Modern Inference (Statistical Learning) - University of Milano-Bicocca

Homework 5

To submit via e-mail by 24/04/2020 h 14:00.
You will have to submit your solution by the deadline. Feel free to choose the format for your solution (.txt, .tex, .pdf etc.), the nicer the better, and the number of files in attachment (one or more, but not too many). Answer with clarity and precision. All R code must be reproducible. For theoretical questions, try to provide a well-reasoned mathematical argument. Simulations can help form your intuition; but a purely empirical answer will only receive partial points. It is encouraged to discuss the problem sets with others, but every group needs to turn in a unique write-up. Use of sources (people, books, internet and so on) without citing them in homework sets results in failing grade.

1 Combining multiple split

Prove the Theorem at page 16 of the slides "Data Splitting Inference".

2 Find the important variables

Download the datasets `lowXy.Rdata` and `highXy.Rdata`.
`lowXy.Rdata`: response `y`, variables `X1,...,X19`, $n = 40$.
`highXy.Rdata`: response `y`, variables `X1,...,X2000`, $n = 200$.
For each dataset, find the set of important variables. A truly important variable is a variable with nonzero coefficient, otherwise is a null variable. Consider the variables only: it does not matter whether the intercept term is zero or not. Let S^* and N^* be the sets of truly important variables and truly null variables, respectively. Given your selected set $\hat{S}$ of important variables, type I selection errors $\hat{S} \cap N^*$ (selecting null variables) are more serious than type II selection errors (not selecting important variables).