Machine Learning Challenge

Dear applicant,

Congratulations, you made it to the CeloAI Machine Learning Challenge 🎉!
The required time for the tasks is approximately *3-4 hours*.

Goal

The goal of the challenge is to find out more about your technical capabilities, and we try to answer specific questions that we can't ask in an interview. We also don't want to ask too many technical questions in a face-to-face interview so that we are not personally biased in a potentially stressful situation.

We want to be transparent and give you some insights into what we look at and how we evaluate:

- Clean, simple, and understandable code
- Analytical / problem-understanding / problem solving skills
- Machine learning modeling, data preparation, data quality assurance, and evaluation skills
- Ability to execute and implement machine learning models
- Ability to explain the main design decisions of our solution and why other alternatives were not selected
- Ability to challenge your solution (business wise and technically) and identify potential more efficient solutions

This challenge gives you the possibility to shine and show your best.

Challenge

The challenge is to implement the complete machine learning pipeline that classifies gestures by training a model that consumes time series data.

Data

We use an existing dataset that contains gesture movement data recorded by an external system. The following gestures have been recorded:

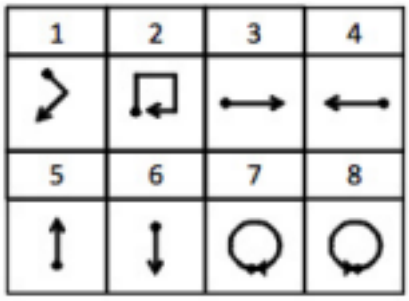

Figure 3: Gesture vocabulary adopted from [KKM+06]. The dot denotes the start and the arrow the end

(taken from uWave: Accelerometer-based personalized gesture recognition and its applications)

The dot denotes the start, and the arrow the end of the gesture. You can download the dataset as a zip-file using the following URL:
http://zhen-wang.appspot.com/rice/files/uwave/uWaveGestureLibrary.zip

If you download and unpack the data, you will get a couple of .rar files. The dataset is structured as follows:

- On the top level, each .rar file includes the gesture samples collected from one user on one day. The .rar files are named as U$userIndex ($dayIndex).rar, where $userIndex is the index of the participant from 1 to 8, and $dayIndex is the index of the day from 1 to 7.
- Inside each .rar file, there are .txt files recording the time series of acceleration of each gesture. The .txt files are named as [somePrefix]$gestureIndex-$repeatIndex.txt, where $gestureIndex is the index of the gesture as in the 8-gesture vocabulary, and $repeatIndex is the index of the repetition of the same gesture pattern from 1 to 10.
- In each .txt file, the first column is the x-axis acceleration, the second y-axis acceleration, and the third z-axis acceleration. The unit of the acceleration data is G, or acceleration of gravity.

Task 1

The first task is to implement a machine learning pipeline, including data preparation, preprocessing, feature extraction, modeling, training, and evaluation to perform gesture detection for the given dataset. That is, given a time-series, classify it in one of the 8 classes. Overall, we are interested in the way you build your machine learning pipeline:

1. We want you to use Python and numpy for all aspects of the "data science workflow" data preprocessing / feature extraction / training / inference. (a few exceptions below).
2. Please implement and train a logistic regression model by hand (plain numpy).
3. (optional) If you want to use other ML-models (neural networks, support vector machines), feel free to use existing libraries (but please, do not invest too much time here).
4. You can (should) use libraries for visualizing the evaluation results or any kind of interesting insights you found during your data exploration phase (whatever you feel makes sense to visualize).

The pipeline that you build should be easily runnable and should automatically run (i.e., a single execution command should be sufficient) all the necessary steps.

Task 2

In the second task we would like you to think about the aspects that would be needed to transform your machine learning pipeline into a modern architecture for a production ready product.
Therefore, please **think** of the following aspects in advance:

- How would you design a devops pipeline using e.g. Github Actions for a Python package? Which functionalities would you include to ensure code quality and consistency?
- Assuming the pipeline you implemented will be deployed as a product. Now the customer also wants to enable real time classification and consume an API that returns the classification results. How would you fit that into the existing architecture?
- The whole system has been a huge success and also other customers want to use it. How would you adapt everything to be able to serve multiple customers with this product? Especially keep in mind scalability and data privacy.
- What would you recommend to automatically transfer machine learning models to production by running microservices for inferencing?

Write down your action plan (max. 1 page) describing the necessary steps you would take to make your machine learning pipeline production ready. You can also consider using other tools or frameworks to simplify development, deployment, and evaluation.

Presentation

Once you have completed the technical part of your challenge, please prepare your content for us in a single zip-file. Please include the following:

- Your complete code and all necessary assets to run it
- A quick README to explain how we run your code
- Your action plan for task 2

During your meeting we would follow the below structure:

1. You walk us through your code and explain the various design decisions you made (max. 30 min)
2. We may ask you some questions regarding your code and design decisions
3. We discuss your action plan to make your machine learning pipeline work in an production environment
4. We ask some general questions about you
5. We answer all your questions that you may have for us