Kaggle Competition 2 IFT 6390

November 19, 2021

1 Introduction

This document describes the problem and instructions for the second Kaggle competition. In this case, the goal is to classify locations around the world as crop or non-crop land, using remote sensing (satellites) and meteorological data. Automatic classification of crop lands offers an opportunity to improve the reaction to challenges such as climate change, agriculture and food security.

The data set that we have selected for this competition is a pre-processed subset of CropHarvest (Tseng et al., 2021), a recently published satellite data set with agricultural class labels, conceived to stimulate machine learning research into agriculture, climate change and food security applications. The subset for this competition has over 60,000 data points labelled as *crop* or *non-crop*. Every data point consists of time series of 12 time steps (aggregation of monthly measurements) with multispectral imagery from satellites (Sentinel-1 and Sentinel-2), meteorological data and topographical data. The complete data set contains more locations and annotations, for example specific crop types.

Contrary to the first competition, where the main goal was to implement a logistic regression classifier, the spirit of this competition is more exploratory and therefore you are free to train any algorithm of your choice as well as further process data provided or even extend it. The evaluation will be based on the performance on a held out test set and a *short* written report.

The competition, including the data, is available here, on Kaggle:

https://www.kaggle.com/t/0de158a910d94826827391a189962f1d.

2 Important dates and information

Please take into consideration the following important deadlines:

- **November 26th 23:59** Deadline to enter the competition on Kaggle.
- **December 15th 23:59** Competition ends. No more Kaggle submissions are allowed.
- **December 17th 23:59 Reports** and **code** are due on Gradescope.

2.1 Note on sharing and plagiarism

:

- You *are* allowed to discuss general techniques with other teams.
- You are **not** allowed to share any of your code. This behavior constitutes plagiarism and it is very easy to detect. All teams involved in sharing code will receive a grade of 0 in the competition.
- You are **not** allowed to generate the test labels by means other than a machine learning algorithm trained by yourself or your team, and you are **not** allowed to use test data for training. This behaviour will be considered plagiarism too and it will imply a grade of 0 in the competition. The data set for this competition is publicly available. Therefore, we have implemented measures that would make such practices easy to detect. Note that at the end of the competition you have to submit a piece of code that trains a machine learning algorithm and outputs the list of predictions on the test set.

3 Join the competition

IFT6390 students must do the competition alone (1-person team) (**IFT3395** students will work in teams of 2 or 3).

Create a Kaggle account if you are not registered, and join the competition by following this link:

https://www.kaggle.com/t/0de158a910d94826827391a189962f1d

Important note: The maximum amount of submissions per day and is 2.
Please register your Kaggle username in this form:

https://forms.gle/xhQJWGgmfRc8FfJ58

4 Baselines

The evaluation metric for this competition is the **F1-score**. The F1-score is a commonly used metric for binary classification, and is defined as the harmonic mean of the precision and recall:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} = \frac{TP}{TP + 0.5 \cdot (FP + FN)},$$

where TP is the number true positives (detection), FP the false positives (false alarm) and FN the false negatives (miss).

As in the first competition, your grade will improve as you beat a set of reference baselines that will be visible on Kaggle. These baselines are:

- a dummy classifier that simply predicts the most frequent class in the training set.

- A weak machine learning algorithm.
- A stronger machine learning algorithm.

To participate in the competition, you must provide a list of predicted outputs for the instances on the Kaggle website. You can submit 2 predictions per day over the course of the competition, so we suggest you start early, allowing yourself enough time to submit multiple times and get a sense of how well you are doing.

5 Methods

You are allowed to use any Python libraries of your wish and any machine learning algorithm seen in class, or not. You are also allowed to post-process the data and even extend the data set. The guiding principle is that the algorithm must be implemented by yourself or your team and the predictions must be made by a trained model. Any form of inspection or derivation of the test labels, or using test data for training, is strictly prohibited and will be penalised with a grade of 0, as stated above.

6 Report

In addition to your methods, you have to submit a short report explaining your methods and decisions. The **minimum length** of the report is **2 pages**, and the **maximum length** is 3 pages.

You are free to choose the style and format of the report, but we highly recommend using the template of NeurIPS conference. The structure and content of the report is also free, but we recommend including the following:

- Project title
- Full name, student number and Kaggle username (mandatory)
- Introduction: briefly describe the problem and summarize your approach and results.
- Methods: explain the motivation for the algorithmic choice and describe it.
- Results: present your results and relevant comparisons.
- Discussion: discuss the pros/cons of your approach and methodology and suggest ideas for improvement.
- Statement of Contributions. add the following statement: "I hereby state that all the work presented in this report is that of the author" (mandatory).
- References: very important if you use ideas and methods that you found in some paper or online; it is a matter of academic integrity.

Reports will be evaluated according to clarity and correctness. Reports that convincingly motivate the decisions and methods chosen, clearly describe the approach and present the results in a descriptive and insightful way will be given higher grades.

Submission Instructions

- You must submit the code developed during the project to Gradescope. The code must be well-documented. The code should include a README file containing instructions on how to run the code. Running the code must generate a CSV file with predictions on the test set, as is submitted to Kaggle.
- The report in pdf format must be submitted to Gradescope too.

7 Evaluation Criteria

Marks will be attributed based on the following criteria:

1. You will be assigned points for each one of the 3 baselines that you beat.
2. You will be assigned points depending on your final performance at the end of the competition, given by your ranking in the private leaderboard.
3. You will be assigned points depending on the quality and technical soundness of your final report (see above).

References

Tseng, G., Zvonkov, I., Nakalembe, C. L., and Kerner, H. Cropharvest: A global dataset for crop-type classification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.