Statistics and Application 统计学与应用, 2019, 8(1), 26-30
Published Online February 2019 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2019.81004

Hans汉斯

Application of Cross-Validation in Model Selection

—Take OLS and RR as Examples

Yanshan Cao

Yunnan University of Finance and Economics, Kunming Yunnan
Email: 547382297@qq.com

Received: Dec. 26th, 2018; accepted: Jan. 9th, 2019; published: Jan. 16th, 2019

Abstract

This paper reviews the origin and development of cross-validation, and summarizes the previous research results. On this basis, leave-one-out cross-validation is used to solve some problems for model selection. OLS and RR were used to analyze the reaction of acetylene data, establishing appropriate models and selecting the optimal model. At the same time, the rationality and reality of the model selection were discussed.

Keywords

Cross-Validation, OLS, RR, Model Selection

交叉验证法在模型选择中的应用

——以 OLS 和 RR 为例

曹延姗

云南财经大学，云南 昆明
Email: 547382297@qq.com

收稿日期：2018年12月26日；录用日期：2019年1月9日；发布日期：2019年1月16日

摘　要

文章回顾了交叉验证法的起源及发展，对已有研究成果进行总结和归纳。在此基础上，选择其中较为常用的留一交叉验证法对实例进行模型选择。分别采用最小二乘估计法和岭回归法对乙炔的反应数据进行

文章引用: 曹延姗. 交叉验证法在模型选择中的应用[J]. 统计学与应用, 2019, 8(1): 26-30.
DOI: 10.12677/sa.2019.81004

分析，建立相应的模型，并对所建立的模型进行选择，同时探讨选择的合理性及现实性。

关键词

交叉验证，最小二乘估计，岭回归，模型选择

1. 交叉验证法概述

交叉验证是一种模型选择方法，在使用这种方法时，不需要做任何的假定，加之操作简便，故其具有更广泛的适应性，应用也较为普遍。

早先，人们在使用统计模型解决实际问题时，同一数据集既被用来进行模型训练，又被用来进行预测误差估计。但上世纪 30 年代，Larson (1931) [1]发现采用既做训练集又做预测估计往往会使得估计结果偏向于乐观。为了克服这个问题，Stone (1974) [2]提出了交叉验证的方法，并证实用一部分数据做训练集，余下的数据用来进行误差估计，得到的估计结果更为优良。在这之后，各国学者纷纷提出了交叉验证的多种方法，交叉验证有了巨大的发展。

首先受到大家关注的是留一交叉验证法，由 Stone，Allen，Geisser (1974) [3]等人提出，它的基本思想是每次从个数为 n 的数据中取出一个数据作为测试集，而将剩下的 $n-1$ 个数据作为训练集，这一操作重复进行 n 次，而这 n 次结果的平均值就是预测误差的估计值。这是交叉验证法中最常见的一种方法。Geisser (1975) [4]提出了 V 折交叉验证法，这种方法首先把数据集平均分成 V 份，然后每次选择这 V 份中的一份作为测试集，其他的 V-1 份作为训练集，重复进行 V 次这样的操作，再把 V 次结果的平均值作为预测误差的估计。Devroye 和 Wagner (1979) [5]提出了 hold-out 法，其主要思想是将数据集进行一次切分，一部分做训练模型，另一部分做测试估计误差。这是最为简单的一种交叉验证法。Shao (1993) [6]提出了留 P 交叉验证法，它与留一交叉验证法较为类似，其基本思想是每次从 n 个数据中随机取出 P 个作为测试集，将余下的 $n-p$ 个数据作为训练集，这一操作重复进行 C_n^p 次，最终所得的平均值即为预测误差的估计。此外，Shao 还提出了平衡不完全交叉验证，可以将它看成留 P 交叉验证法的变异形式，它可以克服当 P 较大时计算比较复杂的不足，其基础是建立在平衡不完全集上的，从而可以保证每个数据在训练集和测试集中都具有相同的地位。Dietterich (1998) [7]提出 5*2 交叉验证法。其主要思想是将数据集 V 平均分成 $V_1^{(1)}$和 $V_1^{(2)}$两部分，先用 $V_1^{(1)}$做训练集，$V_1^{(2)}$做测试集，然后再交换，$V_1^{(2)}$做训练集，$V_1^{(1)}$做测试集，这样所得即为第一折。接下来将数据集 V 重新打乱，平均分为新的 $V_2^{(1)}$和 $V_2^{(2)}$两部分，用 $V_2^{(1)}$做训练集，$V_2^{(2)}$做测试集，再交换，$V_2^{(2)}$做训练集，$V_2^{(1)}$做测试集，所得为第二折。将这种操作重复进行 5 次，就可以得到 10 个训练集和测试集。Dietterich 特别指出，如果超过 5 次对折，各个集合之间共享的样本会过多，计算出来的预测误差估计极易相互依赖，无法增加新的信息，而在五折之前，这种数据共用的问题尚可容忍。

