滑器. 定义为：若对每个 x，存在向量 $\boldsymbol{l}(x)=[l_1(x),\cdots,l_n(x)]^{\mathrm{T}}$，使得 $r(x)$ 的估计可写为

$$\hat{r}(x)=\sum_{i=1}^{n}l_i(x)y_i \tag{7}$$

则估计 $\hat{r}$ 为一个线性光滑器[11]. 显然权重 $l_i(x)$ 随着 x 而变化，这与信号处理中的“自适应滤波器”非常相似.

3 核回归

核方法[12]定义为

$$\hat{r}(x)=\sum_{i=1}^{n}l_i(x)Y_i \tag{8}$$

权重 l_i 由式(9)给出

$$l_i=K\left(\frac{x-x_i}{h}\right)\Big/\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right) \tag{9}$$

这里 h 是带宽，K 是一个核，满足 $K(x)\geqslant 0$，以及

$$\int K(x)\mathrm{d}x=1,\int xK(x)\mathrm{d}x=0,\int x^2K(x)\mathrm{d}x>0, \tag{10}$$

常用的核函数见表1.

表1 常用的核公式

Table 1 Frequently-used kernel formula

核	公式
boxcar	$K(x)=0.5*I(x)$
Caussian	$K(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$
Epanechnikov	$K(x)=\frac{3}{4}(1-x^2)I(x)$
Tricube	$K(x)=\frac{70}{81}(1-\lvert x\rvert^3)^3 I(x)$

以boxcar核为例,帽子矩阵为

$$\boldsymbol{L}=\begin{bmatrix}1/2 & 1/2 & 0 & 0 & 0\\ 1/3 & 1/3 & 1/3 & 0 & 0\\ 0 & 1/3 & 1/3 & 1/3 & 0\\ 0 & 0 & 1/3 & 1/3 & 1/3\\ 0 & 0 & 0 & 1/2 & 1/2\end{bmatrix} \tag{11}$$

显然,这可视作最小二乘法与内插函数法的折中.

为了估计带宽 h,首先必须估计风险函数,一般可采用缺一交叉验证得分

$$CV=\hat{R}(h)=\frac{1}{n}\sum_{i=1}^{n}\left[y_i-\hat{r}_{-i}(x_i)\right]^2 \tag{12}$$

这里 $\hat{r}_{-i}(x_i)$ 为未用第 i 个数据所得到的估计,使 CV 最小的 h,即为最佳带宽. 为了加速运算,可将式(12)重新写为

$$\hat{R}(h)=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i-\hat{r}(x_i)}{1-L_{ii}}\right]^2 \tag{13}$$

这里 L_{ii} 是光滑矩阵 $\boldsymbol{L}$ 的第 i 个对角线元素. 另一种方法是采用广义交叉验证法,规定

$$GCV(h)=\hat{R}(h)=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i-\hat{r}(x_i)}{1-v/n}\right]^2 \tag{14}$$

这里 $v=\mathrm{tr}(L)$.

4 局部多项式回归

采用核回归常会碰到下列2个问题[13]:1)若 x 不是规则设计的,则风险会增大,称为设计偏倚(design bias);2)核估计在接近边界处会出现较大偏差,称为边界偏倚(boundary bias). 为了解决这2个问题,可采用局部多项式回归.

局部多项式回归[14]可视作核估计的一个推广,首先定义权函数 $\omega_i(x)=K[(x_i-x)/h]$,选择 $a=\hat{r}(x)$ 来使得下面的加权平方和最小

$$\sum_{i=1}^{n}\omega_i(x)(y_i-a)^2 \tag{15}$$

利用高等数学知识,可以看出解为

$$\hat{r}(x)=\frac{\sum_{i=1}^{n}\omega_i(x)y_i}{\sum_{i=1}^{n}\omega_i(x)} \tag{16}$$

可见式(16)正好是核回归估计. 这表明核估计是由局部加权最小二乘得到的局部常数估计. 因此,若利用一个 p 阶的局部多项式而不是一个局部常数,就可能改进估计,使曲线更光滑. 定义多项式

$$P_x(u;a)=a_0+a_1(u-x)+\frac{a_2}{2!}(u-x)^2+\cdots+\frac{a_p}{p!}(u-x)^p \tag{17}$$

