



# 数据科学导论第 14 讲——重抽样方法

王小宁

中国传媒大学数据科学与智能媒体学院

2025 年 06 月 19 日



# 目录

重抽样概述

基本概念

交叉验证法

自助法



## 重抽样概述



# 重抽样

- 重抽样（resampling）方法是统计学上一个非常重要的工具，它通过反复从训练集中抽取样本，然后对每一个样本重新拟合一个感兴趣的模型，来获取关于拟合模型的附加信息。
- 本节介绍两种最为重要且常用的重抽样方法：
  - ① 交叉验证法（cross-validation）
  - ② 自助法（Bootstrap）

# 例子

- MASS 包含的 Boston 数据集。
- Boston 数据集记录了波士顿周围 506 个街区的 medv（房价中位数），以及与 medv 相关的 13 个变量包括 rm（每栋住宅的平均房间数）、age（平均房龄）和 lstat（社会经济地位低的家庭所占比例）等，下表给出了该数据集的部分信息。

# 数据集

```
library(MASS)  
head(Boston)
```

##		crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptrat
## 1	0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1	296	15	
## 2	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17	
## 3	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17	
## 4	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18	
## 5	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18	
## 6	0.02985	0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18	
##	medv											
## 1	24.0											
## 2	21.6											
## 3	34.7											
## 4	33.4											
## 5	36.2											
## 6	28.7											

```
dim(Boston)
```

```
## [1] 506 14
```

```
names(Boston)
```

```
## [1] "crim"      "zn"        "indus"     "chas"      "nox"       "rm"  
## [8] "dis"       "rad"       "tax"       "ptratio"   "black"     "lstat"
```

- 若我们使用这 13 个变量建立多元线性回归模型来预测 medv，那么如何知道模型的预测效果如何，或者说，如何估计模型的测试误差？
- 另外，若想得到估计系数的分布情况，又该如何操作？



# 基本概念

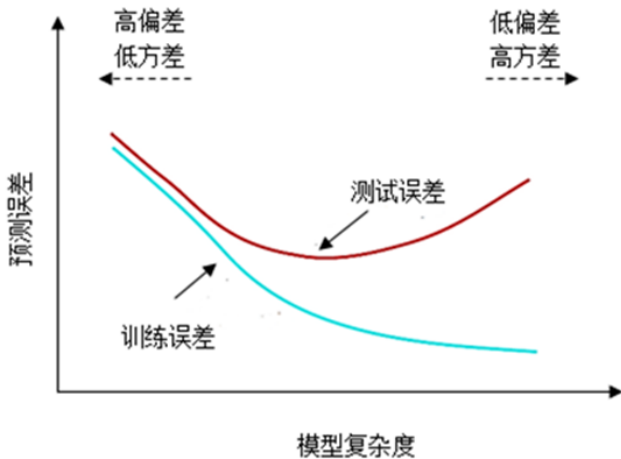


# 训练误差和测试误差

- 在模型训练过程中，一般可通过训练误差和测试误差来衡量模型的拟合精度。
- 训练误差：将一个统计学习方法用于某些观测集上进行训练，得到的模型重新用于这部分观测集进行预测得到的平均误差
- 测试误差：将该模型用于一个新的观测集上（这些观测在训练模型时是没有用到的）来预测对应的因变量所产生的平均误差，它衡量了模型的推广预测能力。

# 关系

- 通常而言，随着模型复杂度的增加，模型的训练误差会一直减小并趋向于 0（最后的模型就是逐点拟合，即出现了过拟合 (overfitting)，如下图的曲线所示。
- 模型的测试误差的变化则如下图上方的曲线所示，通常在模型过于简单时，误差偏高，此时模型欠拟合 (underfitting)。随着模型复杂度的增加，测试误差会先减少后增加。
- 不管是欠拟合还是过拟合，模型的推广预测能力都较差，因此存在一个中等复杂的模型使得测试误差达到最小，目标就是要找到这个最优的模型。



