DS4B: Final Project

A. Fidalgo

Due date: July 29, 2019

This document provides the instructions for the final project that students must present for this class.

In short

In this project, you must provide:

- your **best** vector of predictions of length `140 x 7 = 980`, necessarily called `our.predictions`,
- a fully reproducible `Rmd` file giving and explaining **all** the steps and R code used to obtain that vector.

A `train.data` is provided for you to estimate and select relevant models. There are 7 different responses to be estimated. Models may (should) vary for each response.
Based on these models, the predictions are made on `test.data` for which you have the predictors but not the responses.
Your vector of predictions will be evaluated with the root mean square error criterion both for the overall performance and for each individual response.

Data

The file `project_data.Rdata` contains two data frames, `train.data` and `test.data`. These names are self-explanatory. Their dimensions are the following.

```
dim(train.data)
```

```
## [1] 200  18
```

```
dim(test.data)
```

```
## [1] 140  11
```

The data used in this problem comes from a study on water quality. Samples were taken from sites on different rivers where the quantities of eight chemical substances were recorded: the maximum pH value (`mxPH`), the minimum oxygen value (`mnO2`) as well as the mean values of chloride (`Cl`), nitrates (`NO3`), ammonium (`NH4`), orthophosphate (`oPO4`), phosphate (`PO4`) and chlorophyll (`Chla`).
Further information for each observation includes three categorical variables: the season when the sample was taken (`season`), the river size (`size`) and the flow velocity (`speed`).
In total, that are 11 predictors for the frequency of 7 plants. In the `train.data`, these plants are note `a1` to `a7`.
The `test.data` has the same structure but does not contain the frequencies for each of the 7 plants. Your goal is precisely to estimate them for the 140 observations.
Data summaries and visualizations can help you gain insights into the data.

Some specifics

Models and their selection

Because there are 7 responses to estimate, you can/should use 7 models. Available techniques include linear models, trees, random forests, etc...
For every model that you try, provide a short description. Of course, you do not need to describe sub-versions such as yet another polynomial. However, you cannot simply present only your chosen model.
The process of selection must also be described. This include parameter selection such as the degree of polynomial or the size of the tree. Importantly, it also includes methods to evaluate the models (cross-validation).
Your 980 predictions must be stacked into a vector so that they can be compared with the actual values and evaluated with the root mean square error.
Your scored on the quality of the predictions will be determined by comparing your RMSE with mine.

NA's

Both data frames contain NA's. To deal with them, there are various options:

- remove the observations with NA's,
- fill the blanks,
- use a mix of these two.

Of course, whatever method is chosen in the train data, must similarly be used in the test data. Recall, however, that the test data must **never** be used in the estimation process.

In this particular case, it is acceptable to delete train observations that seem too noisy to be useful in the estimation.

```
train.data %>% is.na %>% rowSums %>% table
```

```
## .
##   0   1   2   6
## 184   7   7   2
```

As for the missing values in some variables, you can fill them with an appropriate choice. Here, you have many options. These range from the simplest (and least accurate) such as replacing with the mean of the variable, to the most sophisticated such as replacing each missing value with the prediction of a complex model. Intermediate solutions include exploiting correlations between variables. No matter what option you take, you must be clear about your choice.

```
test.data %>% complete.cases %>% table
```

```
## .
## FALSE  TRUE
##    18   122
```

Recall that the test data also contains missing values. Hence, your predictions will start by filling them by using the same approach that you used with the train data.

Intermediate submissions

At any moment until the deadline, you can submit a vector of predictions (and its explaining `Rmd`). I will then evaluate the RMSE and tell you how satisfactory it is.

Questions and further details

It is hoped that this description is sufficient for the task. However, if you have questions, please send me an email.
If necessary or relevant, I will update this file with further details. To make the changes clear, I will gather them in the next section along with their date.

Updates

June 28

Description of the project.