则局部多项式的思想是:选择使下列局部加权平方和

$$\sum_{i=1}^{n}\omega_i(x)\left[y_i-P_x(x_i;a)\right]^2 \tag{18}$$

最小的 a,估计 $\hat{\boldsymbol{a}}=(\hat{a}_0,\hat{a}_1,\hat{a}_p)^{\mathrm{T}}$ 依赖于目标值 x,最终有

$$\hat{r}(x)=P_x(x;\hat{a})=\hat{a}_0(x) \tag{19}$$

当 p 等于0时,等于核估计;当 $p=1$ 时,称为局部线性回归(local linear regression)估计[15],由于其算法简单且性能优越,较为常用.

5 基于正则化的回归

为了描述方便,这里假设数据点为 $[(x_0,y_0),(x_1,y_1),\cdots(x_{n-1},y_{n-1})]$. 在风险函数(2)后增加一项惩罚项,一般设为 $r(x)$ 的二阶导数

$$J=\sum_{i=0}^{n-1}\left[y_i-\hat{r}(x_i)\right]^2+\lambda\int\left[r''(x)\right]^2\mathrm{d}x \tag{20}$$

λ 控制了解的光滑程度:当 $\lambda=0$ 时,解为内插函数;当 $\lambda\to\infty$ 时,解为最小二乘直线;当 $0<\lambda<\infty$ 时,$\hat{r}(x)$ 是一个自然三次样条. 需要注意下列事项:首先三次样条表示曲线在结点(knot)之间是三次多项式,且在结点处有连续的一阶和二阶导数;其次一个 m 阶样条为一个逐段 $m-1$ 阶多项式,所以三次样条是4阶的($m=4$);第三,自然样条表示在边界点处二阶导数为0,即在边界点外是线性的;第四,样条的结点等于数据点.

为了加速计算,将数据点重新排序,假设 a,b 为样本点 x 的上下界,令 $a=t_1\leqslant t_2\leqslant\cdots\leqslant t_{n-1}=b$,这里 t 是 x 重新排序后的点,称为结点. 可用B样条基(B-spline basis)[16]作为该三次样条的基,即

$$\hat{r}(x)=\sum_{i=0}^{n-m}P_i b_{i,m}(t)\quad t\in[t_{m-1},t_{n-m}] \tag{21}$$

P_i 称为控制点,共 $n-m$ 个,形成一个凸壳. $n-m$ 个 B 样条基可通过如下计算,首先初始化:

$$b_{j,0}(t)=\begin{cases}1 & \text{if } t_j<t<t_{j+1}\\ 0 & \text{otherwise}\end{cases} \tag{22}$$

然后对 $i=1$，逐步 +1，直到 $i=m-1$，重复迭代下式：

$$b_{j,i}(t)=\frac{t-t_j}{t_{j+m-1}-t_j}b_{j,i-1}(t)+\frac{t_{j+m}-t}{t_{j+m}-t_{j+1}}b_{j,i+1}(t) \tag{23}$$

若结点等距，则称 B 样条是均匀的(uniform)，否则称为不均匀. 如果两个结点相等，计算过程会出现 0/0 情况，此时默认结果为 0.

令矩阵 $\boldsymbol{B}$ 的第 (i,j) 元素 $b_{ij}=b_j(x_i)$，矩阵 $\boldsymbol{\Omega}$ 的第 (i,j) 元素 $\Omega_{ij}=\int b_i^n(x)b_j^n(x)\mathrm{d}x$，则控制点可由式(24) 求得

$$\boldsymbol{P}=(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}+\lambda\boldsymbol{\Omega})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Y} \tag{24}$$

可见，样条也是一个线性光滑器.

