【机器学习】Bootstrap详解

苔执
机器学习话题下的优秀答主

Bootstrap简介

Bootstrap方法是非常有用的一种统计学上的估计方法，是斯坦福统计系的教授Bradley Efron（我曾有幸去教授办公室约谈了一次）在总结、归纳前人研究成果的基础上提出一种新的非参数统计方法。Bootstrap是一类非参数Monte Carlo方法,其实质是对观测信息进行再抽样，进而对总体的分布特性进行统计推断。

因为该方法充分利用了给定的观测信息，不需要模型其他的假设和增加新的观测，并且具有稳健性和效率高的特点。1980年代以来，随着计算机技术被引入到统计实践中来，此方法越来越受欢迎，在机器学习领域应用也很广泛。

首先，Bootstrap通过重抽样，可以避免了Cross-Validation造成的样本减少问题，其次，Bootstrap也可以用于创造数据的随机性。比如，我们所熟知的随机森林算法第一步就是从原始训练数据集中，应用bootstrap方法有放回地随机抽取k个新的自助样本集，并由此构建k棵分类回归树。

具体讲解

下面我们用一个例子具体介绍bootstrap的原理和用法：

假设我们有两个金融资产X和Y，我们现在想要合理配置这两个资产，使得其资产组合的风险最小。也就是找到一个α，使得$Var(\alpha X+(1-\alpha)Y)$最小。这个问题几十年前马尔可维茨已经在其投资组合理论里给出了解答，最优的α表达式如下：

$$\alpha=\frac{\sigma_Y^2-\sigma_{XY}}{\sigma_X^2+\sigma_Y^2-2\sigma_{XY}},$$

$$\text{where } \sigma_X^2=\operatorname{Var}(X), \sigma_Y^2=\operatorname{Var}(Y), \text{ and } \sigma_{XY}=\operatorname{Cov}(X,Y).$$

但是现实生活中实际上我们并不知道σ_X^2,σ_Y^2以及σ_{XY}的值，故而只能通过X和Y的一系列样本对其进行估计。并用估计值$\sigma^{\hat{2}}_X$,$\sigma^{\hat{2}}_Y$以及$\sigma_{\hat{X}Y}$代替σ_X^2,σ_Y^2以及σ_{XY}插入公式：

$$\hat{\alpha}=\frac{\hat{\sigma}_Y^2-\hat{\sigma}_{XY}}{\hat{\sigma}_X^2+\hat{\sigma}_Y^2-2\hat{\sigma}_{XY}}.$$

所以我们唯一的任务就是合理地估计$\sigma^{\hat{2}}_X$,$\sigma^{\hat{2}}_Y$以及$\sigma_{\hat{X}Y}$，传统方法中我们一般会考虑直接使用样本方差（sample variance）去估计σ_X^2,σ_Y^2以及σ_{XY}的值，**然而自从有了Bootstrap之后，我们有了另一种方法与途径，可以更好地去做估计总体的分布特性，即不仅可以估计α，还可以估计α的方差、中位数等值。**下面就讲讲Bootstrap究竟是如何做到这一点的：

Bootstrap步骤：

1. 在原有的样本中通过重抽样抽取一定数量（比如100）的新样本，重抽样（Re-sample）的意思就是有放回的抽取，即一个数据有可以被重复抽取超过一次。

2. 基于产生的新样本，计算我们需要估计的统计量。

在这例子中，我们需要估计的统计量是α，那么我们就需要基于新样本的计算样本方差、协方差的值作为$\sigma^{\hat{2}}_X$,$\sigma^{\hat{2}}_Y$以及$\sigma_{\hat{X}Y}$，然后通过上面公式算出一个$\hat{\alpha}$

3. 重复上述步骤n次（一般是n>1000次）。

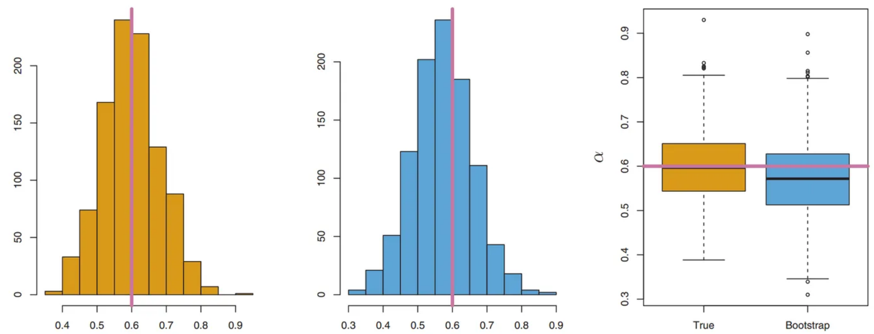
在这个例子中，通过n次（假设n=1000），我们就可以得到1000个α_i。也就是$\alpha_1, \alpha_2, \cdots, \alpha_{1000}$。

