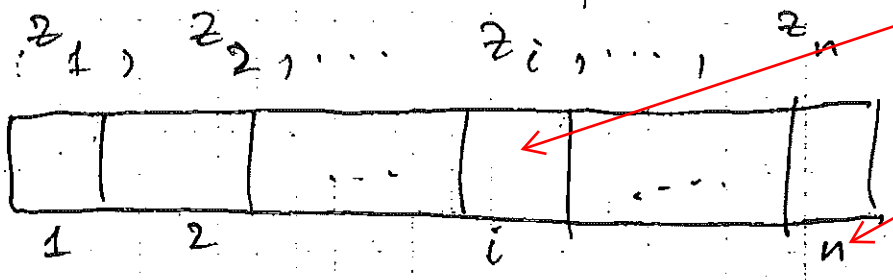
2017-09-26 Discussed on 20-Oct-2020

Numbers of unique values in bootstrap sample [4]



cells correspond to ROW of the data frame

n is n.obs (sample size)

Let X_j be a (categorical) rv.

$\Pr(X_j = i) = 1/n$ for $i = 1, \ldots, n$.

X_j is multinomial rv.

Resampling z_i's with replacement is equivalent to generating categories $X_1, X_2, \ldots, X_n$

Define $Y_i = \begin{cases} 1 & \text{if } X_j = i \text{ for some } j. \\ 0 & \text{if } X_j \neq i \text{ for all } j. \end{cases}$

$S_n = \sum_{i=1}^{n} Y_i$; want $E(S_n) = \sum_{i=1}^{n} E(Y_i) = n \cdot 0.632$.

$E(Y_i) = \Pr(Y_i = 1) = 1 - \Pr(Y_i = 0)$ ★ $= 1 - 0.368 = 0.632$.

$\Pr(Y_i = 0) = \Pr(\text{none of the } X_j\text{'s fell into category } i)$

$= \Pr\left(\bigcap_{j=1}^{n} [X_j \neq i]\right) = \prod_{j=1}^{n} \Pr(X_j \neq i) =$

since multinomial trials are iid

$= \left(1 - \frac{1}{n}\right)^n \approx e^{-1}$ for large n. ≈ 0.368 ★

DATE:

PAGE #:

"Greedy" search: target best improvement ~~at~~ one step at a time.

suppose we have 3 features

X_1, X_2, X_3

true model is $Y = X_1 + X_2 + \varepsilon$.

suppose $Corr(Y, X_3) > Corr(Y, X_1) > 0$

$Corr(Y, X_3) > Corr(Y, X_2) > 0$.

E.g., X3 = X1+X2+U, where U is a bit of noise

=> X_3 is picked at step 1

=> X_1 and X_2 are never chosen by forward selection.

Recap: Total SS $= \sum_{i=1}^{n} (Y_i - \bar{Y}_n)^2$:

does not depend on d or the covariates

$$R^2 = 1 - \frac{\text{Resid SS}}{\text{Total SS}} = 1 - (1 - R^2)$$

$$R^2 \text{ adj} = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)} \cdot \frac{(n-1)}{(n-1)}$$

$$= 1 - \frac{RSS/(n-1)}{TSS/(n-1)} \cdot \frac{n-1}{n-d-1}$$

$$= 1 - (1 - R^2) \cdot \frac{(n-1)}{n-d-1} \rightarrow (n-1-d+d)$$

$$\underbrace{\left(1 + \frac{d}{n-d-1}\right)}$$

=> extra penalty (over unpenalized R^2) is $(1-R^2) \cdot \frac{d}{n-d+1}$