表面上看，基于核的估计与基于正则化的估计原理与模型均不一致，但是 Silverman 证明了如下定理，样条估计 $\hat{r}(x)$ 可视作如下所示的一种渐近的核估计

$$l_i(x)\approx\frac{1}{f(x_i)h(x_i)}K\left(\frac{x_i-x}{h(x_i)}\right) \tag{25}$$

式中 $f(x)$ 是 x 的密度函数.

$$h(x)=\left[\frac{\lambda}{nf(x)}\right]^{1/4} \tag{26}$$

$$K(x)=\frac{1}{2}\exp\left[-\frac{|x|}{\sqrt{2}}\right]\sin\left(\frac{|x|}{\sqrt{2}}+\frac{\pi}{4}\right) \tag{27}$$

显然，若样本 x 是规则设计，则 $f(x)=1$，$h(x)=(\lambda/n)^{1/4}=h$，$l_i(x)\propto K[(x_i-x)/h]$，即此时样条估计可视作形如式(27)的渐近核估计.

6 正态均值模型

令 $\varphi_1,\varphi_2,\cdots$ 为一个标准正交基，则显然 $r(x)$ 可以展开为 $r(x)=\sum_{i=1}^{\infty}\theta_i\varphi_i$，定义

$$Z_j=\frac{1}{n}\sum_{i=1}^{n}y_i\varphi_j(x_i) \tag{28}$$

则随机变量 Z_j 是正态分布，且均值与方差满足：

$$E(Z_j)=\theta_j\quad V(Z_j)=\sigma^2/n \tag{29}$$

可见，若估计出 θ，则可近似求得 $\hat{r}(x)\approx\sum_{i=1}^{n}\theta_i\varphi_i$. 因此正态均值模型将 n 个样本的函数估计问题转换为估计 n 个正态随机变量 Z_j 的均值 θ 的问题[17].

若直接令 $\hat{\theta}=Z$，则显然得到一个很差的估计，下面给出风险更小的估计. 首先，必须做出一个关于 $\hat{\theta}$ 的风险估计，Stein 给出下列定理：令 $Z\sim N(\theta,V)$，$\hat{\theta}=\hat{\theta}(Z)$ 为 θ 的一个估计，并令 $g(Z_1,\cdots,Z_n)=\hat{\theta}-Z$. 则 $\hat{\theta}$ 的风险的一个无偏估计为

$$\hat{J}(z)=\mathrm{tr}(V)+2\mathrm{tr}(VD)+\sum_i g_i^2(z) \tag{30}$$

式中 $g_i=\hat{\theta}_i-z_i$，且 D 的第 (i,j) 个元素为 $g(z_1,\cdots,z_n)$ 的第 i 个元素关于 z_j 的偏导数[18].

假设 $\hat{\theta}=\boldsymbol{b}Z=(b_1Z_1,\cdots,b_nZ_n)$，式中 $\boldsymbol{b}$ 称为调节器，根据 $\boldsymbol{b}$ 的设置，存在下列 3 种情况：

① $\boldsymbol{b}=(b,b,\cdots,b)$，称为常数调节器(constant modulator)，此时令式(30)最小的 称为 James-Stein 估计；

② $\boldsymbol{b}=(1,\cdots,1,0,\cdots,0)$，称为嵌套子集选择调节器(nested subset selection modulator)，此时令式(30)最小的 $\hat{b}$ 称为 REACT 方法. 需要注意的是，若基选择傅立叶基，则该方法类似于频域低通滤波器方法.

③ $\boldsymbol{b}=(b_1,b_2,\cdots,b_n)$ 满足 $1\geqslant b_1\geqslant b_2\geqslant\cdots\geqslant b_n\geqslant 0$，称为单调调节器(monotone modulator)，该方法理论最优，但是需要的运算量太大，几乎不实用.

7 小波方法

小波方法[19] 适用于空间非齐次(spatially inhomogeneous)函数，即函数的光滑程度随着 x 会有本质性的变化. 它可视作正态均值模型的推广，但存在两点区别：一是采用小波基代替传统的正交基，因为小波基较一般的正交基具有局部化的优点，能实现多分辨率分析；另一点是采用了一种称为“阈”的收缩方式.

