Data Science Lab: Process and methods
Politecnico di Torino

Exam rules A.Y. 2020/2021

Last update: December 13, 2020

1 Exam composition

The exam includes an individual project (assigned before each written exam) and a written part. There will be a project assignment per exam call (i.e. the winter session will have two assignments). The final score is defined by considering the evaluation of the individual project and the written part.

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

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

- Individual project: max. 16 points (6 performance, 10 report)
- Written part: max. 16 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 10, (ii) the grade of the written part is greater than or equal to 8, and (iii) the overall grade is greater than or equal to 18. If the final score is strictly 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 at least 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 16, subdivided as follows:

- performance of the proposed solution: 6 points
- quality of the report: 10 points

Project validity. As specified in Section 1, each exam call has an associated project assignment: there will be 2 in January, 1 in July, and 1 in September. **The score achieved in the project is valid only for the associated call**. This is important for the January session! If you decide to take the second call, you also must solve the second project assignment, whether you already solved the first assignment or not.

Out-of-syllabus 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.

Withdrawal. The project score is recorded and assigned if and only if (i) the student ID is present in the leaderboard (i.e. the student did at least one submission) and (ii) the project report is sent. In case any of the two is missing, the project is considered as withdrawn and the exam call failed.

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 call 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 several baseline scores, as low and high thresholds. Only the lowest baseline is made available on the leaderboard.

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 is considered. Then, points are normalized against all the solutions by other participants that reach the same baseline

Note that the lowest 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.

As stated in Section 2.1 (withdrawal), if no solution is submitted before the deadline the project is considered as rejected and the exam call is failed.

2.3 Report submission

Only one report submission 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 must be structured using the following sections and subsections:

1. Problem overview
2. Proposed approach
   (a) Preprocessing
   (b) Model selection
   (c) Hyperparameters tuning
3. Results
4. Discussion

5. References

The relevance of the contents and the quality of the presentation are evaluated.

Constraints on the report. The report must comply with the following constraints:

- The report **must** be generated using the standard IEEE conference template (available on the course website in LaTeX and Word format). You are *strongly* encouraged to use the LaTeX version (either locally, or on Overleaf).
- The report **must** follow the division in sections and subsections introduced in 2.3.
- The report **must** be, at most, 4 pages long. You are not allowed to change font size, margins or anything else in the provided template.

Grading. The grading of the report will be based on:

- Relevance of the content
- Quality of the presentation
- Significance and relevance of images and tables
- Presence of key steps for the successful completion of the task
- Additional steps that helped achieve improved performance

A grade of 0 points will be assigned for the report in case of adoption of a non-complying template (i.e. a template different from the IEEE conference one, or a modified template). The maximum grade for the report is 10 points.

2.4 Software submission

The software solution 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 in a single **ZIP file** (other extensions will not be accepted). Please note that the software submissions will only be used for automated reproducibility tests. Make sure that any and all contents that require evaluation (e.g. images, tables) are included as a part of the report.

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.

The written exam takes place with a PC using the *Exam* university framework integrated with the *Respondus* proctoring tool. **The correct execution of the exam on the Respondus tool will be verified and irregular behaviors will be prosecuted following the university rules.**

Rules. The following is a list of strict rules for the written exam.

- Only students regularly booked through the "Portale della Didattica" are admitted for the written part.
- Students must show their own identity document with a photo.
- The written part lasts **80 minutes**.
- The written part includes up to **16 questions**. The maximum total score is 16/32.
- Books, notes, electronic devices of any kind (smartphones, smart watches, other PC/laptop/tablet, calculators, etc.), apart from the electronic device used for the exam, **are not allowed**.
- In case of connectivity problems (or other technical problems) the university rules in the "Guidelines for online exams" will be followed.
- 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 an 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.

Provided material. All the formulas contained in both the lectures slides and laboratory texts will be provided. If needed, the definition will be included in the text of the exercise. For what concerns the Python language, questions will require the knowledge of the language rather than the actual API functions.