Machine Learning Project

Deadline: May, 14th 2021

Description of the dataset

The **spotify-extr.txt** file contains 10.000 songs collected from Spotify Web API. It has been extracted from the dataset available on the Kaggle website. Moreover, we have created a qualitative output variable, called `pop.class`, that quantifies the popularity of the tracks. Its levels `A`, `B`, `C` and `D` were obtained by slicing the quantitative output `popularity` variable with the thresholds: 20, 40, 60, 100. More precisely, the levels `D`, `C`, `B` and `A` respectively correspond to a value of the quantitative `popularity` variable in the intervals : $[0, 20[$, $[20, 40[$, $[40, 60[$ and $[60, 100]$.

The explanatory variables are: `valence` (the positiveness of the track), `year` (the release year of track), `acousticness` (the relative metric of the track being acoustic), `danceability` (the relative measurement of the track being danceable), `duration` (the length of the track in milliseconds), `energy` (the energy of the track), `instrumentalness` (the relative ratio of the track being instrumental), `key` (the primary key of the track, `Bb` meaning A $\sharp$ or B $\flat$), `liveness` (the relative duration of the track sounding as a live performance), `loudness` (the relative loudness of the track in the typical range [-60, 0] in decibel), `mode` (a binary value representing whether the track starts with a major [encoded `1`] chord progression or not [encoded `0`]), `speechiness` (the relative length of the track containing any kind of human voice), `tempo` (the tempo of the track in Beat Per Minute).

We consider here the classification problem: to predict the popularity class (`pop.class`).

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. Start with some unidimensional descriptive statistics of the dataset. Can you see anomalies?
2. 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?
3. Consider the quantitative variables, except `popularity` and perform a principal component analysis. Can you see clusters? Do they correspond to the popularity 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 (logistic regression) with/without penalization, SVM, an optimal tree, random forest and neural networks. Justify your choices (e.g. 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 popularity classes: direct classification or regression+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, 14th**. 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).