不妨假定父小波为 φ，母小波为 ψ，同时规定下标 (j,k) 的意义如下：

$$f_{j,k}(x)=2^{j/2}f(2^jx-k) \tag{31}$$

为了估计函数 r，用 $n=2^J$ 项展开来近似 r，

$$r(x)\approx\sum_{k=0}^{2^{J_0}-1}\alpha_{J_0,k}(x)\varphi_{J_0,k}(x)+\sum_{j=J_0}^{J}\sum_{k=0}^{2^j-1}\beta_{j,k}\psi_{j,k}(x) \tag{32}$$

这里 J_0 是任取常数，满足 $0\leqslant J_0\leqslant J$. α 称为刻度系数，β 称为细节系数. 那么如何估计这些系数？首先计算

$$S_k=\frac{1}{n}\sum_i\varphi_{J_0,k}(x_i)y_i \tag{33}$$

$$D_{jk}=\frac{1}{n}\sum_i\psi_{j,k}(x_i)y_i \tag{34}$$

S_k、D_{jk} 分别称为经验刻度系数与经验细节系

数，可知 $S_k \approx N(\alpha_{j0,k}, \sigma^2/n)$，$D_{jk} \approx N(\beta_{j,k}, \sigma^2/n)$，可估计方差为

$$\hat{\sigma} = \sqrt{n}\{\text{median}(|D_{j-1,k} - \text{median}(D_{j-1,k})| : k = 0, \cdots, 2^{j-1} - 1)\}/0.6745 \quad (35)$$

然后根据 S_k、D_{jk}、$\hat{\sigma}$ 可得 α 与 β 的估计如下：

$$\hat{\alpha}_{j0,k} = S_k \quad (36)$$

β 的估计形式稍许复杂，采用硬阈与软阈的方式分别为

$$\hat{\beta}_{jk} = \begin{cases} 0 & |D_{jk}| < \lambda \\ D_{jk} & |D_{jk}| \geq \lambda \end{cases} \quad (37)$$

$$\hat{\beta}_{jk} = \text{sign}(D_{jk})(|D_{jk}| - \lambda), \quad (38)$$

之所以采用阈的形式，是因为稀疏性(sparse)的思想[20]：对某些复杂函数，在小波基上展开时系数也是稀疏的. 因此，需要采用一种方式来捕获稀疏性. 然而，传统的 l_2 范数不能捕捉稀疏性，相反，L_1 范数与非零基数能够较好地捕捉稀疏性. 例如，考虑 n 维向量 $\boldsymbol{a} = (1, 0, \cdots, 0)$ 与 $\boldsymbol{b} = (1/n^{1/2}, \cdots, 1/n^{1/2})$，有 $\|\boldsymbol{a}\|_2 = \|\boldsymbol{b}\|_2 = 1$，可见，$L_2$ 范数无法区分稀疏性. 反之，$\|\boldsymbol{a}\|_1 = 1$，$\|\boldsymbol{b}\|_1 = n^{1/2}$，因此，$L_1$ 范数能提取稀疏性；另外，若令非零基数为 $J(\theta) = \{\#(\theta_i \neq 0)\}$，则 $J(a) = 1$，$J(b) = n$，因此，非零基数也能提取稀疏性. 最后，在正则化估计中若惩罚项分别为 L_1 范数或非零基数，则最优估计恰好对应着软阈估计与硬阈估计.

最后，需要解决阈估计中 λ 的计算问题，这里介绍两种最简单的方式：一是通用阈值(universal threshold)，即对所有水平的分辨率阈值均一致，

$$\lambda = \hat{\sigma}\sqrt{\frac{2\log n}{n}} \quad (39)$$

另一种是分层阈值(level-by-level threshold)，即对不同分辨率采用不同阈值，一般是通过最小化下式求得

$$S(\lambda_j) = \sum_{k=1}^{n_j}\left[\frac{\hat{\sigma}^2}{n} - 2\frac{\hat{\sigma}^2}{n}I(|\hat{\beta}_{jk}| \leq \lambda_j) + \min(\hat{\beta}_{jk}^2, \lambda_j^2)\right] \quad (40)$$

$$\lambda_j \in [0, (\hat{\sigma}/\sqrt{n_j})\sqrt{2\log n_j}]$$

式中 $n_j = 2_{j-1}$ 为在水平 j 的参数个数.

