Final Exam AMA-DS, January 2023. Practice

Pedro Delicado

2023-01-09

Human Development Index

In the `HDI.Rdata` file you will find some data published by the UNITED NATIONS DEVELOPMENT PROGRAM in its *Human Development Report* corresponding to the year 2017. After reading this file with

```
load("HDI.Rdata")
```

you will find the *data frame* `HDI` with the following the following variables:

Variable	Description
HDI	Human Development Index (HDI)
Median_age	Median age (years)
Old_age_dep_rat	Old-age (65 and older) dependency ratio (per 100 people ages 15-64)
Exp_years_school	Expected years of schooling (years)
Life_expec	Life expectancy at birth (years)
Inf_Mort_rat	Mortality rate, infant (per 1,000 live births)
log_GDP_per_cap	logarithm of the Gross domestic product (GDP) per capita (PPP $: international dollars with purchasing power parity)
Mob_phon_100_peopl	Mobile phone subscriptions (per 100 people)
Agric_employ_perc	Employment in agriculture (% of total employment)
Servi_employ_perc	Employment in services (% of total employment)
Employment_rat	Employment to population ratio (% ages 15 and older)
Continent	Continent

The **country names** are the row names of `HDI` and can be recovered by

```
country <- row.names(HDI)
```

1. Dimensionality reduction (2 points out of 10)

Let X be the data matrix with columns 2 to 11 of `HDI`:

```
X <- as.matrix(HDI[,2:11])
```

Do dimensionality reduction of the columns of X, from 10 dimesnions to $q = 1$, in two different ways:

1.1 Computing the principal curve (as defined by Hastie and Stuetzle, 1989). Select the tuning parameter by using the local continuity meta criteria (with $K' = 5$). Choose `df` in `seq(15,25,by=2)`

1.2 Performing local MDS. Select the tuning parameters by using the local continuity meta criteria (with $K' = 5$). Choose `K` in `seq(3,15,by=2)` and `tau` in `c(.5,1)`.

1.3 Do a *pairs plot* (that is, a matrix of scatterplots) of the matrix having the following 4 columns:

- `HDI$HDI`
- First principal component of X
- 1-dimensional configuration obtained by principal curves.
- 1-dimensional configuration obtained by local MDS.

1.4 Comment your results.

2. Nonparametric regression (2.5 points out of 10)

2.1 Do the scatterplot of $y =$`Old_age_dep_rat` against $x =$ `Median_age`. Do yout think there are some countries which could be considered outliers?

2.2 To identify those possible outliers, fit a simple nonparametric regression model of `Old_age_dep_rat` as a function of `Median_age`. Use smoothing splines with *degrees of freedom* chosen by Generalized Cross Validation.

2.3 Once you have fitted this model, which is the estimated value $\hat{\sigma}$ of the residual standard deviation? If this estimation is not included in the fitted smoothing spline object, you can compute it from the residuals. If the residuals are not directly available, you can use the functions `fitted` to compute the fitted values of your estimated model and use them to compute residuals.

2.4 Add to the scatterplot of $y =$`Old_age_dep_rat` against $x =$ `Median_age` the following elements:

- the points with coordinates $x =$ `Median_age` and $y =$ the fitted values.
- the points with coordinates $x =$ `Median_age` and $y =$ the fitted values $+/- 1.96\,\hat{\sigma}$.
- the names of the countries for which the absolute value of the residuals are larger than $1.96\ \hat{\sigma}$. These points are those you can consider outliers.

2.5 Could you provide any explanation why those countries are outliers?

3. Functional data analysis (2.5 points out of 10)

The Rdata file `HDI_series.Rdata` contains the data frame `HDI_series`, with rows corresponding to the 134 countries (the row names of `HDI_series` are the country names). The 32 columns of `HDI_series` corresponds to years 1990 to 2021. The entry (i, j) of `HDI_series` is the value of HDI for country i at year j. *(Note: The source of data of `HDI` and `HDI_series` are different. Then the values of `HDI_series` for year 2017 do not coincide exactly with those of `HDI$HDI`)*

3.1 Transform the data in `HDI_series` into a functional data set, with countries as individuals and years as argument of the functions:

- First, transform the raw data to a `fdata`object.
- Then, smooth the raw functional data in an optimal way using generalized cross-validation and local linear smoothing.

Call `HDI_LL` to the smoothed functional data.

3.2 Do descriptive statistics of `HDI_LL` (mean, median and standard deviation functions). *(Note: If you have not found `HDI_LL` then use the raw data functional data set)*

3.3 Look for outliers in the functional data set `HDI_LL`. *(Note: If you have not found `HDI_LL` then use the raw data functional data set)*

3.4 Perform Functional PCA on `HDI_LL` and try to give an interpretation of the first two principal functions. *(Note: If you have not found `HDI_LL` then use the raw data functional data set)*

3.5 Do the scatterplot of the scores on the first two principal functions, adding the country name to each point. Comment your results.

4. Interpretable Machine Learning (2 points out of 10)

Consider the problem of predicting `HDI$HDI` from the rest of variables in the data frame `HDI`.

4.1 Found the gam model you consider the *best one* to explain `HDI$HDI` from the rest of variables in the data frame `HDI`. Call if `gamHDI`.

4.2 Fit a random forest to explain `HDI$HDI` from the rest of variables in the data frame `HDI`. Call if `rfHDI`.

4.3 Compute the relevance of each explanatory variable by Shapley values at both models, `gamHDI` and `rfHDI`. Comment your results.

4.4 Consider these two countries: Spain and Qatar. For each of them, provide local explanations for the predicted values when using the random forest `rfHDI` (choose the explanatory methods that you consider most appropriate.) Comment your results.