但这个在实际操作中非常困难,很多时候我们都只有以一个样本,也就 是一个训练集,因此我们得不到eta 的很多个估计值,那怎么办呢? bootstrap

bootstrap的意思是:自助,也就是不需要依靠其他的东西,而是仅仅 是这一组数据就可以解决问题

怎么解决呢? 就是我们把我们唯一有的这个数据(所以为什么把这个数 据当成population,因为我们只有这一组数据)当成总体,然后模拟从 总体中抽取样本的过程

这也就能解释,为什么我们要抽取n个样本(因为我们要模拟抽取n个数 据的样本的过程)

之后,为什么要有放回的抽样,我们设想,如果不放回,我们抽取了第 一个样本,和原样本一模一样,在抽取?就仅仅是复制样本一,这是没 有意义的,所以要放回,同时放回保证了每个数据被抽取到的概率都是

这样我们可以获得M个beta,这些beta模拟的是我们原样本的均值,但 他的方差可以拟合我们beta,也就是真实值的方差,这种resample的 beta成为bootstrap distribution

bootstrap distribution的variance可以approximate sampling distribution

structure part is same, but not the mean part

relationship between bootstrap dustribution and sampling distribution

was used for statistic inference

当我们进行sample的时候,我们确实是可以得到population的

某些信息,从population中抽取多次样本,这些样本的均值,这

些均值放在一起就是sampling distribution, 把每个sample当成

每次训练数据进行linear model拟合,从而得到一个 \hat{eta} ,很多个

 \hat{eta} 可以求出eta 的均值和方差

add some changes to data-->a model

add another changes --> another model data pertobation?