8 超完备字典

小波基较标准正交基的改进在于更加局部化，因此能实现对跳跃的捕捉. 然而，虽然小波基非常复杂，但面对各种复杂的函数还是不够灵活. 这种缺陷的根源在于：小波基是标准正交基，任意两个基函数之间正交，这保证了基函数简单完整的同时，也丧失了灵活性.

基追踪(basis pursuit)方法[21]的思想是采用一种超完备(overcomplete)的基，例如对“光滑加跳跃”的函数，传统的傅立叶基能够捕捉光滑部分，但是难以捕捉跳跃部分；采用小波基能轻易捕捉跳跃部分，但是描述光滑部分较为困难. 此时若将“傅立叶基”与“小波基”合并成一个新的基，则显然这种基能够轻松地估计“光滑加跳跃”函数.

但是，这种新的基不再正交，它以牺牲正交性来获得更好的灵活性[22]，故此时用“字典”来描述更精确，而本文为了简便统一仍采用“基”表述.

9 前向神经网络

以一个双层神经网络为例，记网络的输入神经元个数为 m，隐层神经元个数为 n，输出层神经元个数为 q，则网络结构如图 2 所示.

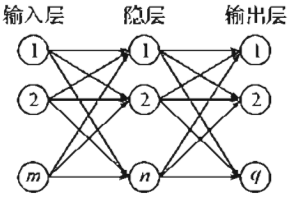

图 2 前向神经网络

Fig. 2 Forward neural network

与上面几节线性方法不同的是，神经网络属于非线性统计数据建模(nonlinear statistical data modeling)，其隐层暗含了“特征提取”的思想，且可视作输入数据在一种“自适应的非线性非正交的基”上的映射. 同样地，此时基牺牲了正交性、线性、不变性，增加了计算负担，但换来了更加强大的灵活性[23].

简而言之，前向神经网络采用了类似基追踪的方法[24]，但基是自适应变化的、非线性的，因此更加灵活. 前向神经网络与基追踪相似之处在于，两者的基都不是正交的，都是根据给定数据而自适应选取的最佳基. 前向神经网络的优势在于无不需预选字典，字典在算法中自动生成，并可作为特征选择的一种方法.

10 径向基函数网络

首先观察径向基函数(RBF)神经元如图 3 所示.

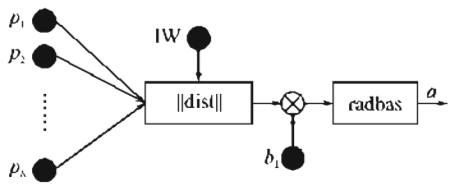

图 3 RBF 神经元图

Fig. 3 Neuron of RBF

图中输入向量 $\boldsymbol{p}$ 的维数为 R，首先 $\boldsymbol{p}$ 与输入层权值矩阵 $\boldsymbol{IW}$ 相减，然后求距离函数 dist，再与偏置 b_1 相乘，最后求径向基函数 radbas(n) = $\exp(-n^2)$，得到神经元的输出为

$$\boldsymbol{a} = \text{radbas}(\| \boldsymbol{IW} - \boldsymbol{p} \| b_1) \tag{41}$$

整个 RBF 网络由两层神经元组成，第 1 层为 S_1 个如图 3 所示的 RBF 神经元，第 2 层为 S_2 个线性神经元，如图 4 所示. 在第 2 层开始时，第 1 层的输出 $\boldsymbol{a}$ 首先经过线性层权值矩阵 $\boldsymbol{LW}$ 后与偏置 b_2 相加，再通过一个纯线性(purelin)函数 purelin(n) = n，得到网络输出 y 为