2. 数据分析

该数据源自一组乙炔的反应数据，总共有 16 个观测值。其中，响应变量向量 y 是正庚烷(n-heptane)转化为乙炔(acetylene)的转化百分比，自变量 x_1 是反应釜的温度(摄氏)，x_2 是氢气 - 乙炔转化百分比，x_3

是接触时间(单位是秒)，如表 1 所示。

Table 1. Data of acetylene reaction
表 1. 乙炔的反应数据表

序号	x_1 (℃)	x_2 (%)	x_3 (s)	y (%)
1	1300	7.5	0.0120	49.0
2	1300	9.0	0.0120	50.2
3	1300	11.0	0.0115	50.5
4	1300	13.5	0.0130	48.5
5	1300	17.0	0.0135	47.5
6	1300	23.0	0.0120	44.5
7	1200	5.3	0.0400	28.0
8	1200	7.5	0.0380	31.5
9	1200	11.0	0.0320	34.5
10	1200	13.5	0.0260	35.0
11	1200	17.0	0.0340	38.0
12	1200	23.0	0.0410	38.5
13	1100	5.3	0.0840	15.0
14	1100	7.5	0.0980	17.0
15	1100	11.0	0.0920	20.5
16	1100	17.0	0.0860	29.5

考虑到各变量单位差异较大，为了同等对待每一变量，在分析前，先对各变量作标准化变换，标准化后的变量分别记为 zy、zx_1、zx_2、zx_3 (MATLAB 程序中为书写方便用 Y、X1、X2、X3 表示)。

令 $A=\left(x_1 \quad x_2 \quad x_3\right)$，利用 MATLAB 求得 x_1、x_2、x_3 的相关关系矩阵为：

$$Cov(A)=\begin{pmatrix} 1 & 0.2236 & -0.9582 \\ 0.2236 & 1 & -0.2402 \\ -0.9582 & -0.2402 & 1 \end{pmatrix}$$

从相关关系矩阵可以看出，x_1 和 x_3 之间的相关系数较大，接近于 1，故我们认为 x_1 与 x_3 之间有线性相关关系。

3. 模型分析

考虑模型 2^3-1 个备选模型，其中的第 s 个模型为：

$$y_i=\beta_0+\sum_{s\in S}x_{i,s}\beta_s+\varepsilon_i, i=1,\cdots,16 \tag{1}$$

其中脚标集 S 取遍 $\{1,2,3\}$ 的所有可能的非空子集。

将上述模型中的变量表示成矩阵的形式：

$$y=X\beta+\varepsilon \tag{2}$$

其中 $y=\left(zy_1,\cdots,zy_{16}\right)'$，$\varepsilon=\left(\varepsilon_1,\cdots,\varepsilon_{16}\right)'$，$X$ 和 β 因模型不同相应有所变化，将在下文中进行详细的讨论。

(一) OLS 法模型估计

普通最小二乘法(OLS)是由德国数学家高斯最早提出和使用的。由于在一定的假设条件下，最小二乘估计量有着非常好的统计性质，从而使它成为回归分析中最有功效和最为流行的方法。

对于线性模型 $y=X\beta+\varepsilon$，OLS 法的思想是，β 的真值应该使误差向量 $e=y-X\beta$ 达到最小，也就是它的长度平方 $Q(\beta)=\|e\|^2=\|y-X\beta\|^2=(y-X\beta)'(y-X\beta)$ 达到最小，求解得到最小二乘估计 $\hat{\beta}=(X'X)^{-1}X'y$。

(二) RR 法模型估计

岭回归(RR)是一种专用于共线性数据分析的有偏估计回归方法，实质上是对最小二乘估计法的一种改良，不再做无偏性的要求。它以损失部分信息、降低精度为代价，从而使得回归系数更符合实际、更可靠。

对于线性模型 $y=X\beta+\varepsilon$，β 的岭回归估计相对于最小二乘估计，在自变量信息矩阵的主对角线元素上人为地加入了一个非负因子，即 $\hat{\beta}=(X'X+kI)^{-1}X'y$，其中岭参数 k 使用 Hoerl & Kennard (1970) [8] 中的方法确定。

由于使用正交验证法，故用训练集进行参数估计，用测试集进行预测误差估计。上述两种方法各七种模型形式如表 2 所示。

Table 2. Seven models for model estimation with OLS and RR
表 2. OLS 法和 RR 法模型估计的七种模型表

