Data Science Lab: Process and methods
Politecnico di Torino

Exam rules A.Y. 2019/2020

Last update: January 11, 2020

1 Exam composition

The exam includes an individual project (assigned before each exam session) and a written part. Both the project and written parts must be passed in the same session. The final score is defined by considering the evaluation of the individual project and the written part.

Assigned task. For each exam session, the project task will be published in a separate file. It will also contain information on important dates for the current session.

Final grade. The maximum grade is 32, subdivided as follows:

- Individual project: max. 20 points (12 performance, 8 report)
- Written part: max. 12 points

The final grade is given by the sum of the grades of the two parts. The exam is passed if (i) the grade of the individual project is greater than or equal to 11, (ii) the grade of the project report is greater than or equal to 4, (iii) the grade of the written part is greater than or equal to 7, and (iv) the overall grade is greater than or equal to 18. If the final score is greater than 31 the registered score will be 30 with honor. If the exam is failed, the exam failure is recorded.

2 Individual project

The project consists of designing and implementing a data science process for solving an assigned data analytics task. The proposed task will be either a classification or a regression problem.

The evaluation of the individual project is based on (i) the performance and accuracy of the proposed solution, in terms of standard quality measures (e.g., prediction accuracy) and (ii) the quality of the process (i.e., in-depth analysis of each phase of the designed process and motivation for selecting given techniques and algorithms).

2.1 Project rules

Duration. The problem specifications will be available 14 days before the final submission deadline.

Deliverables. The project is composed by the following three deliverables. For the project to be valid, **all** three deliverables must be uploaded.

1. The analysis result. The latter must be uploaded to the online submission platform provided for the exam session. Further details are provided in Section 2.2.
2. A report describing the steps performed in the project. Further details and guidelines are provided in Section 2.3.
3. The software used to obtain the above result. Further details are provided in Section 2.4.

Points. The maximum grade for the project is 20, subdivided as follows:

- performance of the proposed solution: 12 points
- quality of the report: 8 points

Time validity. The project score expires at the end of the exam session.

Out-of-program methodologies. The adoption of algorithms and methodologies that are not part of the course program is allowed. However, the proposed approach must be extensively motivated in the report and a **further oral examination** may be requested to complete the evaluation.

Additional data sources. It is possible to use additional data sources, provided that they are publicly available and that the sources' public links are cited in the report. In the report, a clear motivation for the adoption of those sources must be provided.

Further verification. In any case, a further assessment (oral or written) on the delivered project report and/or software may be required by the teachers to specific students.

2.2 Result submission

Each exam session will have a dedicated online submission platform. The outcome of the implemented pipeline must be uploaded to the platform. The performance of the proposed solution is evaluated with a specific evaluation metric. The achieved score places the submission on a public leaderboard.

Dataset composition. The dataset of the proposed task is split into (a) a Development set and (b) an Evaluation set. The Development set comes with the information on the target variable and must be used for training and validation. The Evaluation set is used for the evaluation of the submission.

Leaderboard. The leaderboard is provided by the platform. It ranks all the submissions received for the current assignment. The leaderboard contains two baseline scores, as low and high thresholds. As such, the baselines split the leaderboard into three sections.

To enforce the stability and reliability of the solution, the Evaluation set is split into two parts, namely *public* and *private*, and, whenever a new result is submitted, two scores are computed. The two parts have the same statistical distribution. Scores on the *public* part are used to compose the public leaderboard. Scores on the *private* one will not be publicly available.

Final evaluation. Only the scores achieved on the *private* part are considered for the final evaluation. Specifically, the final grade is given by two factors. First, the baseline reached by the proposed solution (either the lowest one, the highest one or none of them) is considered. Then, points are assigned considering the positioning after that all the submissions that reached the same baseline have been normalized. Note that the lower baseline will be publicly available on the leaderboard.

Submission rules. The following is a list of rules enforced by the online submission platform.

- Maximum 100 submissions.
- Minimum 5 minutes between two consecutive submissions.
- Every submission will be recorded. Then, the student is allowed to choose at most 2 of them to be considered for the final evaluation. If more than one is chosen, only the best performing one, on the *private* part, is considered. If no submission is chosen, the best performing one on the *public* part is selected and its *private* score is considered.

2.3 Report submission

One single report is allowed. If two solutions are selected for the evaluation, they both must be described in the same report.

The report should describe the key steps and takeaways of the implemented pipeline. It should focus on the following steps: (1) data exploration, (2) preprocessing, (3) algorithm choice, and (4) tuning and validation. The relevance of the content and the quality of the presentation are evaluated.

Report structure rules.

- The report must contain one section for each of the aforementioned steps. Each section of the report must contain **at most 400 words**.
- The report may include **no more than 8 figures** overall.

The report must be compiled using one of the following templates: TEX, DOCX. Then, it must be exported as a single PDF file and uploaded following the guidelines of the current session.

2.4 Software submission

The software must be written in Python. It must be organized either in a single Jupyter Notebook file or as a single Python script. If two solutions are selected, at most one file per solution can be used (e.g. `solution1.py`, `solution2.py`). All the software files must be uploaded following the guidelines of the current session.

3 Written part

The written part covers the theoretical topics of the course. It includes multiple-choice and box-to-fill questions, based on solving exercises related to theoretical aspects. Several questions may be proposed on a single problem.

Rules.

- Only students regularly booked through the "Portale della Didattica" are admitted to the classroom for the written part. Students must provide their own identity document to be allowed to take the exam.
- The written part lasts 45 minutes.
- The written part includes up to 12 questions.
- Paper books and paper notes are allowed. Instead, no electronic devices (PC, laptop, mobile phone, smart watches, calculators, etc.) are allowed.
- For each question, there is a single correct answer. Only for multiple-choice questions, wrong answers are penalized. Points of each question will be specified in the exam text.

Topics. Questions may address one or more of the following topics. Each question may require a practical application or example.

- Data preparation: discretization, normalization, distance measures.
- Association rules: extraction algorithms (and practical examples), itemset types (e.g. closed, etc.), quality indices.
- Classification: algorithms, quality indices, validation strategies.
- Regression: algorithms, quality indices.
- Clustering: algorithms, quality indices.
- Time series: data characterization.
- Python notions and operations.