$$y = \text{purelin}(\boldsymbol{LW} \times \boldsymbol{a} + b_2) \tag{42}$$

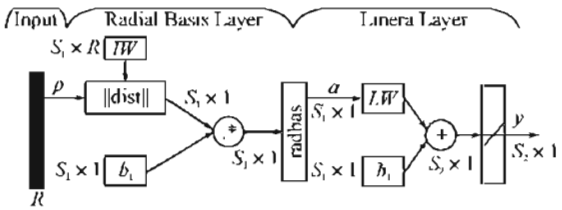

图 4　RBF 神经网络结构图

Fig. 4　Structure of RNN

比较式(41)与式(9)可见，RBF 网络与核方法非常类似，不同之处在于 RBF 网络的 $\boldsymbol{LW}$ 需要通过求解一个方程组，而核方法的权重是直接通过归一化计算求得，因此 RBF 网络预测结果更为逼近完全内插函数估计(注意不是未知函数 r)，而核方法计算更为简便[25].

11　维数灾难

将函数估计推广到高维，则会碰到维数诅咒(curse of dimensionality)[26](图 5)，它意味着当观测值的维数增加时，估计难度会迅速增大. 维数诅咒有两层含义：

一是计算的维数诅咒，指的是某些算法的计算量随着维数的增长而成指数增加. 解决方法通常采用优化算法，例如遗传算法、粒子群算法、蚁群算法等[27].

二是样本的维数诅咒，指的是数据维数为 d 时，样本量需要随着 d 指数增长. 在函数估计中，第二层含义更为重要，这里给予详细解释.

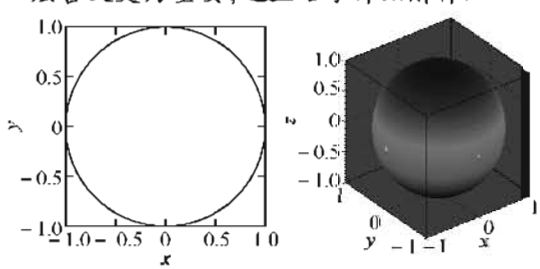

图 5　样本的维数诅咒示意图

Fig. 5　Dimensionality curse of samples

假设一个半径 r 维数为 d 的超球，被一个边长为 $2r$ 维数为 d 的超立方体所包围，假设超立方体内存在一个均匀分布的点，则由于超球的体积为 $2r^d\pi^{d/2}/[d\Gamma(d/2)]$，超立方体的体积为 $(2r)^d$，因此该点同时也落在超球内的概率 P 为

$$P = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} \tag{43}$$

令维数 d 由 2 逐步增长到 20，则对应的概率 P 如图 6 所示. 显然，当 $d = 20$ 时，P 仅为 2.46×10^{-8}. 因此，若在 2 维空间中 1 个样本在半径 r 的意义下能逼近一个正方形，则在 20 维空间内，则需要 $1/2.46 \times 10^{-8} = 4.06 \times 10^7$ 个样本才能在半径 r 的意义下逼近超立方体.

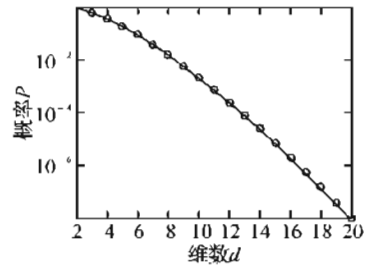

图 6　概率 P 与维数 d 的关系

Fig. 6　The curve of probability P against dimensionality d

因此，在高维问题中，由于数据非常稀少，导致局部邻域中包含极少的数据点[28]，因此估计变得异常困难. 目前还没有较好的办法解决.

12　结　语

将文中阐述的方法归结并示于图 7.

