Final Exam 'Advanced Multivariate Analysis'

Pedro Delicado

2024-01-08

Algerian forest fires

The dataset `firesAlg_tr` (contained in `firesAlg.Rdata`) includes 183 instances that regroup a data of two regions of Algeria, namely the Bejaia region located in the northeast of Algeria and the Sidi Bel-abbes region located in the northwest of Algeria. Each instance corresponds to a different day in one of the two regions.

The dataset includes 10 explanatory attributes and 2 output attributes:

1. temp : temperature noon (temperature max) in Celsius degrees: 22 to 42
2. RH : Relative Humidity in %: 21 to 89
3. wind :Wind speed in km/h: 6 to 29
4. rain: total day in mm: 0 to 16.8 FWI Components

5. Fine Fuel Moisture Code (FFMC) index from the Fire Weather Index (FWI) system: 28.6 to 96
6. Duff Moisture Code (DMC) index from the FWI system: 0.7 to 65.9
7. Drought Code (DC) index from the FWI system: 6.9 to 220.4
8. Initial Spread Index (ISI) index from the FWI system: 0 to 19
9. Buildup Index (BUI) index from the FWI system: 1.1 to 68
10. Fire Weather Index (FWI) Index: 0 to 31.1
11. fire: (output) 0 for "not Fire", 1 for "Fire"
12. region: (output) 0 for Bejaia, 1 for Sidi Bel-abbes

The dataset `firesAlg_test_blinded` (also contained in `firesAlg.Rdata`) has similar structure as `firesAlg_tr` but it does not contain the output variables. There are 61 instances in `firesAlg_test_blinded`.

```r
load("firesAlg.Rdata")
ls()
```

```
## [1] "firesAlg_test_blind" "firesAlg_tr"
```

Second part of the course (5.4 points)

1. (2.8 points) Generalized additive model for a binary variable.

1.1 (1.8 points)

Use `firesAlg_tr` to fit a generalized additive model with response the variable `region` and explanatory variables chosen among the first 10 columns of `firesAlg_tr`.

- Justify the steps you do in the model choice process.
- Indicate clearly which is your finally chosen model.

1.2 (1 point)

To evaluate the performance of your chosen model, I will take use the following quantity:

$$C = G_{\text{test}} - \max\{0, G_{\text{tr}} - G_{\text{test}}\},$$

where G_{tr} is the proportion of good classified instances in the training sample, and G_{test} is the proportion of good classified instances in the blinded test sample (I will compute this quantity later, when grading your exam).

The quantity C will be large when both G_{tr} and G_{test} are large and they are similar to each other. In an overfitted model G_{tr} would be much larger that G_{test} and then C would not be so large.

(*Info:* C is equal to 0.82 for the generalized linear model including all 10 explanatory variables. I've been able to fit a model for which $C = 0.88$.)

Your grade at this item will be

$$\min\left\{1, \max\left\{\frac{C-0.82}{0.88-0.82}, 0\right\}\right\}.$$

2. (1.8 points) Interpretable machine learning.

Consider the dataset obtained by joining 6 of the 10 columns that are common in `firesAlg_tr` and `firesAlg_test_blinded`:

```
cols<- c(1,2,3,7,9,10)
firesAlg_6 <- rbind(firesAlg_tr[,cols],firesAlg_test_blind[,cols])
```

Fit a random forest (with library `ranger`) to explain `FWI` as a function of the other variables in `firesAlg_6`. At the same time, compute the

a. Compute the *Variable Importance* by the reduction of the **impurity** at the splits defined by each variable. (*Hint: Use `set.seed(1234)` before calling the function `ranger`*). Plot the results and comment on them.

b. Compute the Variable Importance by out-of-bag random permutations. (*Hint: Use `set.seed(1234)` before calling the function `ranger`.* This way you fit the same random forest as before). Plot the results and comment on them.

c. Compute the Variable Importance of each variable by Shapley Values. Plot the results and comment on them.

d. Use the DALEX library to do the Local (or Conditional) Dependence Plot for each explanatory variable.

3. (0.8 points) Your own Local (or Conditional) Dependence Plot.

Construct your own Local (or Conditional) Dependence Plot for the explanatory variable `BUI`.

- For doing that, consider the pairs of variables

$$x = \texttt{BUI},\ y = \widehat{\texttt{FWI}},$$

where $\widehat{\texttt{FWI}}$ are the predicted values of `FWI` using the random forest, and use the smoother of your preference.

- Indicate how the required smoothing parameters have been chosen.
- Plot the resulting Local (or Conditional) Dependence Plot over the scatterplot of (x, y).
- Add to the previous plot the Local (or Conditional) Dependence Plot for the explanatory variable `BUI` obtained by DALEX.

First part of the course: Unsupervised learning (3.6 points)

4. (0.9 points) Mixed Gaussian Model.

Consider the dataset `firesAlg_6`. Do a model based clustering of these data assuming a Gaussian Mixture Model, allowing varying volume, shape, and orientation for different components in the mixture. Choose k_{BIC}, the best number of clusters $k \in \{2, \dots, 6\}$ according to BIC. Plot the resulting object from `Mclust` (do 4 different graphics: `BIC`, `classification`, `uncertainty` and `density`).

5. (0.9 points) DBCAN.

Use DBSCAN to find clusters (and outliers) in the data set `firesALG_6`, after **centering and scaling** the variables. Use $\varepsilon = 1$ and `minPts` $= 8$. How many clusters have you obtained? How many outliers? Do a pairs plot of `firesALG_6` coloring the points according to the results of DBSCAN.

6. (1.8 points) Nonlinear dimensionality reduction.

Use a nonlinear dimensionality reduction method at your choice to obtain a 2-dimensional configuration for the data in `firesAlg_6`, after **centering and scaling** the variables.

- Specify how you choose the required tuning parameters.
- Provide graphical representation of the output. In particular, show how the 6 original variables are related with the new 2 dimensions.
- Try to give an interpretation to the new 2 dimensions.