Big Data Science: Examination Project

10th March 2022

Project assignment

1. The project is done in groups of 2. Each student writes up one of the tasks (it should be indicated who wrote what part), but both are responsible (and are graded) on the full content of the report.

2. It is also possible to do the project individually. In that case, the student must do only one task, but the protocol has to be submitted for both tasks.

3. You are allowed to change one (and only one) of the tasks by replacing the dataset with another dataset of your own interest. Take care, in the end you will have to have one classification and one regression case.

4. Write a one-page protocol per task. Include any manipulations of the data that you performed to prepare them for subsequent analysis. This protocol should be submitted to the lecturers as soon as possible, via the assignment in Ufora. The deadline is **1st April 2022**. Deviations from the protocol will be allowed if these are deemed necessary during the analysis of the data. Please write a section with amendments to the protocol in the event of this.

5. Write a concise, to-the-point and clear report of at most 4 pages about your analysis for each task. Exceeding the page limit will be penalized.

6. Choose one of the tasks and bring in elements of big data analysis. Present part of the data analysis in a way that would be necessary if the sample size was vastly larger. You do not need to actually implement this, only the report is sufficient.

   (a) Pay special attention to data visualization (and automation).

   (b) Reflect on the gains/losses between a rather simple and less computationally demanding approach and what can be done by modern statistics if the sample size is a lot bigger than the number of variables and computer power is abundant.

(c) Present a simple parallelized version of the data: divide the data in smaller datasets which are analyzed separately and then merge those results back into a single answer in the most meaningful way. Try to minimize the loss of information and discuss the process of how (not) to divide and recombine.

(d) Report on this in at most 2 pages.

7. Therefore, the total length of the report should not be more than 10 pages (4 pages/task + 2 pages for the big data analysis).

8. Take care to properly reference any source materials that you use. Add the reference list to your report. Also write who did what in a small paragraph here. This does not count in the page limit.

9. When submitting the report, enclose a compressed file containing your source code. The code should be complete and commented, so that it is clear what is the purpose of each block of code.

10. Post your report via the assignment in Ufora. The deadline is **3rd June 2022 at 23:59**. Note that the submission will remain open until June 10th, but submissions after the deadline of June 3rd will receive a penalization of 40% to the grade. This is to avoid that good students would fail the course due to unforeseen issues in last-minute submissions (so, please, avoid last-minute submissions!).

11. The oral examination will be scheduled during the exam period, on **17th and 18th of June**; a detailed time schedule will be available once the written reports are received.

Enjoy and good luck!

General considerations for both tasks

The amount of work carried out for either task is expected to be similar. There is no 'easy' or 'difficult' task. Both provided datasets pose a number of challenges:

- The dimensionality of the data is relatively high, and some problems related to the curse of dimensionality may occur.
- Many of the features extracted from the images might be highly correlated.
- No external expert knowledge of the data may be assumed. The analysis of the data must be done and reasoned from a purely statistical and computational perspective.
- The number of instances might be sufficiently large to cause some difficulties when using algorithms that are not scalable. If the dataset is too large to handle with your computer, take this into account for your analysis and justify how you deal with this challenge (e.g. subsampling, dimensionality reduction, instance selection, prototype generation, approximate methods...).

You are required to carry out at least the following steps for each task:

1. Do a proper partitioning of the data to do a validation of the conclusions and their quality (e.g. cross-validation, repeated splitting in train/test portions...).
2. Clean your data properly (i.e. outlier detection, missing data handling, normalization, etc.).
3. Compare at least three different prediction models.
4. Use several complementary performance measures to evaluate your models. Take into account the characteristics of the data and the problem when choosing these performance measures.
5. Explore at least three dimensionality reduction mechanisms and compare your model (using the same approach as in the previous part) to the baseline models. You can either choose specific feature selection methods (e.g. filter approaches), use model-based approaches such as wrapper or embedded approaches (e.g. RandomForest based importance values, regularization...), or feature extraction algorithms that generate new features (e.g. PCA, Isomap...).
6. Of course, you may optionally explore different ways of improving the results, for instance:
    - Selecting a subset of instances or generating new prototypes
    - Using an ensemble of models

- Applying alternative distance measures or loss functions when training the algorithms
- Augmenting the data by artificially generating more instances
- ...

Task 1: Classification - Fingerprint classification

If you ever looked at your fingertips, you might notice that your fingerprints can be grouped into the 5 classes shown in Figure 1.

For this task, you are provided 5 fingerprint datasets:

- SFinGe: 3 datasets of 10,000 fingerprints generated with a software, with different image qualities (High, Default, and Varying).
- NIST: 2 datasets scanned from 1,650 ink prints each. These correspond to two different captures of the same fingers, although not necessarily in the same order.

The images have already been processed with various methods to extract features that can be used for the classification. All these features are provided together; therefore, they are expected to be redundant. Note that the original images of the SFinGe and the NIST datasets have different sizes, and this leads to different numbers of features for these datasets. Therefore, a classifier or feature selection trained on a dataset cannot be directly applied on another; however, the same methodology might (and should) be followed for all datasets.

For each dataset you are provided a CSV file, whose last column is the class.

Figure 1: Fingerprint classes

Task 2: Regression - Music Analysis

You have a dataset that contains 518 features extracted from 106,574 tracks from 16,341 artists and 14,854 albums. Additionally, 52 metadata features are provided. This dataset was designed in order to be used as a benchmark for various problem, such as genre classification. In this exercise, we will focus on two regression tasks, that target two output values:

- Number of times a track is listened (metadata feature `track_listens`)
- Release time (or year) of a track (metadata feature `album_date_released`)

The rest of the metadata is provided for information purposes, but it should not be used to train the model. It can be used to visualize and interpret the results.