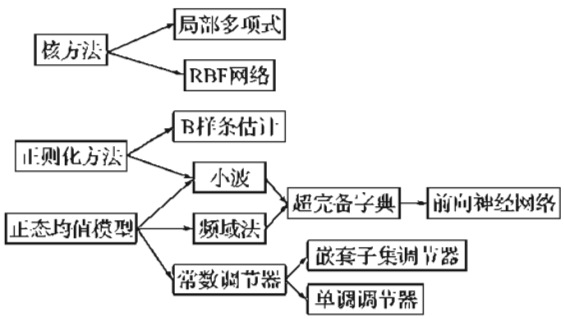

图 7　非参数回归方法

Fig. 7　Survey of non-parametric regression methods

不同类型方法的特点总结如下：

a. 核方法、正则化方法、正态均值模型可以视作最基本最原始的方式. 另外，正则化方法与正态均值模型可视作一类特殊的核方法.

b. 核方法、局部多项式方法、正则化方法、正态均值模型、小波等方法在大多数情况下均非常类似. 这些方法都包含了一个偏倚-方差平衡，所以都需要选择一个光滑参数. 由于这些方法均是线

性光滑器，所以均可以采用第4节中基于CV、GCV的方法.

c. 小波方法一般面向空间非齐次函数. 如果需要一个精确的函数估计，而且噪声水平较低，则小波方法非常有效. 但若面对一个标准的非参数回归问题，而且感兴趣于置信集，则小波方法并不比其它方法明显更好.

d. 超完备字典缺陷是丧失了基的正交性，因此估计系数变得复杂；优点是更为灵活，能够采用稀疏的系数描述复杂函数.

e. 前向神经网络与RBF神经网络是基于不同的模型独立推导出来的，二者不可混淆. 另外，神经网络方法的缺点是一般不考虑置信带，并常用训练误差代替风险函数，容易过拟合；优点是面向应用、思想简单且设计灵活.

f. 理论上，这些方法没有大的差别，特别在用置信带的宽度来评价时. 每种方法都有其拥护者与批评者，没有哪一种方法目前获得应用上的优势. 一种解决方案是对每个问题都利用所有可行的方法，如果结果一致，则选择简单者；如果结果不一致，则必须探讨内在的原因.

g. 所讨论的方法能够用于高维问题，然而，即使通过智能优化算法解决了计算的维数诅咒，仍然面对样本的维数诅咒. 计算一个高维估计相对容易，然而该估计将不如一维情况下那么精确，其置信区间会非常大. 但这并不表示方法失效，而是表示问题的固有困难.

参考文献：

[1] Neumeyer N. A note on uniform consistency of monotone function estimators [J]. Statistics & Probability Letters, 2007, 77(7): 693 - 703

[2] Sheena Y, Gupta A K. New estimator for functions of the canonical correlation coefficients [J]. Journal of Statistical Planning and Inference, 2005, 131(1): 41 - 61.

[3] 张煜东，吴乐南，李铜川，等. 基于PCNN的彩色图像直方图均衡化增强[J]. 东南大学学报，2010，40(1)：64 - 68.

[4] 詹锦华. 基于优化灰色模型的农村居民消费结构预测[J]. 武汉工程大学学报，2009，31(9)：89 - 91.

[5] Wasserman L. All of Nonparametric Statistics [M]. New York: Springer-Verlag, Inc.

[6] 张煜东，吴乐南，吴含前. 工程优化问题中神经网络与进化算法的比较[J]. 计算机工程与应用，2009，45(3)：1 - 6.

[7] Hansen C B. Asymptotic properties of a robust variance matrix estimator for panel data when T is large [J]. Journal of Econometrics, 2007, 141(2): 597 - 620.

[8] Pokharel P P, Liu W F, Principe J C. Kernel least mean square algorithm with constrained growth [J]. Signal Processing, 2009, 89(3): 257 - 265.

[9] Kalivas J H. Cyclic subspace regression with analysis of the hat matrix [J]. Chemometrics and Intelligent Laboratory Systems, 1999, 45(1): 215 - 224.

[10] 张煜东，吴乐南. 基于二维Tsallis熵的改进PCNN图像分割[J]. 东南大学学报：自然科学版，2008，38(4)：579 - 584

[11] Geçkinli N C, Yavuz D. A set of optimal discrete linear smoothers [J]. Signal Processing, 2001, 3(1): 49 - 62.

[12] Antoniotti M, Carreras M, Farinaccio A, et al. An application of kernel methods to gene cluster temporal meta-analysis [J]. Computers & Operations Research, 2010, 37(8): 1361 - 1368.

