The Second Report

Zheng Yuan

2019/3/18

Shao's Method

(1)The Balanced Incomplete CV(n_v) Method

Notations:

Sample size in model fitting: n_c
Sample size in prediction: n_v
So $n_c + n_v = n$
Category I: At least obe nonzero component of β is not in β_α
Category II:β_α contains all nonzero components of β

Dataset splits:

We don't need to enumerate all of the subsets that contain n_v samples. We only need part of them. Let B denote the collection of all b subsets that have size n_v. B satisfies:
(a)every i, $1 \leq i \leq n$, appears in the same number of subsets in B;
(b) every pair (i,j), $1 \leq i < j \leq n$, appears in the same number of subsets in B.

BICV(n_v):

The BICV(n_v) selects a model by minimizing

$$\hat{\Gamma}^{BICV}_{\alpha,n} = \frac{1}{n_v b} \sum_{s \in B} ||y_s - \hat{y}_{\alpha,s^c}||^2$$

over all $\alpha \in A$. Each α here represents a kind of model.

Important results:

(1)Under appropriate conditions, BICV(n_v) is asymptotically correct if $n_c \to \infty$ and $n_v/n \to 1$.
(2)Why BICV(n_v) can beat CV(1): large n_c will lead to a flat cross validation error over all the dels in Category II, where

$$\Gamma_{\alpha,n_c} = \sigma^2 + n_c^{-1} d_\alpha \sigma^2$$

. So it is difficult to find a smallest Γ_{α,n_c}. While with relatively small n_c in BICV(n_v), this problem could be solved. (Why, here n_c is still large, only is relatively small compared with n_v.
Besides, can $n_v/n \to 1$ be relaxed under most cases? (Not all cases, cause he provides a counterexample.)
Even though we should improve $n_v = 1$, but is $n_v/n \to 1$ is too strong compared with $n_v = 1$?

(2)Other CV(n_v) Methods:

Monte Carlo CV(n_v) Ramdonly draw

Analytic Approximate CV(n_v)