WARA M&L
Leaderboard service

Abstract

WARA Media and Language invites the WASP community to propose leaderboard challenges for machine learning. These challenges are essentially benchmarking tasks, where users submit solution attempts for automatic evaluation with respect to a set of criteria. The outcome is reported as a ranked list, i.e., a leaderboard, which is continiously updated as new solutions come in. Leaderboard challenges can, for example, be used to draw attention to a particular task or evaluation method, or as a teaching tool in ML courses.

1 Introduction

Leaderboard challenges are increasingly used to benchmark problems related to prediction, classification, and synthesis. Companies such as Netflix and Kaggle have organised competitions where participants around the world compete for prize money with their solutions. A challenge typically consists of an ML task, a set of evaluation criteria, and a dataset. The dataset is split into training and test sets, both of which are shared publicly, but the class labels of the test sets are withheld. The participants train their models on the training set, and then use their solutions to predict the missing class labels for the test set, which are submitted for evaluation [1]. In some cases, the test set is kept secret, and the subitted solutions are evaluated on this undisclosed data. A central component in these competitions is the leaderboard, which ranks the participants based on their score. When multiple submissions are allowed, the leaderboard will only display the best result for each user. In this way, the participants can also compete with themselves and try to improve their previous best score.

Many universities are hosting challenges for students to widen their knowledge in ML by offering a convenient environment to pre-process the data, train, validate and test submitted algorithms. Columbia University has, e.g., hosted community competitions to predict protein tertiary structure using deep learning in Kaggle*. At Uppsala University, a simple leaderboard challenge was introduced as part of the course 'Statistical Machine Learning'†: The task consisted in the application of different statistical models to classify songs in Spotify playlist data. By supporting the WASP community in the creation of leaderboard challenges, WARA Media and Language can bring together researchers from a diverse set of fields and and disciplines. Leaderboard challenges can also be used on courses to replace the routine assignments and projects. In this way, students can gain hands-on experience of working with different algorithms to solve real-world problems.

2 Implementation and Pipeline

There are several tasks linked to hosting a leaderboard challenge. These can be technical, e.g., running a web server or preparing dataset, or theoretical, e.g., formulating the problem statement or defining appropriate evaluation metrics. This section describes a pipeline to cover these tasks, using a made-up leaderboard challenge for illustration.

*https://www.kaggle.com/competitions/cu-deep-learning-spring19-hw2/overview
†https://www.it.uu.se/edu/course/homepage/sml/project/submit/

2.1 Problem Statement

The first task is to define a problem statement and motivate its relevance. For example, in the area of Autonomous Vehicles, we find problems such as recognising street signs, warning if the driver is getting drowsy, or automatically veering for pedestrians. Each of these can with a bit of care be translated into a crisp problem statement that lends itself to automatic evaluation.

2.2 Datasets

Few factors are as important as the quantity and quality of the training dataset. If one is lucky, there may be suitable open-source datasets, but if other cases the data needs to be manually collected and possibly annotated. Though data insufficiency might be a major issue, WARA M&L has the resources to provide assistance in data collection and annotation.

2.3 Evaluation Metrics

The next task is to define the metrics that will be used score a solution. This can for example be based on the solutions efficiency or precision, but in the case of generative approaches, on the subjective opinion of a human judge. Common metrics for prediction-based challenges are Mean Squared Error (MSE) and Mean Absolute Error (MAE) are used. Similarly, for classification-based challenges some widely used metrics are Precision, Recall, and F-Score. It is helpful if the metric is so simple that it can be realised as, e.g., a Python script that computes the score based on ground-truth data.

2.4 Implementation and hosting

The final task for online challenges is to build and host a web application. The web application should accept input labels from the user, compare them with the ground-truth labels, evaluate the solution using the chosen metric, and update the leaderboard based on the resulting score.

Alternatively, one may use one of several open platforms leaderboard challenges. Kaggle is a common choise for community challenges, but it requires that the dataset is shared publicly. If this is not possible, then one may instead use EvalAI[‡]. This is an open source machine learning platform to host private machine learning challenges, where only authorized users can upload the predictions. The platform also supports challenges where the participants submit their training code in the form of a docker image instead of the generated predictions.

2.5 Sample Challenges

To host the challenge with WARA ML arena, please provide:

- Problem statement
- Target dataset
- Evaluation metric

[‡]https://eval.ai/

Based on these, the WARA ML committee will decide on how to host the challenge. Table 1 shows an example of the type of information needed.

Challenge Name	Mitosis Detection (Cancer Diagnosis)
Problem Statement	Develop systems to detect mitosis (cell divisions) in histology images.
Dataset	MITOS dataset, **Input x:** RGB histology image, **Output y:** total count and locations of mitosis detections.
Evaluation Metrics	Precision, Recall, and F1 score

Table 1: shows the requirements needed to host Mitosis detection challenge[2]

For the challenge outlined in Table 1, a basic Python web application can be built which accepts the numbers and locations of mitosis detection as input and compute the scores i.e, Precision, Recall, and F1 score using the true values. The application then ranks the predictions based on F1 score. Once the application is ready, it can be hosted in a Amazon EC2 instance and the public URL can be sent to participants to submit their results.

In case the organizer of the challenge does not want to share the test data, the participants can be asked to submit their training code in the form of docker image. WARA M&L can provide support through crowd-sourced evaluators, who can manually run the docker image on the test set and generate the leaderboard for the competition.

References

[1] A. Blum and M. Hardt. "The ladder: A reliable leaderboard for machine learning competitions". In: *International Conference on Machine Learning*. PMLR. 2015, pp. 1006–1014.

[2] D. C. Cireşan et al. "Mitosis detection in breast cancer histology images with deep neural networks". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2013, pp. 411–418.