Homework 4

Step functions, splines, smoothing splines and Generalized Additive models

The aim of the homework is to find a good Generalized Additive Model for predicting value of variable `y` in the case of a generated data set. You have training data of 4000 observations in the file `hw4_train.csv`; I'll compute the performance of you best model on an additional test set of 2000 observations and the best 5 students (according to test set RMSE of their models) get extra points (from 5 to 1) for the homework assignment. In order to qualify for extra points, the R script (described below) has to work correctly.

The conditions of the assignment:

1) The modeling should use for model fitting the `gam()` command from `mgcv` package (satisfied automatically if you use the `tidymodels` framework with `gen_additive_mod()`).
2) You final model should use up to 6 variables including
   - a linear spline approximation with maximum 7 interior knots (given by the parameter `knots` of `splines::bs()` command) with respect to one numerical predictor
   - a natural spline approximation with maximal degree of freedom 6 with respect to a second numerical predictor
   - a smoothing spline with respect to a third numerical predictor
   - Up to 3 additional linear terms (a factor variable or collection of dummy variables for it's levels is considered one term)
3) Your submission should contain a report (in .html or .pdf format showing code and output) which describes what you did in order to find a good model and a separate R script, which
   - Reads in the training set from `hw4_train.csv` (file name should be given without any additional path information in the file input command)
   - fits the best model you found to the full training set
   - reads in the test set from 'hw4_test.csv' (note that it does not contain the column for `y`)
   - computes predictions for observations in the test set and stores them in the variable `my_predictions` Your script should work correctly if the script and csv files are in the same directory and before running the script, the working directory is changed to that location (Session->Set Working Directory-> To Source File location) before executing the script.

   You can check your code by using a sample test set with two observations from Moodle: clear the workspace before executing (in RStudio use menu Session->Clear Workspace), then set the working directory and then run the file.

I remind you that you are expected to solve the homework problems by yourself, sharing solutions (and copying parts of other student's solutions) is not allowed.