可以帮助得到一些robust,但是不能理论上的证明

那就会产生两个问题 我们有一个数据,需要将其人工分为train data & test data 怎么选择数据,这是有randomness的 比如我就分为50%50%,每一个train data随着自由度变化,可以画出一 条MSE curve, 那么如果我进行resample10次,将MSE取平均,就能够 一定程度的降低randomness 那proportion怎么解决呢?如果更少的数据,那么我们的模型肯定是更 Validation set approach: 不准确的 每次选择一个数据点作为test error,其余的作为train data,因此模型的 拟合只有n次,最终的MSE就是n次拟合的MSE的均值,这样就完全消 掉了randomness variance偏大 因为 f(+). f(+>)是沙常相近雨. B系1个train date &1个tost data. . eizialiste correlation-> Var (n > ei)= 1/m (n e2 + h (n +) p 2)) = 6+(1)96 | LUVON 青波有 岩類 correlation, Var (前至ei)=デ 是常用3几手所有的data > nearly biased MSE根大 LOOCV confidence interval 偏小<--由于model is highly positively correlated disadvantages Resampling proof point estimate is good Prove: eli] $X^{T}X = \begin{bmatrix} X_{i} & X_{Li} \end{bmatrix} \begin{bmatrix} X_{i} \\ X_{Li} \end{bmatrix} = X_{i}X_{i}^{T} + X_{Li}^{T}X_{Li}$ Modified<--我们可以通过一次拟合就可以得到LOOCV的 $\chi_{ij}^{\tau}\chi_{ij}:=\chi_{i}^{\tau}\chi_{-}\chi_{i}\chi_{i}^{\tau}$ Xy= [xi xij] [yi] = xiyi+xiiyii performance<-- h_i 我们可以提前计算 $\begin{array}{ll}
X_{ij}^{I} y_{ij} = x^{T} y - x_{i} y_{i} \\
\Rightarrow \hat{\beta}_{ij} = (x_{ij}^{T} x_{ij})^{T} x_{ij}^{T} y_{ij} \\
= (x_{i}^{T} x_{i} x_{i}^{T})^{T} (x_{i}^{T} y - x_{i}^{T} y_{i}^{T})
\end{array}$ $\begin{array}{ll}
A^{T} u A^{T} \\
u \cdot v \cdot is \quad \text{column. vector.}
\end{array}$ $A = x x. \quad u = -x_i \quad V = x_i$ $\Rightarrow (x^i x - x_i x_i^i)^{-1} = (x^i x)^{-1} - \frac{(x^i x)^i (-x_i) x_i^i (x^i x)^{-1}}{l + x_i^i (x^i x)^i (-x_i)} \quad h = x (x^i x)^i x^i$ $= (x^i x)^{-1} + \frac{(x^i x)^i x_i x_i^i (x^i x)^{-1} - h_i}{l - h_i} \quad h_{ij} = x_i^i (x^i x)^i x_i$ $\Rightarrow \hat{\beta}_{[i]} = \left((x^i x)^{-1} + \frac{(x^i x)^i x_i x_i^i (x^i x)^{-1}}{l - h_i} \right) (x^i y - x_i y_i) \quad \hat{\beta}$ $= (x^i x)^i x^i y - (x^i x)^i x_i y_i + \frac{(x^i x)^i x_i x_i^i (x^i x)^i x^i y_i}{l - h_i} \quad \hat{\beta}_{ij} = \frac{(x^i x)^i x_i y_i^i - (x^i x)^i x_i y_i^i + (x^i x)^i x_i y_i^i +$ yeij=xi Beij eeij= yi- yeij Looch om $= \hat{\beta} - \frac{(x^{1}x)^{1}x_{1}}{|x|} \left((-h_{1})y_{1} - \chi_{1}^{T}\hat{\beta} + h_{1}y_{1} \right)$ $= \hat{\beta} - \frac{(x^{1}x)^{1}x_{1}}{|x|} \left(y_{1} - x_{1}^{T}\hat{\alpha} \right) e_{1}$ $= \hat{\beta} - \frac{(x^{1}x)^{1}x_{1}}{|x|} e_{1} = \hat{\beta}_{[1]}$ $= \hat{\beta} - \frac{(x^{1}x)^{1}x_{1}}{|x|} e_{1}$ $= \hat{\beta} - \frac{(x^{1}x)^{1}x_{1}}$ B= (xTx) xTy. y= xiB ei= yi-yi training error. 直接相比的话,智是 LOOCV error 大. 两种的数据少. Boli zág $h_i = \frac{\partial \hat{y_i}}{\partial y_i}$ $E[\Sigma \frac{\partial \hat{y_i}}{\partial y_i}]$ degree of freedom Ei + hi ei = hi ei Prove: Ei eli] 同样的,Ridge Regression 也是这样的,不需要一定做n次拟合了 将整个数据分成K层,一般K=5/10,,每次将一个K分出来,作为test data,其余的作为train data K-fold cross validation 他为什么有的时候比LOOCV效果好?因为Variance小

How to combine variable selection and cross validation

experiment: 我们设计50个variable,都独立服从于t分布,而Y服从于伯 努利分布,也就是说,X和Y是independent,当我们训练一个分类器, 他的正确率应该是在50%左右才是正确的

正确的是我们应该先分train data & test data,对train data 进行variable selection,也就是说,我每次variable selection 的数据都不一样,这个 必须不能涉及test data Y,我不在乎你到底对train data做了什么,什么 都可以,也就是说cross validation必须包含variable selection和训练一

怎么分,是 五五分, 还是三七分?

if 我们先用每个单独的X和Y做回归,寻找p最小的20个variable,然后利 用这20个X变量和Y进行训练分类器,进行cross validation,我们发现 这个分类器的正确率在14%,远远低于50%,为什么?

因为我们犯了data snopping 的错误,我们已经看了所有的X和Y数据了 ,但又假装我们没看过,然后去进行训练和cross validation,也就是说 ,我们在期末考试前已经看了考试题和答案了,那我根据这个学习,再 去做期末考试,我的成绩肯定会远远高于没看过期末考试答案的,而 train data相当于去年的期末考试题

不会,因为我PCA只看了X,并没有关注Y,哪怕看了test data X,但你 只是看了期末考试题,不会就是不会

那么,如果是PCA呢,两种做法会有不同吗