4. 最后，我们可以计算被估计量的均值和方差（不用关注最后的具体数值，这与原本的样本有关）：

$$\bar{\alpha} = \frac{1}{1,000} \sum_{r=1}^{1,000} \hat{\alpha}_r = 0.5996,$$

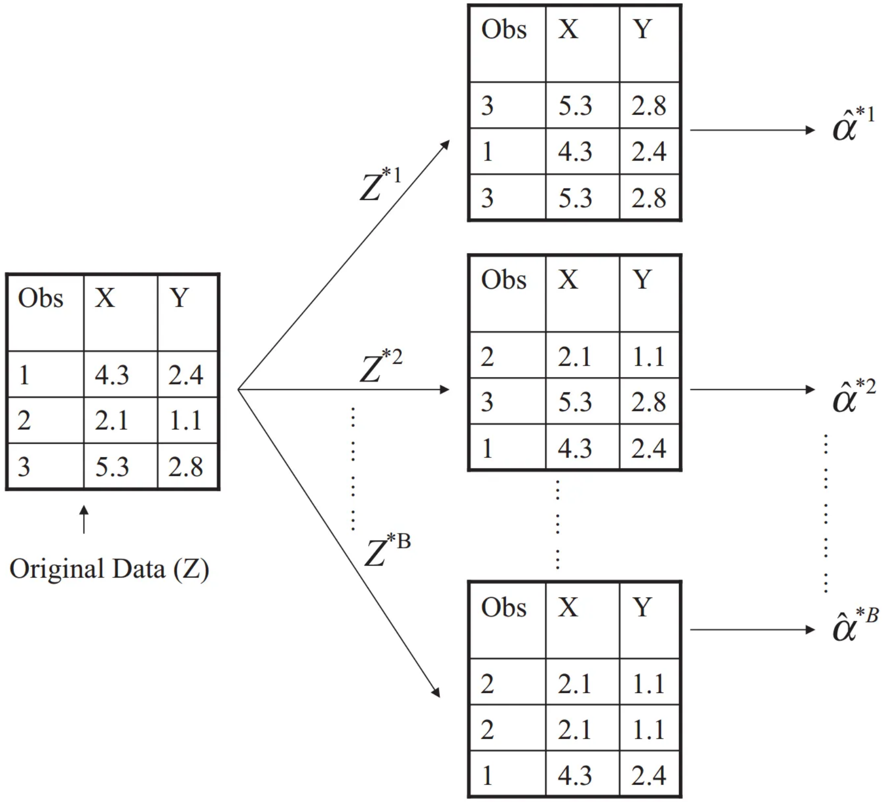
$$\sqrt{\frac{1}{1,000-1} \sum_{r=1}^{1,000} \left(\hat{\alpha}_r - \bar{\alpha}\right)^2} = 0.083.$$

我们发现，通过Bootstrap方法我们竟然不仅可以估计α的值（ 这点普通方法也可以很容易做到），还可以估计α的accuracy也就是其Standard Error。这可是只利用原有的样本进行一次估计所做不到的。那么Bootstrap对于分布特性的估计效果究竟如何呢？请看下图：



左边是真实的α分别，右边则是基于bootstrap方法得到的1000个α的分布，可以看到，二者是比较相近的，也就是说Bootstrap有着不错的估计效果。而且当重复次数增多，Bootstrap的估计效果会更好。

