

广义加性模型简介

蔡苗*

2019-07-01

摘要

广义加性模型是可以假设自变量与因变量或者因变量的函数之间为非线性关系的统计模型。

目录

1 引言	2
1.1 什么是广义加性模型 (GAM)	2
1.2 为什么我们要学会 GAM	2
1.3 实现 GAM 的软件和包	2
2 基本概念	2
2.1 基底函数	2
2.2 有效自由度	2
3 GAM 的统计理论	2
3.1 平滑化	2
3.2 GAM 的估计	3
3.3 选择平滑参数的方法	3
3.4 变量选择	3
4 实际例子	3
参考文献	4

*圣路易斯大学公共卫生学院流行病与生物统计系。电子邮件: miao.cai@slu.edu

1 引言

1.1 什么是广义加性模型 (GAM)

广义加性模型 (Generalized additive Model, GAM) 是一类不假设预测变量和结果变量之间的关系是线性的模型。GAM 最开始由 Hastie and Tibshirani (1986) 提出。

本稿件主要参考 Larsen (2015)。

1.2 为什么我们要学会 GAM

- 可解读性
- 灵活性
- 自动选择参数
- 正则化

1.3 实现 GAM 的软件和包

大部分的统计软件均可以实现 GAM, 如SAS和R。在 SAS 中可以使用 PROC GAM 或者 PROC TPSPLINE 来实现, 在 R 中可以使用 Wood (2017) 开发的 mgcv 包和由 Hastie (2018) 开发的 gam 来实现。相较于 SAS, R 的 GAM 扩展包的功能和灵活性都更强, 因此我们比较推荐使用 R 的 mgcv 包来实现 GAM。

2 基本概念

2.1 基底函数

基底函数 (basis function)。

2.2 有效自由度

有效自由度 (Effective Degrees of Freedom, EDF)。

3 GAM 的统计理论

3.1 平滑化

- 局部回归 (LOESS)
- 平滑样条 (smoothing splines)

- 回归样条

3.2 GAM 的估计

- 局部算分算法
- 带惩罚项的迭代重新加权最小二乘法 (Penalized iterative reweighted least squares, **PIRLS**)

带惩罚项的似然函数

$$2 \log L(\alpha, s_1(x_1), \dots, s_k(x_k)) - \sum_{k=1}^K \lambda_k U_k$$

其中 $\log L(\alpha, s_1(x_1), \dots, s_k(x_k))$ 为对数似然函数, $\sum_{k=1}^K \lambda_k U_k$ 为惩罚项

3.3 选择平滑参数的方法

- 广义交叉验证 (Generalized cross validation criterion, **CGV**)
- 受限极大似然估计 (Restricted Maximum Likelihood, **REML**)

3.4 变量选择

3.4.1 单变量选择

- 信息值 (Information value, IV)
- 证据权重 (Weight of evidence, WOE)

3.4.2 多变量选择

- 分步选择法 (向前和向后法)
- 收缩估计法 (shrinkage)

4 实际例子

参考文献

Hastie, Trevor. 2018. *Gam: Generalized Additive Models*. <https://CRAN.R-project.org/package=gam>.

Hastie, T, and R Tibshirani. 1986. “Generalized Additive Models Statistical Science.”

Larsen, Kim. 2015. “GAM: The Predictive Modeling Silver Bullet.” *Multithreaded. Stitch Fix* 30.

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with R*. Chapman; Hall/CRC.