coreML'23

Twitter Project Description

Prof. Dr. Isabel Valera, Pablo Sanchez-Martin & Batuhan Koyuncu

Summer 2023

1 Introduction

In the following, we describe the practical project that we have prepared for the students of the Machine Learning (ML) core lecture of Summer 2023. The idea of this project is that the students bring into practice the theoretical concepts introduced during the lectures.

This year the project consists on predicting the sentiment of tweets. Such a problem has been mapped into two different ML tasks: (1) a regression task, where the tweet sentiment is represented by a real-valued variable; and (2) a classification task, where the sentiment is represented as a categorical variable taking three different values (positive, negative, and neutral). The evaluation of the Twitter project (for short) is divided into two main parts that contribute differently to your final grade for the course: the report and the challenge. Details on both parts are provided below.

1.1 Participation in the Twitter project

The participation to the Twitter project is fully **voluntary but highly recommended**, as it provides an opportunity for students to gain hands-on experience in ML. Students are expected to work in groups of up 3 members and register their group via CMS (already possible). **Participating in the Twitter project means submitting both the report and the predictions for the challenge**. It is not possible to participate in the challenge but not submitting the report, or the other way around. Refer to Section 5 for the expected timeline.

2 Problem Description

In this project, you will work with a dataset of tweets that have been labeled with both sentiment score and sentiment type. Your task will be to develop machine learning models to predict these labels for new, unseen tweets. The sentiment score is a continuous value between -1 and 1 that represents the degree of positivity or negativity of the tweet. The sentiment type is a categorical value that can be one of three options: positive, negative, or neutral. You will need to develop two models: one for the regression task to predict the sentiment score, and another for the classification task to predict the sentiment type.

2.1 Datasets

We share with you three datasets containing a training set and two test sets. The training set comprises $n_{tr} = 20,000$ tweets, each with a set of features including both the text of the tweets and some additional meta-information about them (e.g., the author id and the number of likes received) and corresponding labels $\mathbf{y} \in \mathbb{R}^2$ for both regression and classification tasks. The test set named `TEST_1` contains $n_{tst1} = 1,000$ tweets without labels and it is already available in CMS. The test set named `TEST_2` contains also $n_{tst2} = 1,000$ tweets without labels and will be available closer to the final submission deadline (due to August 17, 2023). The data have been preprocessed and cleaned, so you can focus on building your machine learning models. Still, you may want to play with different embeddings to represent the text of each tweet $\phi(\mathbf{text})$ (i.e., an alternative representation of the content of the tweets after applying a transformation, similar to a basis function, to the original data) that are obtained using large language models (LLMs), such as BERT [1] from Google.

2.2 Tasks

You will be working on two tasks in this challenge. The first task is to predict the sentiment score of a tweet, which is represented as a real-valued variable. Thus, the first task corresponds to a regression task. The

second task consists on predicting the sentiment type of a tweet. This is a classification task, where you will be predicting one of three categories: 'negative', 'neutral' or 'positive'.

2.3 Jupyter Notebook Example

We have uploaded a Jupyter Notebook in CMS that contains a description of the data and an example on how to load and preprocess the data, visualize it, and use it to train a linear regression and classification model. The notebook also includes instructions on how to format and save your predictions for submission to CMS (both for the leaderboard and the final ranking). We highly recommend that you use this notebook as a starting point for your solution.

3 Report Instructions

Every student group/team participating in the Twitter project should submit a report (.pdf file) using the provided LaTeX template in CMS. The report be **at most 6 pages long (references excluded)** and contain detailed information on the methodology applied to select the final model and make the necessary predictions to participate in the challenge. More specifically, the report should contain information on (at least) the following aspects:

1. **Data analysis & preprocessing:** The report should describe any considered approach (if any) for data analysis and prepocessing used to prepare the input data (features) to the ML model.
2. **ML modeling:** The report should include a short description of the different models (i.e., classifiers and regression models) applied to the data, specifying the used python libraries (if any).
3. **Model selection:** The report should detail the methodology followed to compare the different ML models (and, if applicable, data preprocessing approaches), as well as to select the final model used to make the predictions for the challenge.
4. **Empirical results:** The report should provide a summary and description of the empirical results that have led the students select the final classification and regression models for the challenge.
5. **Others:** The report may contain any additional analysis performed by the students that may be interesting from a practitioner point of view. Examples of such analysis may i) provide a thorough data analysis (for example, data visualization using unsupervised learning techniques); or account for the robustness, explainability or fairness considerations of the different models explored by the students.