[13] Hsieh P F, Chou P W, Chuang H Y. An MRF-based kernel method for nonlinear feature extraction [J]. Image and Vision Computing, 2010, 28(3): 502 - 517.

[14] Katkovnik V. Multiresolution local polynomial regression: A new approach to pointwise spatial adaptation [J]. Digital Signal Processing, 2005, 15(1): 73 - 116.

[15] Baíllo A, Grané A. Local linear regression for functional predictor and scalar response [J]. Journal of Multivariate Analysis, 2009, 100(1): 102 - 111.

[16] Zhang J W, Krause F L. Extending cubic uniform B-splines by unified trigonometric and hyperbolic basis [J]. Graphical Models, 2005, 67(2): 100 - 119.

[17] 张煜东，吴乐南，韦耿，等. 用于多指数拟合的一种混沌免疫粒子群优化[J]. 东南大学学报，2009，39(4)：678 - 683.

[18] Chaudhuri S, Perlman M D. Consistent estimation of the minimum normal mean under the tree-order restriction [J]. Journal of Statistical Planning and Inference, 2007, 137(11): 3317 - 3335.

[19] Labat D. Recent advances in wavelet analyses: Part 1. A review of concepts [J]. Journal of Hydrology, 2005, 314(1): 275 - 288.

[20] Kunoth A. Adaptive Wavelets for Sparse Representations of Scattered Data [J]. Studies in Computational Mathematics, 2006, 12: 85 - 108.

[21] Donoho D L, Elad M. On the stability of the basis pursuit in the presence of noise [J]. Signal Processing, 2006, 86(3): 511 - 532.

[22] Malgouyres F. Rank related properties for Basis Pursuit and total variation regularization [J]. Signal Processing, 2007, 87(11): 2695 - 2707.

[23] 张煜东,吴乐南,韦耿. 神经网络泛化增强技术研究[J]. 科学技术与工程,2009,9(17):4997－5002.
[24] 屠艳平,管昌生,谭浩. 基于BP网络的钢筋混凝土结构时变可靠度[J]. 武汉工程大学学报,2008,30(3):36－39.
[25] Zhang Y D, Wu L N, Neggaz N, et al. Remote-sensing Image Classification Based on an Improved Probabilistic Neural Network[J]. Sensors, 2009, 9: 7516－7539.
[26] Aleksandrowicz C, Barequet C. Counting polycubes without the dimensionality curse [J]. Discrete Mathematics, 2009, 309(13): 4576－4583.
[27] 张煜东,吴乐南,奚吉,等. 进化计算研究现状(上)[J]. 电脑开发与应用,2009,22(12):1－5.
[28] 王忠,叶雄飞. 遗传算法在数字水印技术中的应用[J]. 武汉工程大学学报,2008,30(1):95－97.

Survey of non-parametric estimation methods

ZHANG Yu-dong[1,2], *YAN Jun*[1], *WANG Shui-hua*[1], *WU Le-nan*[1]
(1. School of Information Science & Engineering, Southeast University, Nanjing 210096, China;
2. Brainimaging Lab., Department of Psychology, Columbia University, New York NY 10032, USA)

Abstract: In order to solve the problem of function estimation, we first discuss traditional parametric regression method. Since it needs a priori knowledge to determine the model, the parametric method is not robust and is model-sensitive. Thus, data-driven non-parametric method is introduced, which needs not any a prior knowledge to estimate the unknown function. Eight major non-parametric methods are discussed as kernel method, local polynomial regression, regularization method, normal mean model, wavelet method, overcomplete dictionary, forward neural network, and radial basis function network. These algorithms are compared, and their coherence and inheritance are investigated. Finally, generalize the algorithms to high dimensionality and point out two problems as curse of dimensionality of computation and sample. The former can be settled down by intelligent methods while the latter is problem intrinsic.

Key words: parametric statistics; non-parametric statistics; kernel method; local polynomial regression; regularization method; normal mean model; wavelet; over-complete dictionary; forward neural network; radial basis function network

本文编辑:龚晓宁