Big Data Science: Examination Project

5 March 2020

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. 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.

3. 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 **27th March 2020**. 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.

4. 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.

5. 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 analysed separately and then merge those results back into a single answer in the most meaningful way. Try to minimise the loss of information and discuss the process of how (not) to divide and recombine.

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

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

7. 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.

8. 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.

9. Post your report via the assignment in Ufora. The deadline is **19th May 2020 at 23:30**. Note that the submission will remain open until May 26th, but submissions after the deadline of May 19th 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!).

10. The oral examination will be scheduled during the exam period (on May 28th and 29th); 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.

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 without feature selection. You can either choose specific feature selection methods (e.g. filter approaches), or use model-based approaches such as wrapper or embedded approaches (e.g. RandomForest based importance values, regularization...).
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 - p53 Mutants

Protein p53 prevents cancer formation in various organisms by triggering the death of cells that undergo an uncontrolled cellular proliferation. Therefore, mutations of this protein that disrupt its tumor suppression mechanisms are a very important factor in human cancer.

This dataset contains 5409 numeric features extracted from biophysical models of 31270 mutant p53 proteins. These features represent 2D electrostatic and surface characteristics of the molecules, as well as some 3D distances. The class of every protein (either *active* or *inactive*) is also provided, and was determined experimentally.

The goal of this task is to classify every possible mutation into active or inactive. Note that there is a significant imbalance between the two classes (there are 151 active mutations, and 31269 inactive ones). Your model should avoid predicting everything as the majority class. Therefore, you should use appropriate evaluation measures (e.g. ROC based measures, balanced accuracy...) and possibly specific sampling approaches to deal with imbalanced data.

Task 2: Regression - Gas sensor array drift at different concentrations

You have a dataset that was experimentally obtained with 16 chemical sensors exposed to 6 gases at different concentration levels.

The data is already pre-processed so as to obtain 8 features from each particular sensor; this yields a total of 128 features for every measurement. Every measurement corresponds to one of 6 possible gases at a specific concentration; the goal of this task is to estimate the concentration of a gas given the sensor measurements.

There are several levels of difficulty for this task:

1. Train a separate regression model for every gas
2. Train a single regression model to estimate the concentration of any gases
3. Train a single regression model to estimate the concentration of all 6 gases separately

For processing purposes, the dataset is organized into ten batches, each containing the number of measurements per class and month indicated in the tables below. This reorganization of data was done to ensure having a sufficient and as uniformly distributed as possible number of experiments in each batch (Table).

Batch 1	Months 1 and 2
Batch 2	Months 3, 4, 8, 9 and 10
Batch 3	Months 11, 12, and 13
Batch 4	Months 14 and 15
Batch 5	Month 16
Batch 6	Months 17, 18, 19, and 20
Batch 7	Month 21
Batch 8	Months 22 and 23
Batch 9	Months 24 and 30
Batch 10	Month 36

Table 1: Batches for the regression dataset