Challenge Data

Solve 2x2x2 Rubik's cube

by LumenAI

currently running

Introduction | Public ranking | Academic ranking | Intermediate academic ranking | Final ranking | Submissions

Submit a solution

Description

LumenAi
Real Time Machine Learning

Dates

From Jan. 1, 2019 to Jan. 1, 2020

Challenge context

In this data challenge, we investigate the so called Rubik's cube, the three dimensional puzzle developed by the Hungarian architect Erno Rubik.

Challenge goals

The goal is to design an automatic Rubik's analyzer that estimates the current length of the shortest path to the solution. Algebraic manipulations of this type could be used in different contexts and solve complex problems. Considering a new unseen configuration on the 2x2x2 Rubik's Cube, the goal of the challenge is to predict the length of the shortest path to the solution.

Presentation of the challenge at the Collège de France

You can find the presentation of the challenge made at the Collège de France here (video in French)

Public metric

mean_absolute_error from scikit-learn.
scikit-learn metrics

Course

Not registered with course

Add a course

Files

x_train
input data of the training set
y_train
output data of the training set
x_test
input data of the testing set

The challenge provider

LumenAi
Real Time Machine Learning

Machine Learning

PROVIDER WEBSITE

Data description

3 datasets are provided as csv files, split between training inputs and outputs, and test inputs. Input files consist of 25 columns separated by commas. The first column, denominated as "ID" in the first line (header) represents a unique sample identifier while the 24 other columns denominated as "pos[i]" in the first line (header) are position features representing each the color of facet [i] of a given configuration of the Rubik's Cube. The color is coded by an integer from 1 to 6. The training input file consists of 1,837079 million examples while the test input file comprises 1,837080 million samples. Output files contain the length of the shortest path to the solution from the input configuration, which corresponds to the minimum number of moves separating the considered input configuration and the solution. They simply consist of two columns denominated ID and distance in the first line (header), where ID refers to the unique identifier of the input sample and distance the length of the shortest path to the solution. Solutions file on test input data submitted by participants shall follow the same format as the training output file but with an ID column containing the test samples IDs and with distances in the distance column which can be float and not necessarily integers. The metric used to compute scores is the Mean Absolute Error.

Benchmark description

The benchmark is based on a clustering and majority vote approach as follows: a clustering of the training input data is performed and a mean is calculated by cluster to predict the test set (this strategy is better than a simple random forest with sklearn without tuning).

I want to cancel my participation