## 模型复杂度与模型的预测误差

# 偏差和方差

- 在统计学习中，通常存在三种误差来源，即随机误差、偏差和方差。
- 随机误差是数据本身的噪声带来的，这种误差是不可避免的。一般认为随机误差服从高斯分布，记作  $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ 。
- 若假定  $Y$  是因变量， $X$  是自变量，有  $Y = f(X) + \varepsilon$
- 偏差描述的是模型拟合结果的期望与真实结果之间的差异，反映的是模型本身的精度，可以表示为：

$$Bias(\hat{f}(X)) = E[\hat{f}(x)] - f(X)$$

# 偏差和方差

- 方差描述了模型每一次的拟合结果与模型拟合结果的期望之间的差异的平方，反映的是模型的稳定性，可以表示为

$$Var(\hat{f}(X)) = E[\hat{f}(x_0) - E(\hat{f}(x_0))]^2$$

- 模型在任意一点  $X = x_0$  的均方误差可表示为：

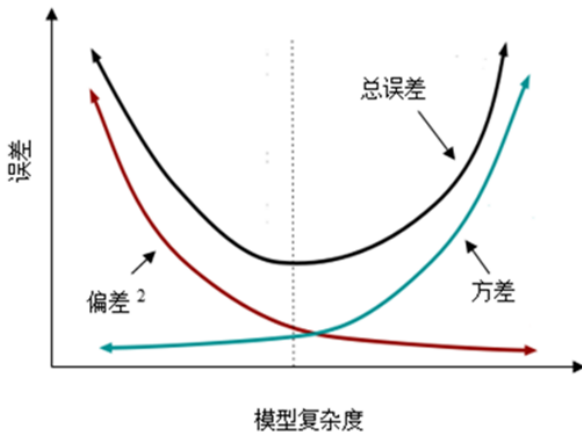
$$Err(X) = E[(Y - \hat{f}(x_0))|X = x_0]$$

$$= Bias^2(\hat{f}(x_0)) + Var(\hat{x}_0) + \sigma_\varepsilon^2$$

## 偏差和方差 2

- 由于随机误差是不可避免的，所以在实际应用中，只能设法减小偏差和方差。
- 在一个实际系统中，偏差与方差往往是无法兼得的。若想降低模型的偏差，就会在一定程度上提高模型的方差，反之亦然。
- 下图给出了模型复杂度与误差的关系，一般而言，当模型较简单时，偏差较大，方差较小，此时模型是欠拟合的。随着模型复杂度的增加，偏差会逐渐减小，而方差会逐渐增大，达到一定程度会出现过拟合的现象。
- 因此，模型过于简单或复杂都是不好的，如何选择一个复杂度适中的模型，即如何对偏差与方差进行权衡（Bias-Variance trade-off）是机器学习中的一个重要问题。

# 模型复杂度与偏差、方差的关系



模型复杂度与方差、偏差的关系



# 交叉验证法



# 交叉验证法 (Cross Validation, CV)

- 用测试误差来衡量模型的推广预测能力，但是一般情况下，并不能事先得到一个测试观测集。
- 幸运的是，现如今已有很多方法可以通过对可获得的训练数据来估计测试误差。
- 采取的方法是：在拟合过程中，保留训练观测的一个子集，先在其余的观测上拟合模型，进而将拟合的模型用于所保留的观测子集上进行预测，从而得到其预测误差的估计。

# 验证集方法-原理

- 把给定的观测集随机地分别不重复的两部分：  
一部分用于训练，称为训练集 (*training set*)  
另一部分用于验证，称为验证集 (*validation set*) 或测试集 (*test set*)；
- 只在训练集上拟合模型，然后将拟合的模型用于验证集上，对验证集中观测的因变量进行预测；
- 在验证集上估计得到的拟合值与真实值的均方误差（回归问题）或分类误差（分类问题）就是该模型的测试误差。



## 验证集方法-弊端

- 最终模型的选取将极大程度地依赖于训练集和验证集的划分方式，因为不同的划分方式会得到不同的测试误差。
- 方法只用了部分数据进行模型的训练。
- 在实际应用中，当用于模型训练的观测越多时，训练得到的模型的效果往往也更好。
- 验证集方法的这种划分使得我们无法充分利用所有的观测，因此对模型的效果有一定的影响。

