Home assignment I (Machine Learning in Economics)

Deadline: 4th April 2023, 23:55 via Moodle

Please write your solution as a properly edited “essay”-type document with explanations and concise tables and figures. I will evaluate ONLY this document. Points will be decreased if the document is unnecessarily lengthy or not transparent.

Also, please upload your R script as well. However, this is just for background, I will check it only if there seems to be some discrepancy (e.g. direct copying from a peer).

1. We examine the properties of ridge regression and LASSO in a very simplified setting. Suppose that $k = n$, i.e. we have the same number of parameters as observations, and there is no intercept term.
   Then OLS minimizes $\sum(y_i - \beta_i)^2$, hence the OLS estimates are $\hat{\beta}_i = y_i$.
   a. Write down the minimized expressions (as functions of λ) in case of the ridge regression and LASSO.
   b. Solve the minimization problems. How do the estimated β_i-s depend on λ? Explain the differences between ridge regression and LASSO based on this example.

2. For a sample of size n, calculate the probability that a sample element does not appear in a bootstrap sample. Find the limit of this expression as $n \to \infty$.

3. Open the twoyear data set in the wooldridge package: data(twoyear, package='wooldridge'), which contains wage data for 6763 people, originating from the National Longitudinal Study of the High School Class of 1971, and used by Kane and Rouse (1995) to study the wage returns of two- and four-year colleges.[1] The dependent variable is lwage (log wage) or the original wage, for details of the explanatory variables see the documentation.[2]
   Unless indicated otherwise, use a random 60% sample for the training set, 20% for the test set and 20% for the hold-out set.

[1] T.J. Kane and C.E. Rouse (1995), Labor-Market Returns to Two- and Four-Year Colleges, American Economic Review 85, 600-614.

[2] e.g. https://rdrr.io/cran/wooldridge/man/twoyear.html

If requested, use five-fold cross-validation on the training and test set combined (i.e. on the 80% of the sample).

a. Use “domain knowledge” to fit at least three different regression models (on the training set) explaining wage with the explanatory variables (and potentially their polynomials and interactions). Evaluate their performance with information criteria and their prediction accuracy on the test set.

b. Similarly, fit three regression models explaining log(wage). Evaluate their performance with information criteria and their prediction accuracy on the test set. Also predict the wage (instead of log wage) using the adjustment covered in class.

c. Compare the best model of part a. and part b. in terms of prediction accuracy for the level of the wage on the test set.

d. Use various λ values in LASSO estimated on the training set to predict log(wage) on the test set. (You should include some polynomials and / or interaction terms in the original set of variables.) Graph the test set MSE as a function of λ.

e. Find the best LASSO model by cross-validation.

f. Fit a pruned regression tree to the training and test set combined (the cost complexity parameter can be determined by cross-validation). Plot and interpret the obtained tree.

g. Fit at least five random forests with different tuning parameters (e.g. m = the number of variables [bagging] or its square root; T = 25 or 500) to the training set and compare their performance on the test set. (The baseline option should be included.) Choose the best model.

h. Fit at least five boosted trees with various tuning parameters (e.g. $\lambda = 0.01$ or $\lambda = 0.001$; different number of trees; the baseline option should be included) and compare their performance on the test set. Choose the best model.

i. Finally, use the holdout set to compare the performance of the best models of part b, e, f, g, h. (In doing so, if it was not done earlier, it is reasonable to refit all the “best” models on the previous training and test set combined, and extrapolate them to the holdout set.)