Machine Learning Project

Deadline: May, 20th 2022

Description of the dataset

The **rain.txt** file contains 688 meteorological observations at a given station, provided by *Météo France* for the challenge *Défi IA 2022*. It has been extracted from the dataset available on the Météo France github[1]. The aim of this project is to predict the amount of rainfall over the next day.

The explanatory variables are:

- Observed meteorological parameters during **the current day**:
  - `date`: the date of the current day,
  - `ff`: the wind speed (in $m.s^{-1}$);
  - `t`: the temperature (in Kelvin K);
  - `td`: the dew point (in K);
  - `hu`: the humidity (in %);
  - `dd`: the wind direction (in degrees);
  - `precip`: the total amount of precipitation (in $kg.m^{-2}$);
- Forecasts of meteorological parameters for **the next day** by the *Météo France* AROME model:
  - `ws_arome`: the wind speed (in $m.s^{-1}$);
  - `p3031_arome`: the wind direction (in degrees),
  - `u10_arome` and `vu10_arome`: the U (from West to East) and V (from South to North) wind components (in $m.s^{-1}$) at the vertical level of $10m$;
  - `t2m_arome`: the temperature at the vertical level of $2m$ (in K);
  - `d2m_arome`: the dew point at the vertical level of $2m$ (in K);
  - `r_arome`: the humidity (in %) ;
  - `tp_arome`: the total amount of precipitation in ($kg.m^{-2}$);
  - `msl_arome`: the sea level pressure (in Pa);

For more details, you can refer to the *Météo France* github.

The response variables are:

- `rain` (quantitative): the total amount of rainfall over the next day (in $kg.m^{-2}$);
- `rain_class` (qualitative): an artificially created categorial variable with three classes that are `no_rain` (if `rain` $= 0$), `low_rain` (if $0 <$ `rain` ≤ 2), and `high_rain` (if `rain` > 2);

We consider here the classification problem: to predict the rain quantity (`rain_class`) during the next day from the explanatory variables.

[1] https://meteofrance.github.io/meteonet/english/data/ground-observations/

Questions

Data analysis

The aim of the section is to control and understand the data, which is a useful preliminary step. The questions below are the basics that you should do. Feel free to complete them with your own ideas.

1. Replace the variable `date` by a categorial variable `month`.
2. Start with some unidimensional descriptive statistics of the dataset. Can you see anomalies?
3. Continue with a multidimensional descriptive analysis. In particular, using visualization techniques (e.g. scatterplot, correlation plot, boxplot), which variable(s) seem to be the most influential on the output? Can you see interactions?
4. Consider the quantitative variables, except `rain` and perform a principal component analysis. Can you see clusters? Are they linked with the `rain_classes`?

Models

Now we consider the prediction problem with a machine learning point of view, i.e. by focusing on the model performance. What best performance can we expect? Below are some guiding questions.

1. First of all, split the data into a training set and a test set. Why is this step necessary when we focus on performance?
2. Here, we consider the classification problem directly. Compare the performance of a linear model with/without penalization, SVM, an optimal tree, random forest, boosting and neural networks. Justify your choices (e.g. which kernel for SVM), and tune carefully the parameters. Interpret the results and quantify the improvement brought by non-linear models.
3. Now, we first consider the regression problem and then classify using the given thresholds. Same question as before.
4. What approach is the best to predict the rain classes: direct classification or "regression plus thresholding"?
5. Interpretation and come-back to data analysis. Are your results consistent with the preliminary data analysis, e.g. about non-linearities, influence of variables (or variable importance)?

Project organization and deliverables

The project has to be done by groups of 4 students. **Deadline: May, 20th**. As deliverables, a pdf report which does not exceed 30 pages is expected. It must include an introduction, an interpretation of the results, a conclusion, etc. Moreover, two Jupyter notebooks are also expected, one in R, the other in Python. Do not forget to comment your code. The deposit will be done in Moodle: each group will upload a zip file containing the report (pdf format) and the two Jupyter notebooks.

The evaluation will take into account the presentation and the writing (clarity, argumentation, etc) of the report, on the consistency of the study, the coherence between both R and Python notebooks and obviously, the interpretations of the results (graphs and others).