Kaggle Competition 1 IFT 6390

October 13, 2021

1 Introduction

For this project, you will take part in a Kaggle competition about detection of extreme weather events from atmospherical data. The goal is to design a machine learning algorithm that can automatically classify whether a set of climate variables corresponding to a time point and location (latitude and longitude) is associated to i) standard (background) conditions, ii) a tropical cyclone or iii) an atmospheric river. Models that are capable of accurately detecting such events are crucial for our understanding of how they may evolve under various climate change scenarios.

The data set that we have prepared for this competition is a relatively small subset of a bigger data set, ClimateNet. The complete data set amounts to nearly 30 GB because it contains climate variables at almost 900,000 locations around the globe. For this competition, we have reduced the number of locations to 120, while keeping the data at all time points. The data set is divided into 47,760 data points for training (data from 1996 to 2010) and 7,320 data points for testing (data from 2011 to 2013). Each data point consists of 16 atmospheric variables such as pressure, temperature and humidity. More details about these variables are available on Kaggle.

The goal of this project is for you to implement and train several classification algorithms. The evaluation will be based on the performance on the held out test set and a written report.

The competition, including the data, is available here:

https://www.kaggle.com/t/ed5755a58be34553a7633bcc2deb8146.

2 Important dates and information

Please take into consideration the following important deadlines:

- **October 8th 23:59** Deadline to enter the competition on Kaggle.

- **October 27th 23:59** Competition ends. No more Kaggle submissions are allowed.

- **November 3rd 23:59** Reports and code are due on Gradescope.

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 data competition."

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/ed5755a58be34553a7633bcc2deb8146

Important note: The maximum amount of submissions per day and is 2.

Please register your Kaggle username in this form: https://forms.gle/CP7Y5PKRWQUBHhHD6

4 Baselines

We ask you to build a **logistic regression** classifier and beat the baselines highlighted in the leaderboard. These baselines are:

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

- a logistic regression classifier.

- the TA's best baseline.

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.

For each of the 3 baselines that you beat, you get extra points.

Your logistic regression classifier must be implemented from scratch. Apart from standard Python libraries, the only libraries allowed are `numpy`, sparse matrices from scipy (`scipy.sparse`), and `pandas`.

5 Other models

You must try at least **1 other model** besides logistic regression, and compare their performances. You are encouraged to implement techniques studied during the course, and look up for other ways to solve this task. Here are a few ideas:

- Support Vector Machines

- Naïve Bayes

- Random Forests

- Hand-crafted features that take into account the nature of the data

The goal is to design the best performing method as measured by submitting predictions for the test set on Kaggle. Your final performance on Kaggle will count as a criterion for evaluation (see Section 6). If a tested model does not perform well, you can still add it in your report and explain why you think it is not appropriate for this task. This kind of discussion is an important feature that we will be using to evaluate your final competition report.

For this part, you are free to use any library of your choice.

6   Report

In addition to your methods, you must write up a report that details the pre-processing, validation, algorithmic, and optimization techniques, as well as providing results that help you compare different methods/models.

The report should contain the following sections and elements:

- Project title

- Full name, student number and Kaggle username.

- Introduction: briefly describe the problem and summarize your approach and results.

- Feature Design: Describe and justify your pre-processing methods, and how you designed and selected your features.

- Algorithms: give an overview of the learning algorithms used without going into too much detail, unless necessary to understand other details.

- Methodology: include any decisions about training/validation split, regularization strategies, any optimization tricks, choice of hyper-parameters, etc.

- Results: present a detailed analysis of your results, including graphs and tables as appropriate. This analysis should be broader than just the Kaggle result: include a short comparison of the most important hyper-parameters and all methods (at least 2) you implemented.

- 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".

- References: very important if you use ideas and methods that you found in some paper or online; it is a matter of academic integrity.

- Appendix (optional). Here you can include additional results, more details of the methods, etc.

You will lose points for not following these guidelines. **The main text of the report should not exceed 6 pages.** References and appendix can be in excess of the 6 pages.

Your must submit your report and your code on Gradescope before **November 3rd 23:59**.

Submission Instructions

- You must submit the code developed during the project. The code must be well-documented. The code should include a README file containing instructions on how to run the code.

- The prediction file containing your predictions on the test set must be submitted online at the Kaggle website.

- The report in pdf format (written according to the general layout described above) and the code should be uploaded on Gradescope.

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).