Big Data Computing Final Project Guidelines

Gabriele Tolomei

Sapienza University of Rome, Italy

Email: tolomei@di.uniroma1.it **Homepage:** https://www.di.uniroma1.it/∼tolomei

Scope of the Document

This document provides students of the **2020-21 Big Data Computing** class (hereinafter, "you") with a list of *guidelines* for developing their projects, which are mandatory for final grading. Please, read through the *whole* document carefully.

1 Project Proposal

No matter what project you decide to work on (more on this in the sections below), **this must first be approved**. In order to get such approval, you have to come up with a written project proposal, i.e., a half-page document, that contains at least the following information:

- The problem/task you are planning to address (e.g., binary classification).
- The dataset(s) you will be using along with their references, or how do you plan to collect data if no dataset is already available for achieving your goal.
- The methods you would like to experiment with (e.g., SVM, logistic regression, random forest, etc.).
- The evaluation framework you will use to assess the quality of each method (e.g., accuracy, precision-recall, ROC, etc.)

Project proposals are *mandatory* and must be submitted via email at tolomei@di.uniroma1.it *before* the actual project submission, using "`BDC 2020-21 Project Proposal`" as the subject line. Although there is no specific deadline for issuing your project proposal, a good rule of thumb is to send it at least one month before the deadline by which you plan to submit your project.

Remember: Do not start working on a project if you didn't get it approved first! In any examination session, if you submit a project on time but its proposal hasn't been previously approved, this will *not* be accepted nor evaluated for grading.

2 Project Requirements

Projects must of course refer to a typical big data task, such as those seen during classes: e.g., clustering, regression/classification, recommendation, graph analysis, using large datasets in *any* application domain of interest.

The development environment must be the same of that used throughout the course, namely PySpark in combination with Databricks Community Edition platform.

Projects can be done either individually or in group of **at most** 2 students, and they should be accompanied by a brief presentation written in english (e.g., a few PowerPoint slides).

There is no restriction on the learning technique(s) you should use to solve your task (e.g., K-means, logistic regression, matrix factorization, artificial neural networks, etc.) In other words, you can experiment with any technique you want – providing it is relevant to your goal and hopefully comparing more than one with each other – including also those not covered in class. In fact, you are strongly encouraged to explore non-standard techniques!

3 Project Selection

There exist several resources that contain many project ideas as well as the related datasets to work on; among those, I would recommend the following ones:

- Kaggle
- UCI Machine Learning Repository
- Awesome Public Datasets

Alternatively, you can also come up with your own project idea, as long as it satisfies the requirements indicated in the section above.

4 Dataset Policy

It is **strictly forbidden** to use datasets which are downloaded without the required permissions and/or coming from illegitimate sources. You can, of course, build your own dataset to work with but that must be collected according to the law, and all the steps performed must be properly documented and acknowledged.

5 Project Evaluation

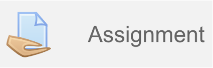
Projects will be evaluated according to the following **5** criteria:

- "**Big-Data-ness**": This criterion will check if the "big data" requirements of the project are actually met. It will provide answers to questions like "*How complex the task is?*" or "*How large is the dataset used?*".
- **Coding**: This criterion will evaluate the style of programming. It will provide answers to questions like "*Is the source code easy to read, reusable, and well documented?*" or "*How much effort would it take for the source code to get into a hypothetical production system?*".
- **Methodology**: This criterion will assess the soundness of the methodological approach used to achieve the project's goal. It will provide answers to questions like "*Does the pipeline implement all the stages needed?*" or "*How reliable are the results obtained?*".
- **Originality/Impact**: This criterion will validate both the novelty of the task that the project aims to solve and the creativity of the solutions implemented. It will provide answers to questions like "*Is the task novel/impactful or has it been already deeply investigated?*" or "*How conventional are the solutions adopted to solve the task?*".
- **Presentation**: This criterion will judge the quality of the presentation attached to the project. It will provide answers to questions like "*Is the project well defined and its goal clearly stated?*" or "*What is the level of understanding of the challenges associated with the specific task?*".

