

基于假设函数空间上样本误差的估计

王 兵, 曾文曲, 陈鹤峰

(广东工业大学 应用数学学院, 广东 广州 510090)

摘要: 在传统回归分析中, 基于已知样本寻找最优函数一般在某个已知函数空间, 如多项式空间, 对数以及指数空间等, 对于样本个数一定的情况下, 样本误差会随着函数空间的增大而增大. 在本文中, 作者用概率测度、覆盖原理在有界函数空间中给出一个样本误差估计的方法.

关键词: 样本误差; 覆盖原理; 有界函数空间; 概率测度

中图分类号: O212.2

文献标识码: A

文章编号: 1007-7162(2006)04-0098-03

1 主要定义

定义 1 设 X 是一紧矩阵空间, $Y=R$ 是一实数空间, $z = \{(x_i, y_i)\}_{i=1}^m$ 是一样本, ρ 是在 $Z = X \times Y$ 上的概率分布, ρ_X 为 X 上的边缘分布, $\rho(y|x)$ 为条件分布, 记 $f_\rho(x) = \int_Y y d\rho(y|x)$.

定义 2 设 X 是一紧矩阵空间, $Y=R$ 是一实数空间. H 为 $\ell(X)$ 的子空间, $f: X \rightarrow Y, f_H = \left\{ f \mid \min_{f \in H} \int_Z (f(x) - y)^2 d\rho \right\}, f_z = \left\{ f \mid \min_{f \in H} \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2 \right\},$

记: $\varepsilon(f) = \int_Z (f(x) - y)^2 d\rho, \varepsilon(f_\rho) = \int_Z (f_\rho(x) - y)^2 d\rho, \varepsilon(f) = \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2,$

定义样本误差: $\varepsilon_H(f_z) = \varepsilon(f_z) - \varepsilon(f_H).$

定义 3 设 S 是一矩阵空间, $\forall \gamma > 0$ 定义 $N(S, \gamma)$ 为半径为 γ 且在 S 中能覆盖 S 的最小个数, 当 S 是紧的时, 则这个数是有限的.

定义 4 设 $M > 0$ ρ 是在 $Z = X \times Y$ 上的概率测度, 称函数集 H 是从 X 到 R 有界的, 如果满足: 对于所有属于 H 的 f 都有 $|f(x) - y| \leq M$.

2 主要定理

定理 1 [Bennett] 设 $\{\xi_i\}_{i=1}^m$ 是概率空间 Z 上的相互独立的随机变量, 且数学期望 $E(\xi_i) = \mu_i$, 方差 $D(\xi_i) = \sigma_i^2$. 记 $A^2 = \sum_{i=1}^m \sigma_i^2$, 如果对于每一个 i 都有 $|\xi_i - \mu_i| \leq M$ 几乎处处成立, 则对于任意的正数 $\varepsilon > 0$ 有

$$P\left\{\sum_{i=1}^m |\xi_i - \mu_i| \geq \varepsilon\right\} \leq \exp\left\{-\frac{\varepsilon}{M} \left(1 + \frac{A^2}{M\varepsilon}\right) \log\left(1 + \frac{M\varepsilon}{A^2}\right) - 1\right\}.$$

本定理及其证明见参考文献 [1].

定理 2^[2] 设 H 是 $\ell(X)$ 的紧 M 有界子集, 则对于任意正数 $\varepsilon > 0$ 有

收稿日期: 2006-03-27

作者简介: 王兵 (1978-), 男, 在读研究生, 主要研究方向为概率统计, 分形几何理论.

$$P\left\{\sup_{f \in H}(\varepsilon(f) - \varepsilon(f_z)) \leq \varepsilon\right\} \geq 1 - N\left(H, \frac{\varepsilon}{8M}\right) \exp\left\{-\frac{m\varepsilon^2}{8M^4}\right\}.$$

