Compétitions d'apprentissage automatique avec le package R rchallenge

Adrien Todeschini, Robin Genuer

Compétitions d'apprentissage automatique avec le package R rchallenge

Adrien Todeschini [1] & Robin Genuer [2]

[1] *INRIA - IMB - Univ. Bordeaux, 33405 Talence*
Adrien.Todeschini@inria.fr
[2] *ISPED - Univ. Bordeaux - INSERM - INRIA, 33076 Bordeaux*
Robin.Genuer@isped.u-bordeaux2.fr

Résumé. En apprentissage automatique, les performances empiriques obtenues sur données réelles sont déterminantes dans le succès d'une méthode. Ces dernières années ont vu l'apparition d'un grand nombre de compétitions d'apprentissage automatique. Ces challenges sont motivés par des applications industrielles (prix Netflix) ou académiques (challenge HiggsML) et mettent en compétition chercheurs et *data scientists* pour obtenir les meilleures performances. Nous avons souhaité confronter les étudiants à cette réalité en leur soumettant un challenge dans le cadre du cours d'apprentissage automatique. Leur classement est affiché sur une page web mise à jour automatiquement permettant une émulation parmi les étudiants. L'historique des résultats leur permet également de visualiser leur progression au fil des soumissions. De plus, le challenge peut se poursuivre en dehors des sessions encadrées favorisant l'autonomie et l'exploration de nouvelles techniques d'apprentissage et outils informatiques. Le système que nous avons mis en œuvre est disponible sous forme de package R afin d'être réutilisé par d'autres enseignants. S'appuyant sur les outils R Markdown et Dropbox, il ne nécessite aucune configuration réseau et peut être déployé très facilement sur un ordinateur personnel.

Mots-clés. appentissage automatique, compétition, enseignement, package R

Abstract. In machine learning, empirical performance on real data are crucial in the success of a method. Recent years have seen the emergence of a large number of machine learning competitions. These challenges are motivated by industrial (Netflix prize) or academic (HiggsML challenge) applications and put in competition researchers and data scientists to obtain the best performance. We wanted to expose students to this reality by submitting a challenge in the context of the machine learning course. The leaderboard is displayed on an automatically updated web page allowing emulation among students. The history of the results also allows them to visualize their progress through the submissions. In addition, the challenge can continue outside of the supervised sessions promoting independence and exploration of new learning techniques and computer tools. The system we have implemented is available as an R package for reuse by other teachers. Building on the R Markdown and Dropbox tools, it requires no network configuration and can be deployed very easily on a personal computer.

Keywords. machine learning, competition, teaching, R package

1 Introduction

En apprentissage automatique et fouille de données, les performances empiriques obtenues sur données réelles sont déterminantes dans le succès d'une méthode. Dans l'industrie (banque, santé, marketing, défense, etc.), l'apprentissage automatique est utilisé pour la prise de décisions associées à des coûts ou des risques. Il est alors primordial de faire la preuve des bonnes performances dans un contexte réel. Dans le secteur académique, les conférences du domaine mettent également l'accent sur les résultats obtenus sur données réelles dans la sélection des articles conjointement aux aspects théoriques.

Ces dernières années ont vu l'apparition d'un grand nombre de compétitions d'apprentissage automatique. Ces challenges sont motivés par des applications industrielles (prix Netflix [1]) ou académiques (challenge HiggsML [2]) et mettent en compétition chercheurs et *data scientists* pour obtenir les meilleures performances sur un ou plusieurs critères d'évaluation mesurés par exemple sur un ensemble test. Outre le prestige, les compétitions sont parfois récompensées d'un prix. Celui-ci est parfois très important (1M$ pour Netflix) et certaines équipes se regroupent pour partager leurs méthodes et savoir-faire et obtenir de meilleurs résultats. Pour les industriels, l'investissement est intéressant car le travail récolté est le fruit d'une participation collective. Les directions d'exploration sont ainsi démultipliées et peuvent également fusionner à l'image des méthodes d'ensemble en apprentissage telles que les forêts aléatoires.