自变量	y	X	β	ε
zx_1	$(zy_1,\cdots,zy_{16})'$	$(1 \quad zx_1)$	$(\beta_0 \quad \beta_1)'$	$(\varepsilon_1,\cdots,\varepsilon_{16})'$
zx_2	$(zy_1,\cdots,zy_{16})'$	$(1 \quad zx_2)$	$(\beta_0 \quad \beta_2)'$	$(\varepsilon_1,\cdots,\varepsilon_{16})'$
zx_3	$(zy_1,\cdots,zy_{16})'$	$(1 \quad zx_3)$	$(\beta_0 \quad \beta_3)'$	$(\varepsilon_1,\cdots,\varepsilon_{16})'$
zx_1、zx_2	$(zy_1,\cdots,zy_{16})'$	$(1 \quad zx_1 \quad zx_2)$	$(\beta_0 \quad \beta_1 \quad \beta_2)'$	$(\varepsilon_1,\cdots,\varepsilon_{16})'$
zx_1、zx_3	$(zy_1,\cdots,zy_{16})'$	$(1 \quad zx_1 \quad zx_3)$	$(\beta_0 \quad \beta_1 \quad \beta_3)'$	$(\varepsilon_1,\cdots,\varepsilon_{16})'$
zx_2、zx_3	$(zy_1,\cdots,zy_{16})'$	$(1 \quad zx_2 \quad zx_3)$	$(\beta_0 \quad \beta_2 \quad \beta_3)'$	$(\varepsilon_1,\cdots,\varepsilon_{16})'$
zx_1、zx_2、zx_3	$(zy_1,\cdots,zy_{16})'$	$(1 \quad zx_1 \quad zx_2 \quad zx_3)$	$(\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3)'$	$(\varepsilon_1,\cdots,\varepsilon_{16})'$

(三) 留一交叉验证法模型选择[9]

在前文中已经简单介绍过，在使用留一交叉验证法时，每次从 n 个数据中取出一个数据作为测试集，而将其他 $n-1$ 个数据作为训练集，这一操作重复进行 n 次，而这 n 次结果的平均值就是预测误差的估计值。使用 MATLAB 进行交叉验证，得到上述 14 个模型的预测误差如表 3 所示。

Table 3. Prediction error of models with OLS and RR
表 3. OLS 法和 RR 法估计的模型的预测误差表

方法 / 自变量	OLS	RR
zx_1	0.1380	0.1391
zx_2	1.0164	1.0043
zx_3	0.2037	0.2047
zx_1、zx_2	0.1386	0.1393
zx_1、zx_3	0.1457	0.1441
zx_2、zx_3	0.2040	0.2025
zx_1、zx_2、zx_3	0.1485	0.1478

如果只根据预测误差的最小值来判定，在上述 14 个模型中应当选择只含变量 $x_1(zx_1)$的最小二乘估计模型，但考虑到 x_1、x_2、x_3 对 y 均有影响，只用一个变量可能不能很好地反映 y 受到影响的各个方面。根据前面进行的变量间相关性的分析，我们知道 x_1 与 x_3 之间存在线性关系，当这两个变量同时出现在模型中时，根据预测误差值可以明显地看出岭回归法估计的模型优于最小二乘法估计的模型。故如果我们希望 x_1、x_2、x_3 三个变量均出现在模型中，则应选择 RR 法估计的模型。如果是选择含两个变量的模型，结合变量间的相关性，选择含变量 x_1、x_2 的最小二乘估计模型。

4. 合理性探讨

我们注意到响应变量 y 与自变量 x_2 均为转化百分比，是成分数据，而自变量 x_1 (温度)与 x_3 (时间)并不是，这样直接建立的线性模型可能效果并不是十分好。成分数据是一种在社会、经济、技术等多领域应用十分广泛的数据类型，结合前人的研究，这类数据可以考虑利用对数变换来进行建模，一方面可以起到降维的作用，另一方面解释起来也更为合理。

5. 结语

本文通过采用最小二乘估计法和岭回归法对一个具体的实例进行线性模型估计，并采用留一交叉验证法对模型进行选择，但仅凭交叉验证法估计的预测误差进行判断所得结果未必理想，还需要适当结合实际情况，所得结果才更令人信服。此外，对于成分数据直接进行线性回归可能并不是最为恰当的，还需要我们多加探索，认真实验。

参考文献

[1] Larson, S.C. (1931) The Shrinkage of the Coefficient of Multiple Correlation. *Journal of Educational Psychology*, **22**, 45-55. https://doi.org/10.1037/h0072400

[2] Stone, M. (1974) Cross-Validatory Choice and Assessment of Statistical Prediction. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **36**, 111-147. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

[3] Geisser, S. (1974) A Predictive Approach to the Random Effect Model. *Biometrika*, **61**, 101-107. https://doi.org/10.1093/biomet/61.1.101

[4] Geisser, S. (1975) The Predictive Sample Reuse Method with Applications. *Journal of the American Statistical Association*, **70**, 320-328. https://doi.org/10.1080/01621459.1975.10479865

[5] Devroye, L.P. and Wagner, T.J. (1979) Distribution-Free Performance Bounds for Potential Function Rules. *IEEE Transactions on Information Theory*, **25**, 601-604. https://doi.org/10.1109/TIT.1979.1056087

[6] Shao, J. (1993) Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association*, **88**, 486-494. https://doi.org/10.1080/01621459.1993.10476299

[7] Dietterich, T. (1998) Approximate Statisitical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Computation*, **10**, 1895-1924. https://doi.org/10.1162/089976698300017197

[8] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, **12**, 69-82. https://doi.org/10.1080/00401706.1970.10488635

[9] Celisse, A. (2008) Model Selection in Density Estimation via Cross-Validation. *Density Estimation*, **14**, 1-39.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2325-2251，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：sa@hanspub.org