引理 1 让 $H = S_1 \cup S_2 \cup \dots \cup S_b$ 且 $\varepsilon > 0$ 则有如下结论

$$P\left\{\sup_{f \in H}(\varepsilon(f) - \varepsilon(f_z)) \geq \varepsilon\right\} \leq \sum_{j=1}^l P\left\{\sup_{f \in H}(\varepsilon(f) - \varepsilon(f_z)) \geq \varepsilon\right\}.$$

证明: 记 $L_z(f) = \varepsilon(f) - \varepsilon_z(f)$, 则有: $\sup_{f \in H} L_z(f) \geq \varepsilon \Leftrightarrow \exists j \leq l$ s.t. $\sup_{f \in H} L_z(f) \geq \varepsilon$ 由事件和的概率小于等于事件概率的和可得^[4].

引理 2^[1] 如果 $|f_j(x) - y| \leq M$, $j = 1, 2$ 则 $|L_z(f_1) - L_z(f_2)| \leq 4M \|f_1 - f_2\|_\infty$.

现在来证明定理 2

证明: 记 $l = N\left(H, \frac{\varepsilon}{4M}\right)$, 考虑 $f_1, f_2, \dots, f_b$ 使得 D_j 为以 f_j 为中心, 以 $\frac{\varepsilon}{4M}$ 为半径覆盖 H .

由定义 4 $|f(x) - y| \leq M$, $|L_z(f) - L_z(f_j)| \leq 4M \|f - f_j\| \leq 4M \frac{\varepsilon}{4M} = \varepsilon$ 对于所有的 $f_j \in D_j$ 都成立, 即 $\sup_{f \in D_j} L_z(f) \geq 2\varepsilon \Rightarrow L_z(f_j) \geq \varepsilon$

对于所有 $j = 1, \dots, l$

$$P\left\{\sup_{f \in D_j} L_z(f) \geq 2\varepsilon\right\} \leq P\left\{L_z(f_j) \geq 2\varepsilon\right\} \leq \exp\left\{-\frac{m\varepsilon^2}{2M^4}\right\},$$

由定理 1 [Bennett] 可得: $P\left\{\frac{1}{m} \sum_{i=1}^m \xi_i(z_i) - \mu\right\} \leq \exp\left\{-\frac{m\varepsilon^2}{2M^2}\right\}$, 其中 ξ 为 Z 上的随机变量, 满足 $E(\xi) = \mu$, $|\xi(z) - \mu| \leq M$, 在引理 1 中, 用 $\frac{\varepsilon}{2}$ 代替 ε 即可.

3 样本误差 $\varepsilon_H(f_z)$ 的估计

现在考虑 f_z 和 f_H 的接近程度, 也就是 $\varepsilon_H(f_z)$ 的最优估计. 主要结论如下

定理 3 设 H 为 $\ell(X)$ 上的 M -有界紧子集, 则, 对于任意的 $\varepsilon > 0$ 有

$$P\left\{\varepsilon_H(f_z) \leq \varepsilon\right\} \geq 1 - \left[N\left(H, \frac{\varepsilon}{16M}\right) + 1\right] \exp\left\{-\frac{m\varepsilon^2}{32M^4}\right\}.$$

引理 3^[1] 设 $M > 0$ 且 $f: X \rightarrow Y$ 是 M -有界, 则对于任意的 $\varepsilon > 0$ 有

$$P\left\{L_z(f) \geq -\varepsilon\right\} \geq 1 - \exp\left\{-\frac{m\varepsilon^2}{2M^4}\right\}.$$

定理 3 的证明:

由定义 2 可以看到

$$\varepsilon_H(f_z) = \varepsilon(f_z) - \varepsilon(f_H),$$

$$\varepsilon_H(f_z) = \varepsilon(f_z) - \varepsilon_z(f_z) + \varepsilon_z(f_z) - \varepsilon_z(f_H) + \varepsilon_z(f_H) - \varepsilon(f_H),$$

由于 $\varepsilon_z(f_z) - \varepsilon_z(f_H) \leq 0$ 即得

$$\varepsilon_H(f_z) \leq [\varepsilon(f_z) - \varepsilon_z(f_z)] + [\varepsilon_z(f_H) - \varepsilon(f_H)].$$