# 留一交叉验证法 (LOOCV)-原理

- 1. 对于给定的样本容量为  $n$  的观测集, 令  $i = 1, 2, \dots, n$ .
  - ① 将观测  $(x_i, y_i)$  作为验证集, 剩下的观测  $\{(x_1, y_1), \dots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \dots, (x_n, y_n)\}$  作为训练集。
  - ② 在  $n - 1$  个观测组成的训练集上拟合模型, 然后将拟合的模型用于验证集上, 对验证集中的观测利用  $x_i$  预测它的因变量量  $y_i$ , 于是就能得到测试误差的一个渐进无偏的估计:
    - a) 回归问题:  $MSE_i = (y_i - \hat{y}_i)^2$
    - b) 分类问题:  $Err_i = I(y_i \neq \hat{y}_i)$
- 2. 将 1. 得到的  $n$  个测试误差的估计取平均值即得到测试均方误差的 LOOCV 估计:
  - a) 回归问题:  $CV_{(n)} = \frac{1}{n} \sum_{i=1}^n MSE_i$
  - b) 分类问题:  $CV_{(n)} = \frac{1}{n} \sum_{i=1}^n Err_i$

# LOOCV

- 优点

- ① 由于每一个数据都会单独作为验证集，所以说在训练集和验证集上的划分不存在随机性，因此，多次运用 LOOCV 方法总会得到相同的结果。
- ② 由于每一次的训练都使用了几乎所有的（个）观测，所以拟合得到的模型的偏差较小。

- 缺点

- ① 由于需要拟合模型  $n$  次，当  $n$  很大，或者每个单独的模型拟合起来耗时很长时，计算成本将非常大。于是，就有了一种折中的方法—— $K$  折交叉验证法（ $K$ -fold CV）。

## K 折交叉验证法

- 1. 对于给定的样本容量为  $n$  的观测集，随机地将其分为  $K$  个大小相当的组，或者说折 (fold)，令  $k = 1, 2, \dots, K$ ：
  - ① 将第  $k$  折的所有观测视为验证集，剩余  $k - 1$  折的观测均视为训练集；
  - ② 在训练集上拟合模型，然后将拟合的模型用于验证集上，对验证集中观测的因变量量进行预测。同 LOOCV 方法一样，这时可以得到测试误差的一个估计  $MSE_k$  (回归问题) 或  $Err_k$  (分类问题)。
- 2. 将 1 中得到的  $K$  个测试误差的估计取均值即得到测试均方误差的  $K$  折 CV 估计：
  - a) 回归问题:  $CV_{(K)} = \frac{1}{K} \sum_{i=1}^n MSE_k$
  - b) 分类问题:  $CV_{(K)} = \frac{1}{K} \sum_{i=1}^n Err_k$

## K 折交叉验证

- 如何确定  $K$ ?
- 这其实就是涉及偏差-方差权衡的问题。
- $K$  越大，即每次用于拟合模型的训练集包含的观测越多，模型的偏差就越小。
- 特别的，对于 LOOCV，由于每一次的训练都包含了  $n - 1$  个，即近乎所有的观测，故能提供一个近似无偏的测试误差估计。
- $K$  越大，就意味着每一次用于拟合模型的训练集的观测数据越相似。
- 特别的，对于 LOOCV，每一次训练的观测数据几乎是相同的，因此这样拟合得到的结果之间是高度（正）相关的。



- 由于许多高度相关的量的均值要比相关性相对较小的量的均值具有更高的波动性，因此  $K$  越大，得到的测试误差估计的方差也将更大。
- 考虑到上述问题，在实际应该用中一般选取  $K = 5$  或  $K = 10$ ，因为根据经验，这两个取值会使得测试误差的估计不会有过大的偏差或方差。





# 自助法

# 自助法 (Bootstrap) 介绍

- 自助法 (Bootstrap) 是 Efron 在 1979 年提出的一种重抽样方法，是统计学上一种广泛使用且非常强大的方法，可以用于衡量一个指定的估计量或统计方法中不确定的因素。
- Bootstrap 的基本原理是，对已有观测数据进行重抽样得到不同的样本，对每个样本进行估计，进而对总体的分布特性进行统计推断。
- 所谓重抽样，就是指有放回的抽取，即一个观测有可能被重复抽取多次。




