DS5110 HW 5 - Due March 20

Kylie Ariel Bemis

3/9/2020

Instructions

Create a directory with the following structure:

- `hw5-your-name/hw5-your-name.Rmd`
- `hw5-your-name/hw5-your-name.pdf`

where `hw5-your-name.Rmd` is an R Markdown file that compiles to create `hw5-your-name.pdf`.

Do not include data in the directory. Compress the directory as `.zip`.

Your solution should include all of the code necessary to answer the problems. All of your code should run (assuming the data is available). All plots should be generated using `ggplot2`. Missing values and overplotting should be handled appropriately. Axes should be labeled clearly and accurately.

To submit your solution, create a new private post of type "Note" on Piazza, select "Individual Student(s) / Instructor(s)" and type "Instructors", select the folder ***"hw5"***, go to Insert->Insert file in the Rich Text Editor, upload your `.zip` homework solution. Title your note "[hw5 solutions] - your name" and post the private note to Piazza. **Be sure to post it only to instructors**

Part A

Problem 1

Choose one of the "miniposters" created by your fellow classmates and posted on Piazza for Homework 3. Cite both the name of the student whose miniposter you chose and the original source of the dataset used in that miniposter.

Download and import that dataset into R, put it into a tidy format (if necessary), and print the first ten observations of the dataset.

Problem 2

To the best of your ability, reproduce the figures from the miniposter you chose. You may contact the author of the original miniposter; if you do, cite and describe any information you receive from them.

(If you are contacted for information on reproducing figures from your own miniposter, you may provide it, but you are not obligated respond.)

Part B

Problems 3–5 use the `PimaIndiansDiabetes2` data from the `mlbench` package. Install the `mlbench` package from CRAN and use `data(PimaIndiansDiabetes2)` to load the dataset.

Problem 3

Write a function that performs cross-validation for a linear model (fit using `lm`) and returns the average root-mean-square-error across all folds. The function should take as arguments (1) a formula used to fit the model, (2) a dataset, and (3) the number of folds to use for cross-validation. The function should partition the dataset, fit a model on each training partition, make predictions on each test partition, and return the average root-mean-square-error (RMSE).

(Do NOT use a pre-existing function such as `caret::train()` that already performs cross-validation.)

Using 5-fold cross-validation, report the cross-validated RMSE of the model you proposed in Homework 4, Problem 5.

Problem 4

Now consider only the predictors `triceps`, `diabetes`, `pressure`, `insulin`, and `glucose`. Use cross-validation to perform stepwise model selection with these variables to predict `mass`. Then create a plot showing the RMSE at each step of variable selection.

What was the most predictive model using stepwise selection with these variables? How does it compare to the one proposed in Homework 4?

(Keep any transformations you used in previous models, but you do not need to do any further checks for linearity. You do not need to handle missing data.)

Problem 5

Can you report the final cross-validated RMSE for the "best" model that you found in Problem 4 as a good measure of the RMSE we could expect on new data? Why or why not?