Note: Teamwork projects are expected to be more "complex" than single-person projects, so as to justify the need for a two-person effort.

6 Project Submission

Projects will be submitted for grading to the Moodle web page of the course. More specifically, for each examination session there will be an *Assignment* entry on Moodle, which will allow you to upload your project material. Moodle assignments are identified by the following icon:



For example, suppose you are ready to submit your project on the June 2021 session. On the Moodle web page of the course, you will see an entry called "*June 2021 Exam Session*", along with an assignment corresponding to that session named "*June 2021 Project Submission Week*". This assignment has a **one-week** time window, within which you must submit your project (e.g., from **June, 12 at 00:00** to **June, 18 at 23:59**[1]). More generally, everyone who wants to submit her/his project *must* upload the material within the deadline established by the project submission week of a specific examination session.

The project material *must* be packaged inside a **single archive file**, namely a (compressed) folder (e.g., `.tar`, `.tgz`, `.bz2`, etc.) containing the following **two items**:

- A notebook file (`.ipynb`) with all the source code of the project;
- A presentation (e.g., PowerPoint or PDF slides) with a description of the project, the main choices made to accomplish the task, and the results obtained.

Please, consider that the notebook must be "ready-to-execute": in other words, it must contain everything to be run properly and successfully (e.g., environment setup, library dependencies, etc.)

To ease the grading process, it may be helpful to setup a naming convention for the project submission. The uploaded folder should be named as follows: `X(_Y).Z`, where:

- `X` and `Y` are *student IDs* of the project team;
- `Z` = Archive extension (e.g., `tar`, `tgz`, `bz2`, etc.).

In the case of a single-person team, the folder will be named as `X.Z`, whilst in the case of two-person team, `X` and `Y` will be the ID of the first and second student, *after they have been alphabetically ordered by their last name.*

For example, if the project is done by a single student whose ID is "12345", then `X`=`12345` and the folder will be named `12345.tgz` (or similar extension). Instead, if the project is done by two students: *Clark Kent* whose student ID is "67890" and *Bruce Wayne* whose student ID is "12345", then the project folder will be named `67890_12345.tgz` (or similar extension). In addition, files within the project folder (i.e., the notebook and the presentation) should follow the same naming convention.

Note: Teamwork projects must be submitted **only once by one member** of the team.

7 Project Discussion

Right after the project submission deadline of an examination session and before the next one, there will be an oral discussion session. Note, however, that **not all** the submitted

[1] Central European Time (CET) or Central European Summer Time (CEST), if not otherwise specified.

projects will be automatically accepted for such an oral discussion. In fact, all the submitted projects will be first preliminarily assessed and only those which are considered qualified will be shortlisted for presentation.

The oral session is composed of **two parts**:

- An *oral presentation* (supported by few slides): you will describe the main goal of your project, a comparison of the proposed approaches, and the main results obtained (**max. 20 minutes**)[2];
- A *project demo*: you will be asked to show a working demo of your project by running (part of) your notebook, and to motivate the choices you made.

The whole session will be in english. Questions about any other topic addressed during the course may also be asked, but those can be answered either in english or in italian, as you prefer.

The oral session is public, and therefore *everyone* is welcome to join it!

[*Due to the current restrictions posed by the COVID-19 emergency, oral discussions will take place remotely via Google Meet or Zoom. You will be notified in advance on how to attend the examination.*]

8 Honor Code

The basic principle under which you are expected to operate is that you should submit your own work. Of course, if you are part of a two-member project team this principle will naturally extend to the team as a whole. More specifically, attempting to take credit for someone else's work by turning it in as your own constitutes *plagiarism*, which is a serious violation of basic academic standards.

To observe the honor code, you are invited to follow the rules below:

- You must not submit or look at solutions or program code that are not your own;
- You must not share your solution code with other students, nor ask others to share their solutions with you;
- You must indicate on your submission any assistance you received;

Please, be aware that all submissions are subject to automated plagiarism detection.

As a final remark, many forms of collaboration – if legitimate and properly acknowledged – are acceptable and indeed encouraged.

[2] In case of two-person teams, each member of the team *must* be actively involved in the oral presentation.