Compétitions d'apprentissage automatique avec le package rchallenge

R. Genuer[a] et **A. Todeschini**[b]

[a]Equipe SISTM
INRIA - ISPED - Univ. Bordeaux
33076 Bordeaux
Robin.Genuer@isped.u-bordeaux2.fr

[b]Equipe CQFD
INRIA - IMB - Univ. Bordeaux
33405 Talence
Adrien.Todeschini@inria.fr

Mots clefs : Appentissage automatique, Compétition, Enseignement

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

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

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 la professionnalisation, l'autonomie ainsi que l'émulation entre étudiants.
Le challenge donné aux étudiants du Master 2 MIMSE de l'Université de Bordeaux utilisant notre package est consultable à l'adresse `http://goo.gl/KRuYn0`. 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.

2 Le package rchallenge

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

— **R Markdown** [1] : 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** [2] hébergé sur Github [7]. Son installation s'effectue sous R avec la commande suivante [8] :

```
> devtools::install_github("adrtod/rchallenge")
```

Références

[1] 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.
[2] Todeschini, A. et Genuer, R. (2015). rchallenge : A simple datascience challenge system using R Markdown and Dropbox. R package version 1.0.
[3] Wickham, H. et Chang, W. (2015). devtools : Tools to Make Developing R Packages Easier. R package version 1.7.0.

6. `https://www.dropbox.com/`
7. `https://github.com/adrtod/rchallenge`
8. nécessite le package **devtools** [3]