# Bootstrap

- 本质上, Bootstrap 方法, 就是将一次的估计过程, 重复上千次上万次, 从而便得到了上千个甚至上万个的估计值, 于是利用重复多次得到的估计值, 就可以估计其均值、标准差、中位数等。
- 尤其当有些估计量的理论分布很难证明时, 可以利用 Bootstrap 方法进行估计。

## • Efron

Stanford University




# Bradley Efron

Professor of Statistics and Biomedical Data Science

Search this site...

[Home](#) [Papers](#) [Talks](#) [Other Works](#) [Interviews](#)



*Computer-Age Statistical Inference*


CONTACT INFO

- BRAD AT STAT.STANFORD.EDU
- SEQUOIA HALL

390 Jane Stanford Way  
Stanford, CA 94305-4065

NEW COURSE NOTES

- Winter 2019: STATS 305B Part 1
- Winter 2019: STATS 305B Part 2
- Winter 2019: STATS 305B Part 3

 MORE INFO  
Curriculum Vita

# 问题

- 假设现在有一部分包含  $X$  和  $Y$  的容量为  $n$  的样本，记为  $Z$ ，想对其建立线性回归模型，那么如何对斜率参数  $\theta$  进行估计呢？
- 在传统的方法中，一般会使用所有已有的样本进行估计得到。
- 但若采用 Bootstrap 方法，我们便可以更好地去估计总体的分布特征，即不仅可以估计  $\theta$ ，还可以估计  $\theta$  的方差、中位数等值。
- 那么 Bootstrap 是如何做到的呢？

# 基本步骤

- 1. 指定重抽样次数  $B$ , 对于  $b = 1, 2, \dots, B$ :
  - ① 在原有的样本中通过重抽样的方式得到一个与原样本大小相同的新样本, 记为  $Z_b^*$ ;
  - ② 基于新产生的样本, 计算我们需要的估计量  $\hat{\theta}_b$ 。
- 2. 对于 1. 中得到的  $B$  个  $\hat{\theta}_b$ , 就可以计算被估计量  $\theta$  的均值和它的标准误差:

a). 均值  $\bar{\theta} = \frac{1}{B} \sum_{i=1}^B \hat{\theta}_b$

b). 标准误差:  $SE(\hat{\theta}) = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}_b - \bar{\theta})^2}$

# 说明

- 在线性回归的情况下，Bootstrap 可能不是特别有用，因为很容易根据公式导出估计量的分布，但是当估计量的分布很难导出的时候，可以利用 Bootstrap 就显得很有用，可以估计  $\hat{\theta}$  方差，分位数等。
- Bootstrap 的强大之处在于，他可以简便地应用于很多统计方法中（用于创造数据的随机性），包括对一些很难获取的波动性指标的估计。
- 例如，随机森林算法的第一步就是从原始的训练数据集中，应用 Bootstrap 方法有放回地随机抽取  $n$  个新的自助样本集，并由此构建  $K$  棵分类回归树。

## 本周推荐

- 书籍：《心理学与生活（第 19 版）》，人民邮电出版社，2016
- 练习：《R 语言实战（第 12 章）》人民邮电出版社，2016



# 期末作业

- 作业要求：
  - ① 按照 CGSS2015 问卷内容提出一个问题，然后在此问题上问题分析说明、已有研究简述、数据分析和建模分析、得出结论等步骤。
  - ② 使用 R 语言分析数据（采用“期末 word/tex 模板.Rmd”），可生成 pdf 或 word 文档，不可生成 html 文档，期末提交作业时将生成文档和源文档同时提交。
- 提交时间和邮箱：
- **2020 年 7 月 25 日 24 时**, [xiaoningwang@cuc.edu.cn](mailto:xiaoningwang@cuc.edu.cn)
- 邮件主题：学号 + 姓名 + 期末作业



谢 谢!