Récemment, des plateformes de compétition en ligne ont démocratisé le recours à cette forme de *crowdsourcing*. Kaggle [3], leader du domaine, propose des dizaines de challenges internationaux suivis par des milliers de participants. Kaggle est également utilisé dans l'enseignement grâce à une section ouverte aux universités [4]. En France, la plateforme Datascience.net [5] connaît un certain succès depuis 2013.

L'obtention de bonnes performances est une tâche compliquée et pluridisciplinaire faisant intervenir prétraitements, extraction de features, comparaison et sélection de modèles ou méthodes etc. Les plateformes servent alors de vitrine aux jeunes statisticiens et *data scientists* où leurs talents peuvent être attestés objectivement par la mise en compétition.

Ce contexte nous a motivés à proposer notre propre challenge aux étudiants de l'Université de Bordeaux. Les bénéfices de cette approche sont multiples :

— **Professionnalisation** : Le challenge confronte les étudiants à une situation proche des conditions réelles où les critères de sélection sont établis sur un classement. Le projet global allant du prétraitement des données aux prédictions finales permet de développer les compétences requises par le marché de l'emploi.

— **Autonomie** : A l'instar des MOOC (*Massive Online Open Courses*), les étudiants

1. http://www.netflixprize.com
2. http://higgsml.lal.in2p3.fr/
3. http://www.kaggle.com
4. http://inclass.kaggle.com/
5. http://www.datascience.net

sont libres dans leur progression et dans l'organisation de leur temps. Cela encourage l'auto-formation et la collaboration contrairement aux formes plus traditionnelles d'enseignement où le contenu est reçu de manière passive. Cela permet également l'expression des talents dans une discipline qui demande des compétences diverses allant de la théorie mathématique à la maîtrise des outils informatiques.

— **Emulation** : L'aspect compétitif et moderne de l'interface web apporte un côté ludique et stimulant adapté à une génération d'étudiants tournés vers le numérique.

2 Le package R rchallenge

Pour mettre en place notre challenge, nous avons mis en œuvre une solution très simple s'appuyant sur les outils suivants :

— **R Markdown** (Allaire *et al.*, 2015) : offre une syntaxe simplifiée pour mettre en forme des documents contenant à la fois du texte, des instructions R et leurs sorties textuelles ou graphiques. Disponible avec l'environnement de développement RStudio, son édition est très simple et s'apprend très rapidement. Nous l'utilisons pour produire une page html dynamique servant de portail au challenge.

— **Dropbox** [6] : un service de stockage et de partage de copies de fichiers locaux en ligne très populaire. Nous l'utilisons pour récupérer les soumissions des participants et pour héberger la page web.

Cette solution ne requiert aucune configuration réseau, ne dépend d'aucune plateforme externe et peut être installée très facilement sur un ordinateur personnel. Afin de faciliter son déploiement par d'autres enseignants, nous l'avons rendue diponible dans le package R **rchallenge** (Todeschini et Genuer, 2015) disponible sur GitHub [7] et sur le CRAN. L'installation du package s'effectue sous R avec la commande suivante :

```
> install.packages("rchallenge")
```

Une fois le package installé, il suffit de le charger et d'installer un nouveau challenge dans un dossier Dropbox (e.g. `Dropbox/mychallenge`) avec les commandes suivantes :

```
> library("rchallenge")
> setwd("~/Dropbox/mychallenge")
> new_challenge(template = "fr")
```

Un challenge en français [8] prêt à l'emploi est à présent disponible dans le dossier `Dropbox/mychallenge` avec le contenu listé dans le tableau 1.

6. `https://www.dropbox.com/`
7. `https://github.com/adrtod/rchallenge`
8. Pour une page web en anglais, utiliser `new_challenge(template = "en")`

`challenge.rmd`	Script R Markdown pour la page web en français.
`data`	Répertoire contenant les jeux de données `data_train` et `data_test`.
`submissions`	Répertoire pour les soumissions. Il devra contenir un sous-répertoire par équipe dans lequel celles-ci pourront soumettre leurs prédictions. Les sous-répertoires seront partagés avec Dropbox.
`history`	Répertoire où l'historique des soumissions sera stocké.