不仅是α的标准差，如果我们想要估计α的中位数、分位数等统计量，也是可以通过Boostrap方法做到的，其整个流程可以用下面一张图诠释：



本质上，Bootstrap方法，是将一次的估计过程，重复上千次上万次，从而便得到了得到上千个甚至上万个的估计值，于是利用这不止一个的估计值，我们就可以估计

α

均值以外的其他统计量：比如标准差、中位数等。

本文部分图片来源：《An Introduction to Statistical Learning with Applications in R》

说在后面

关于机器学习的内容还未结束，请持续关注该专栏的后续文章。

更多内容请关注我的专栏：R Language and Data Mining

或者关注我的知乎账号：温如

编辑于 2017-01-16 10:34

「真诚赞赏，手留余香」

赞赏

还没有人赞赏，快来当第一个赞赏的人吧！

机器学习　数据分析　数据挖掘

欢迎参与讨论

64 条评论　默认　最新

已重置
看题目我还以为是个前端框架😂
2017-01-13　回复　12

张嘉鱼
看题目我还蒙圈了一会2333
2017-01-13　回复　3

没有CO2的可乐
我也是，点进来发现。。。😓
2017-01-14　回复　2

展开其他 2 条回复 ›

Dylan
在我的印象中Bootstrap好像是一套自迭代的方法体系，应用自由而广泛！另外作者讲的这个估计方法，本人总感觉可以解析解得到最终期望值，因为理论上迭代次数越多，就会得到一个比较稳定的样本组合分布。不过我也没仔细研究过，下来抽时间看下😁
2017-01-12　回复　5

清川 ‣ **Jamie**
我也不清楚😳，我是学机器学习模型的evaluation的时候涉及到bootstrap方法的，一般来说对于数据量很小的情况，这种方法测试出的模型性能更准
2023-01-10　回复　喜欢

Jamie ‣ **清川**

请问如果不需要统计量的分布，那么bootsrapping比直接用全部数据求统计量有什么有点吗？（比如估计精度更高或者可以纠偏？）😳求讲解
2023-01-10 回复 喜欢

展开其他 2 条回复 ›

cortex
不过R计算for循环真的龟速啊！(^·ɪ·^)
2017-01-13 回复 2

蔺笑天
点赞~ 文中提到的“真实alpha的分布”，请问这里真实是什么意思，怎么得到的真实alpha分布？如果知道了真实数据，alpha不就直接可以求出最优值了吗，为什么还有个分布？
2020-05-17 回复 1

清川
估计这部分是验证性实验，真实alpha分布是实验设置下自己生成的数据，用来检验bootstrap方法的效果
2022-02-17 回复 4

观一半月
我觉得booststrap的到的是估计值是对一个依赖于某个估计值的不可知值的估计。
boostrap --估计--> 真实分布生产的无数个抽样集的某个估计值的分布 --估计--> 真实值
2023-10-14 回复 喜欢

展开其他 1 条回复 ›

良良
请问boostrap采样n次和直接把样本复制n份有什么区别和优点呢?
2017-03-09 回复 1

了然
bootstrap采样是有放回的，得到的bootstrap样本里面可能会有相同的观测
2019-10-14 回复 2

朴华
额, 这个和bagging的做法是一回事吧
2023-08-10 回复 1

Barry zh
。。。。Bagging = Bootstrap AGGregation😳😳
01-24 回复 1

何舜成
居然没提经验分布函数......
2017-01-13 回复 1

gggggssnv
感觉就是书上讲了一堆都不如一句话：用经验分布替代总体分布。。。
2020-03-11 回复 10

gggggssnv ‣ 寻李白
频率分布是总体分布的非参数估计，自助法就是用频率分布替换总体分布，然后抽样
2021-09-06 回复 3

展开其他 2 条回复 ›

夏目沉吟
周志华的书,讲bootstrap也还不错....
2017-01-13 回复 1

老玩家
您好，请问 是周老师的哪本书？就是讲bootstrap挺好的？ 我想实现bootstrap法进行内部验证
2020-04-13 回复 1

Edvard hua
点赞，诚意满满。公式和符号都是插图显示出来了，这可是要花不少时间的。
2017-01-12 回复 1

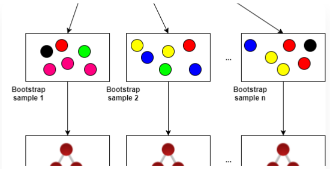
白粥丶
这个感觉和统计里的均值的均值很像
点击查看全部评论 ›
2023-06-30 回复 喜欢

欢迎参与讨论

文章被以下专栏收录

Data Science with R&Python
基于R和Python的数据科学笔记本

推荐阅读

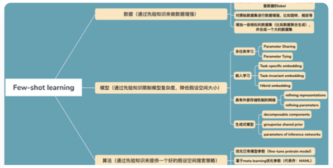


机器学习中Bootstrap的应用

jinpeng.li

深入机器学习系列18FactorizationMachines&O

星环科技



从Few-shot Learning再次认识机器学习

Honda 发表于AI_知识...

机器学习中不同类型的学习范式

AISeer