3.1 Report grading

The report will contribute to the final grade of each student in the team as:

$$\text{Final grade} = \max(\text{Exam grade};\ 0.75 * \text{Exam grade} + 0.25 * \text{Twitter report grade}),$$

where there are four possible grades for the Twitter report:

- [**0 (out of 10) points**] If a major methodological mistake (e.g., selecting the ML model on the data used to train it) is detected.
- [**5 points**] If a subset of the models introduced in the lectures and tutorials are correctly applied, evaluated and reported.
- [**7.5 points**] If a comprehensive application of the techniques covered in the lectures are correctly applied, evaluated and reported.
- [**10 points**] If the students go one step beyond the course material. The may, e.g., perform an interesting data analysis of the problem (see point 5. above) and/or apply methodology that goes beyond what has been introduced in the lectures (e.g., apply and describe other classification methods beyond the ones covered in the lectures).

4 Challenge

Twitter project will be maintained in a challenge format similar to a Kaggle-like competition. This means you will see your performance and ranking in a leaderboard which will be updated at multiple time stamps during the semester. After each update of the leaderboard you will be able to see your model's performance on the test set and your ranking in the whole competition. In the following part, we will deliver the details about the challenge.

Throughout the semester, you are encouraged to work with your training data, `TWEETS_TRAIN`, to update your models, try different approaches, or perform better model validation and hyper-parameter tuning. Your participation in the challenge will be evaluated in two folds.

1. **Leaderboard:** For the first part of the challenge, we have a **leaderboard** which will be updated multiple times during the semester. Throughout the semester, you can submit your sentiment predictions of the tweets from the test set, `TWEETS_TEST_1`, in CMS. In this leaderboard, you will be able to see the performance evaluation of your team's model and its ranking among the models of other teams (for both prediction tasks). The idea of the leaderboard is twofold: i) give you a realistic estimate of the team raking for each of the tasks; and ii) incentivize healthy competition among teams.

2. **Final Assessment:** For the second and final evaluation of the challenge, you are supposed to submit your sentiment predictions of the tweets from the test set, `TWEETS_TEST_2`, which will be shared with you towards the end of the semester. You need to submit your predictions again through CMS. Your participation in the challenge will be assessed by your model's performance in this step, independently of the performance in the `TWEETS_TEST_1`.

4.1 Performance evaluation

The teams will be ranked for each of the Twitter sentiment analysis tasks (i.e., the classification and regression tasks) based on the following metrics:

- For the regression task

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{true} - y_i^{pred})^2}. \tag{1}$$

- For the classification task

$$\text{(macro) Average f1-score} = \frac{1}{K}\sum_{k=1}^{K} F_1(k). \tag{2}$$

  Above, $F_1(k)$ denotes the f1-score computed for each individual to class k (where precision and recall are computed in a 1-vs-all manner).

The top 5% of the teams–ranked according to the above metrics evaluated on the `TWEETS_TEST_2` dataset–for each of the tasks, will get a bonus (one extra point in the German grading system) on their final grade.

4.2 Submission

To submit your predictions for both the leaderboard and final submission, put your .npy files containing the predictions for both regression and classification tasks in the **same zip file**. Please name files inside the .zip are named as stated at the end of the provided Jupyter notebook.

5 Timeline

In the following, we detail the key dates that should not be missed if interested in joining the Twitter project:

- **Team registration:** via CMS due by 22.05.23. Teams of up to 3 students should be registered in CMS by then. Later changes in a team will only be possible under request.

- **Leaderboard submission:** via CMS due by 09.06.23. Every team interested in participating in the Project should submit their current predictions by this date. This is a necessary condition to participate in the project and allows students to get familiar with the challenge evaluation, as well as get feedback about their ranking.

- **Final project submission** via CMS due by 17.08.23. the students will need to submit both their report and final predictions (on `TWEETS_TEST_2`) by this date. We will not accept any late submissions.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.