STAT 5443: Computational Statistics

Homework 2 – Due December, 8 2021

Get hold of the prostate cancer data, available in the *R scripts and examples* folder in Blackboard. The response variable is `lpsa` (column 9), while columns 1–8 contain different predictors. (Column 10 is not relevant for this problem and can be discarded).

Use package `leaps` to perform a "best-subset" linear regression analysis. Clearly, the model that minimizes the residual sum of squares is the full model. As we discussed in class, using that model to predict out-of-sample responses is not a good idea. One possible approach to prevent overfitting is to select a model based on minimizing information-theoretic criteria, like AIC or BIC. What model would you select, based on AIC? What model would you select, based on BIC? Another approach that we discussed in class is to use cross-validation. Use 5- and 10-fold cross-validation to determine the best model to use for predictions. Provide, in both cases, the final model to be used to predict future cases.