Statistical Learning, Homework #2

Veronica Vinciotti, Marco Chierici

Released: 20/04/2021. Due: 05/05/2021

This homework deals with model/variable selection methods and decision trees.

You should submit an RMarkdown file and a pdf file of the report. The RMarkdown file should reproduce exactly the pdf file that you will submit.

Note that:

- your code should run without errors (except for minor adjustments such as file paths);
- you should discuss/justify each choice that you make and provide comments on the results that you obtain.

Exercise 1

Consider the "fat" dataset provided for this Homework (tab-separated `fat.tsv`). It contains percent body fat, age, weight, height and body circumference measurements for 252 male subjects. Our goal is to predict body fat (variable `y` in the dataset) from the other explanatory variables.

1. Load the data and perform a first exploratory analysis
2. Split the data into train/test
3. Perform least squares regression to predict `y` from the other variables. Discuss the results of the model and compute the test MSE.
4. Apply ridge regression and the lasso to the same data. For each method:
    - Plot the coefficients as a function of λ
    - Plot the cross-validation MSE as a function of λ and find the optimal λ
    - Compute the test MSE of the optimal model
    - Examine the coefficients of the optimal model
5. Critically evaluate the results you obtained. If they look suspicious, think about a possible cause. For example, examine the coefficients of the least square regression model (estimate and sign), together with the R^2 value; compute the pairwise correlations between the variables, . . .
   Think of a modification of the analysis in light of your findings and repeat steps 1-4 of your new analysis. Comment on the new results.

Exercise 2

In this question, you will revisit the `Hitters` dataset. The goal is to predict the salary of baseball players, as a quantitative variable, from the other explanatory variables.

1. Split the data into training/test sets.

2. Fit a decision tree on the training data and plot the results. Choose the tree complexity by cross-validation: plot the cross-validation deviance versus the number of terminal nodes and prune the tree if applicable. Finally, evaluate the optimal model by computing the test MSE.

3. Apply bagging on the training portion of the data and evaluate the test MSE. Does bagging improve the performance?

4. When we grow a random forest, we have to choose the number m of variables to consider at each split. Remember that bagging is a particular case of random forest with m equal to the number of explanatory variables $nvar$. Set the range for m from 1 to $nvar$. Define a matrix with $nvar$ rows and 2 columns and fill it with the test error (1st column) and OOB error on training data (2nd column) corresponding to each choice of m. Save the matrix as a dataframe and give it suitable column names. Compare OOB errors with test errors across the m values. Are the values different? Do they reach the minimum for the same value of m?

5. Reach a conclusion about the optimal random forest model on the training data and evaluate the model performance on the test data. Identify the variables that are important for prediction.

6. Fit a regression tree on the training data using boosting. Find the optimal number of boosting iterations, both by evaluating the OOB error and the cross-validation error. Produce plots with OOB error and CV error against the number of iterations: are the two methods leading to the same choice of the optimal number of iterations? Reach a conclusion about the optimal model, evaluate the test MSE of this model and produce a partial dependence plot of the resulting top N variables (N of your choice).

7. Draw some general conclusions about the analysis and the different methods that you considered.