Datascience competitions with the R package `rchallenge`

Adrien Todeschini, Robin Genuer

Inria informatics mathematics

université de BORDEAUX

June 25, 2015
4ème Rencontres R, Grenoble

Datascience challenges

- Growing success of machine learning challenges: Netflix ($1M)
- Popularized thanks to: Kaggle

Education side

- Emulation
- Autonomy

Examples in statistical learning courses

Master 2 MIMSE (Adrien); **Master 2 Biostat (Robin)**

- Context: proteomic dataset
- Problem: supervised classification (2 classes: healthy/cancer).
- Dimension:
  - $p = 1000$ variables with 500 fake
  - $n = 100$ observations in training set

    $n = 100$ in test set (with unknown outputs)
- Objectives: prediction of test set outputs and variable selection

`https://dl.dropboxusercontent.com/u/50849929/challenge_fr.html`

Instructions for participants

1. Create and give the name of one Dropbox account to the admninistrator

2. Download datasets

3. Program classifiers with good perfomance.
   Two criteria:
   - prediction error rate
   - bad variable selection rate

4. Submit test prediction files in the shared Dropbox folder

Requirements and set up

- `rmarkdown`
  [Allaire et al., 2015]
- 
- `rchallenge`
  [Todeschini and Genuer, 2015]

Create a new challenge in a Dropbox folder

```
setwd("~/Dropbox/mychallenge")
new_challenge()
```

Add participants

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

Content of the folder

- `challenge.rmd`: R Markdown script of the webpage
- `data`: directory with training, test and quiz data
- `submissions`: directory of submissions (one subdirectory per team)
  - `team_foo`: (Dropbox-) shared with team foo
  - `team_bar`: (Dropbox-) shared with team bar
- `history`: directory of submissions history (one subdirectory per team)

Edit challenge.rmd

Edit metadata, R code chunks and text

````
---
title: "Challenge"
output:
  html_document:
    highlight: tango
    theme: spacelab
    toc: yes
---

```{r echo=FALSE, message=FALSE, warning=FALSE}
library(rchallenge)
data_dir = "data"
deadline = as.POSIXct("2015-09-01 23:59:59")
                                        ...
```

Welcome to the challenge webpage!

# Objectives
**Binary classification**: predict the status of a patient
(cancer v.s. healthy) based on the abundance of proteins.
                                        ...
```
````
````

Publish html page in Dropbox and automate the updates

```
publish()
```

- Your **Public** Dropbox folder must be enabled
- Give public link to your `Dropbox/Public/challenge.html` file to participants.

Automate, using:

- **crontab** on Unix
- **Task Scheduler** on Windows

Automated tasks

- Fully autonomous system is set up (no further administration)
- With each update, the program automatically performs:
  1. `store_new_submissions`
  2. `compute_metrics`
  3. `print_leaderboard`
  4. `plot_history` and `plot_activity`

Classement

Le classement ainsi que les scores affichés sont calculés sur l'ensemble des données test.

Seul le meilleur score par équipe parmi toutes les contributions est retenu.

L'équipe `baseline` correspond au score du meilleur classifieur parmi `predict_all_bad` ou `predict_all_good` qui tient lieu de référence à améliorer.

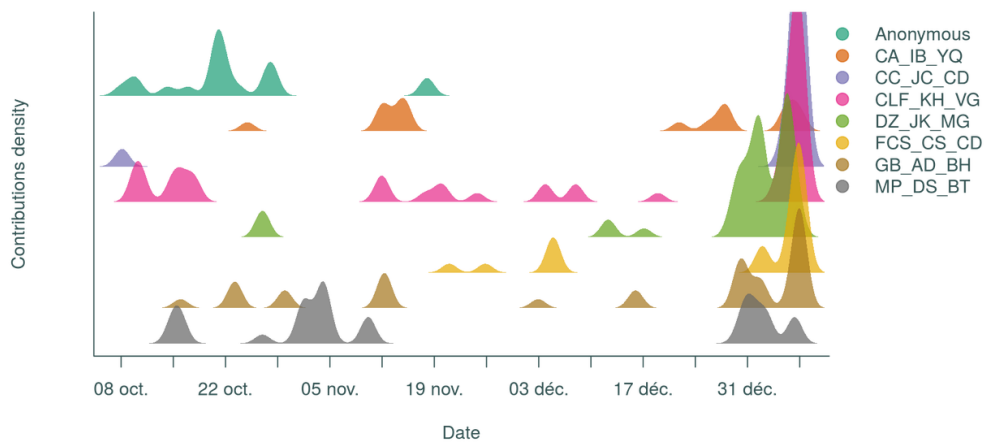
Taux d'erreur

Dernière mise à jour : `08 janv. 2015 00:00`

Rang	Equipe	Contributions	Date	Score
1. →	CLF_KH_VG	56	06/01/15 16:27	0.220
1. →	CC_JC_CD	35	07/01/15 15:38	0.220
3. →	Anonymous	20	18/11/14 16:55	0.225
3. ↑	GB_AD_BH	34	07/01/15 19:50	0.225
5. ↓	MP_DS_BT	34	04/11/14 18:44	0.230
5. ↓	FCS_CS_CD	26	07/01/15 18:40	0.230
7. →	CA_IB_YQ	18	28/12/14 15:17	0.235
8. →	DZ_JK_MG	48	05/01/15 11:35	0.275
9. →	baseline	4	07/10/14 16:28	0.300

Nombre de contributions

Dernière mise à jour : `08 janv. 2015 00:00`



Observations

- Tested for you: it works !
- Autonomous system: almost no administration after the setup
- Evaluation ?
- Frequency of updates ?
  - Low frequency $\Rightarrow$ encourages cross validation but low reward for students
  - High frequency $\Rightarrow$ encourages overfitting but stimulating (immadiate reward)
- Quiz set ?

Strengths/drawbacks

Strenghts

- Simple and effective
- Generalizable to other courses

Drawback

- Computer must be switched on for updates

Future work

- Submit R function and evaluate execution time
- Interactive plots with `ggvis`
- Common leaderboard for several metrics
- Interactive webpage using **Shiny**, without Dropbox

References

Allaire, J., Cheng, J., Xie, Y., McPherson, J., Chang, W., Allen, J., Wickham, H., and Hyndman, R. (2015).

***rmarkdown**: Dynamic Documents for R*.

R package version 0.5.1.

Todeschini, A. and Genuer, R. (2015).

***rchallenge**: A simple datascience challenge system using R Markdown and Dropbox*.

R package version 1.1.

Collaborate via GitHub!

`http://adrtod.github.io/rchallenge`