TABLE 1 – Contenu du dossier après installation d'un nouveau challenge.

Le challenge fourni par défaut est un problème de classification binaire sur le jeu de données *German Credit Card*[9]. Vous pouvez personnaliser le challenge de deux façons :

— *A la création du challenge* : grâce aux options de la fonction `new_challenge`.

— *Après la création du challenge* : en remplaçant manuellement les jeux de données dans le dossier `data` ansi que les prédictions du dossier `submissions/baseline` par vos propres données et en personnalisant le script `challenge.rmd` selon vos besoins.

Enfin, pour compléter l'installation vous devez :

1. Créer et partager un sous-répertoire du dossier `submissions` pour chaque équipe participant au challenge.

```
> new_team("team_foo", "team_bar")
```

2. Publier la page html dans le dossier `Dropbox/Public`[10].

```
> publish()
```

3. Donner le lien public vers le fichier `Dropbox/Public/challenge.html` aux participants.
4. Automatiser la mise à jour.

Pour l'étape 4, sous un système Unix, vous pouvez utiliser l'utilitaire **cron**[11] en ajoutant la ligne suivante à votre table de planification à l'aide de la commande `crontab -e`[12] :
`0 * * * * Rscript -e 'rchallenge::publish("~/Dropbox/mychallenge/challenge.rmd")'`
Cette instruction mettra à jour toutes les heures la page html dans votre dossier `Dropbox/Public` de façon automatique.

9. `https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)`
10. Il est nécessaire d'avoir activé le dossier Public sous Dropbox, voir `https://www.dropbox.com/enable_public_folder`.
11. `http://fr.wikipedia.org/wiki/Cron`
12. Faire attention aux guillemets et apostrophes.

Sous Windows, vous pouvez utiliser le **Planificateur de tâches**[13] pour créer une tâche avec une action *Démarrer un programme* configurée comme suit :

Programme/script : `Rscript.exe`

options : `-e rchallenge::publish('~/Dropbox/mychallenge/challenge.rmd')`

Dès lors, un système de challenge entièrement autonome est mis en place ne nécessitant aucune opération d'administration supplémentaire. A chaque mise à jour, le programme effectue automatiquement les tâches suivantes à l'aide des fonctions disponibles dans notre package listées dans le tableau 2.

`store_new_submissions`	Lecture des fichiers soumis et sauvegarde des nouveaux fichiers dans l'historique.
`print_readerr`	Affichage des erreurs de lecture éventuelles.
`compute_metrics`	Calcul des scores pour chaque soumission de l'historique.
`get_best`	Sauvegarde du meilleur score par équipe.
`print_leaderboard`	Affichage du classement.
`plot_history`	Graphiques d'évolution des scores par équipe.
`plot_activity`	Graphiques d'activité par équipe.

TABLE 2 – Fonctions du package.

Le challenge donné aux étudiants du Master 2 MIMSE de l'Université de Bordeaux utilisant notre package est consultable en ligne[14]. Soulignons enfin le fait que notre package peut être adapté à d'autres types de cours dès lors qu'une soumission peut être évaluée numériquement.

Références

ALLAIRE, J., CHENG, J., XIE, Y., MCPHERSON, J., CHANG, W., ALLEN, J., WICKHAM, H. et HYNDMAN, R. (2015). ***rmarkdown*** *: Dynamic Documents for R.* R package version 0.5.1.

TODESCHINI, A. et GENUER, R. (2015). ***rchallenge*** *: A simple datascience challenge system using R Markdown and Dropbox.* R package version 1.1.

13. `http://windows.microsoft.com/fr-fr/windows/schedule-task`
14. `https://dl.dropboxusercontent.com/u/25867212/challenge_mimse2014.html`