在引理 2 中, 对于函数 f_H , 用 $\frac{\varepsilon}{2}$ 代替 ε 即考虑 $\varepsilon_z(f_H) - \varepsilon(f_H) \leq \frac{\varepsilon}{3}$ 的概率为

$$P\left\{\varepsilon_z(f_H) - \varepsilon(f_H) \leq \frac{\varepsilon}{2}\right\} \geq 1 - \exp\left\{-\frac{m\varepsilon^2}{8M^4}\right\}.$$

同样, 在定理 2 中, 用 $\frac{\varepsilon}{2}$ 代替 ε 得至少以 $1 - N(H, \frac{\varepsilon}{16M}) \exp\left\{-\frac{m\varepsilon^2}{32M^4}\right\}$ 的概率使得下式成立,

$$\sup_{f \in H} L_z(f) = \sup_{f \in H} \{\varepsilon(f) - \varepsilon_z(f)\} \leq \frac{\varepsilon}{2}, \text{ 即: } \varepsilon(f_z) - \varepsilon_z(f_z) \leq \frac{\varepsilon}{2}.$$

由上两个界的估计, 可得: $\varepsilon_H(f_z)$ 至少以

$$\left[1 - N(H, \frac{\varepsilon}{16M}) \exp\left\{-\frac{m\varepsilon^2}{32M^4}\right\}\right] \left[1 - \exp\left\{-\frac{m\varepsilon^2}{8M^4}\right\}\right] \geq 1 - \left[N(H, \frac{\varepsilon}{16M}) + 1\right] \exp\left\{-\frac{m\varepsilon^2}{32M^4}\right\}$$

的概率成立, 这即为样本误差的一个有效估计.

4 定理 3 的应用

对于任意给定的 $\varepsilon > 0$ 和置信区间 $\delta > 0$ 可以用定理 3 给出样本个数所满足的条件.

由 $P\left\{\varepsilon_H(f_z) \leq \varepsilon\right\} \geq 1 - \delta$ 可得样本个数 m 需满足:

$$m \geq \frac{32M^4}{\varepsilon^2} \left\{ \ln(1 + N(H, \frac{\varepsilon}{16M})) + \ln(\frac{1}{\delta}) \right\},$$

只要取 $\delta = \left\{N(H, \frac{\varepsilon}{16M}) + 1\right\} \exp\left\{-\frac{m\varepsilon^2}{32M^4}\right\}$, 解得 m , 即得样本个数 m 所满足的条件.

5 结束语

作者运用样本空间分布上的概率测度以及函数空间的覆盖原理, 给出了样本误差的一个较好估计. 值得一提的是, 函数空间分解中的覆盖数可以通过一种叫做再生核希尔伯特空间得到. 读者可以参阅参考文献 [2-3].

参考文献:

- [1] T Zhang Statistical behavior and methods based on convex risk minimization[J]. Ann Stat 2004, 32: 56-85.
- [2] F Cucker, S Smale On the mathematical functions of learning[J]. Bull Amer Math Soc, 2001(39): 1-49.
- [3] M Anthony, P L Bartlett Network Learning: Theoretical Foundations[M]. London: Cambridge University press, 1999.
- [4] 汪荣鑫. 数理统计[M]. 西安: 西安交通大学出版社, 1986.

Estimation of Sample Error Based on Hypothesis Space

WANG Bing ZENG Wen-qu CHEN He-feng

(Faculty of Applied Mathematics, Guangdong University of Technology, Guangzhou 510090, China)

Abstract In traditional regress analysis, a good regression function is found in some function space based on known samples, such as polynomial function space, logarithm space and exponent space. When the sample number is certain, the sample error will be big with the widening of function space. In this paper, a method of sample error estimation is presented by probability measure and covering theory in a bounded function space.

Key words sample error; covering theory; bounded function space; probability measure