Final Project
Machine Learning Applications
Bachelor in Data Science and Engineering
Academic Year 2020/2021

April 16, 2021

1 Introducción

In this project, students will use the knowledge and techniques acquired during the course to solve a machine learning task on text documents. Students can work in groups of a maximum of four people. It is important that, regardless of how each group chooses to distribute the work, all the components of the group know the complete project. The evaluation of the final project will be carried out through the delivery of a report and a group presentation followed by questions.

The project consists of the following tasks:

- Task 1. Natural Language Processing, Topic Modeling and Graphs
- Task 2. Machine Classification or Regression using characteristics extraction or selection techniques
- Task 3. Recommendation Systems

For the execution of the final project, students must choose to implement either task 2 or 3, depending on their preferences and the possibilities of the database used.

2 Data set creation

For the completion of the final project, it is preferred that students work on their own data set. Said data set can be a collection of documents available in the open, or it can be created from the collection of Internet pages. Opting for the second of the proposed alternatives (i.e., creating your own database) will positively affect the final grade of the project.

Bear in mind that for the completion of the final project it will be necessary for your data set to consist of at least several thousand documents in text format, in order to facilitate the construction of sufficiently representative topic models. In addition, each document must have additional metadata that can be used in the representation of the graph, as well as for the implementation of Tasks 2 or 3. For instance, to carry out Task 3 the data set needs to contain rating information or in Task 2 you need any additional variable (categorical or real) to be used as target of a classification or regression task.

In any case, check with one of the professors of the course, once you have decided on the collection of documents to use, to obtain guidance about its viability or possible difficulties in the implementation of the tasks. Make the selection of the data set early, to avoid possible delays that could jeopardize the presentation of the project within the established deadline.

3 Task 1: Text Preprocessing, Topic Modeling and Graph Visualization

This task will consist of the thematic analysis of the collection provided. The steps you must follow in your work are as follows:

- Step 1: Implementation of a pipeline for the preprocessing of the texts. For this task you can use the usual libraries (NLTK, Spacy), or any other library that you consider appropriate.
- Step 2: Extraction of themes and vector representation of the documents using the LDA algorithm.
- Step 3: Calculation of semantic distances between documents, and calculation and representation of a graph (it is recommended to use Gephi)

In the report you must include a description of the preprocessing pipeline used. Likewise, you must describe the topic model obtained, and explain how you have carried out the selection of the number of topics. Finally, you must include at least a screenshot of a graph, and explain how it was generated (i.e., how you calculated the links, the position of the nodes, the color criteria, or any other aspect that you consider relevant).

4 Task 2: Machine Classification or Regression

Implementation and **evaluation** of the performance of a classifier or regression model for the dataset used. Use one of the metadata available in the dataset as your target variable: a categorical variable if you opt for a classification task, or a real type variable for regression. Note that discrete but ordered variables (such as dates, scores, etc.) can also be used as target variables for a regression task.

For this task, you will need to compare the performance by using the TFIDF representation or the document vectors provided by LDA as input variables. In addition, you must use for your work some of the feature extraction or selection algorithms described in the course, analyzing their impact on the results obtained. Use the usual metrics for performance analysis, i.e., error rates, ROC curves, confusion matrices, etc., if you pose a classification task, or the root mean square error if you choose a regression model.

To adjust the hyperparameters of the classification or regression models, you must use a validation methodology that must also be explained in the report.

5 Task 3: Recommender Systems

In this case, you will have to implement a collaborative filtering system where you can explore neighborhood based versions (either user based, content based, or both) or latent based methods such as ALS.

To properly complete this task, you will have to select any or several of the above methods, train them (selecting their parameters adequately), and evaluate its performance. All these steps have to be clearly explained in the report.

Of course, for the implementation and evaluation of these approaches you can use the Surprise library.

6 Guidelines for deliverables

Students must provide the following deliverables for the evaluation of the final project:

1. Descriptive report of the work carried out in .pdf format and a maximum length of 12 pages (excluding only cover and references).

2. Python script with the implemented code duly commented

The report should not include in any case the implemented code, but it should consist of four main sections:

- Task 1 (max. 6 pages)
- Task 2 or 3 (max. 5 pages)
- Code user's manual (max. 1 page).
- Acknowledgment of authorship. Inexcusably, the report must respect the principle of recognition of authorship. If you have used extraneous code snippets or any material from external sources, you must clearly specify this in the report. Failing to do so, may result in the loss of the entire grade for the final project.

7 Grading

The maximum mark of the final project is 4 points, which will be distributed as follows:

- Project execution and documentation (Report): 3 points.
  - Task 1. Natural Language Processing, Topic Modeling and Graphs: $1,75$ points
  - Task 2. Machine Classification or Regression using characteristics extraction or selection techniques: $1,25$ points
  - Task 3. Recommendation Systems: $1,25$ points

  For each of the Tasks (1 and 2/3) the following aspects will be considered:
  - Methodology (45%): methodological correctness of your implementation. This includes the correct application of the methods, but also other aspects, such as normalization, hyperparameter validation, selection of evaluation metrics, strategy for evaluating the performance, etc.
  - Memory quality (40%): the most important aspect that will be assessed is the discussion of your results for which you are encouraged to provide graphical representations supporting your conclusions. Formal presentation will also be taken into account.
  - Code quality (15%): Organization, code efficiency, adequate comments will be taken into account here.
- Virtual presentation: 1 point.

  The assigned time for your presentation will be 10 min. All team member participants should contribute to the presentation, and individual grades could be granted for this item.
- Virtual presentation Q&A.

  Each presentation will be followed by questions by the teachers that can be addressed individually to the different members. The objective of this phase is to verify that all members of each group have participated effectively in the final project, and are aware of its implementation and results in sufficient detail. If during this phase it is found that a member of the team has significant knowledge gaps about the project, the teaching team could apply a penalty factor to that member, which could even result in the total loss of the project's qualification.

All Projects must be handed in via Aula Global. The deadline will be Monday, May 17, at 23:55.

Presentations will take place virtually during the usual class slots of May 18 and May 19. An additional slot will be available during the morning of May 19.

Students that can present using this morning slot are kindly requested to do so, so that other students that may have strong constraints can present during the afternoon sessions.

A doodle will be published soon for the teams to book 20 min slots for project defense.