

Ressources : R et Mathématiques pour les data sciences

Laude

08/10/2020

En guise d'introduction

Philosophie du programme maths/R

C'est un programme exigeant, demandant une attention soutenue et des efforts personnels.

Les concepts qui seront abordés sont **universels et ne seront pas obsolètes avant longtemps**.

Le monde de l'entreprise, les startups, l'innovation et la recherche sont aujourd'hui reliés par un socle mathématique et informatique commun qu'il serait risqué de négliger.

L'important sera de progresser par rapport à votre niveau initial en **décryptage d'expressions mathématiques** et en **"coding"**.

Compétences requises

Les compétences requises seront :

- do not panic !
- savoir mettre de côté ses *a priori*
- disposer de souvenirs diffus concernant des mathématiques de niveau bac + 2
- savoir manipuler un ordinateur sans trop de craintes
- savoir lire, appréhender et tenter d'analyser un texte complexe en conservant une attention soutenue (comme en philosophie)

Conseils précieux

Lisez, relisez, relisez à nouveau de nombreux documents et posez-vous des questions de fond.

Rechercher des réponses sur le net puis posez des questions (mêmes bizarres ou triviales) au formateur sur Discord ou par email.

N'oubliez pas que les deux cours sont en forte interaction et se complètent.

OBJECTIF du cours de maths / Syllabus

- ☐ Apprendre à interpréter les expressions mathématiques présentes dans les papiers traitant du BigData, des Data sciences ou de la Business Intelligence. S'initier à l'analyse critique de documents comportant des aspects mathématiques liés à ces disciplines.
 - ☐ Modalités
 - ☐ Rappel visuel de différents concepts mathématiques [présentation non formelle]
 - ☐ Manipulation très sommaire de Markdown et LaTeX
 - ☐ Focus sur les relations entre l'intelligence artificielle et les probabilités [présentation formelle]

- ☐ Focus sur l'algèbre linéaire et bilinéaire (avec la notion de différentiation) [cours formel]
- ☐ Etude (survol et commentaire) de thèses comportant des aspects mathématiques
- ☐ Début de la préparation du rapport en classe (par les étudiants avec des échanges avec le formateur)
 - ☐ Analyse de documents fournis
 - ☐ Sélection des documents à analyser dans le rapport
 - ☐ Premières tentatives d'analyse
- ☐ création d'une vidéo/pitch de 3 minutes + 1 slide unique, présentés par chaque groupe à la classe en fin de formation (ces vidéos et le slide devront être livrés avec le rapport, avant la dernière séance)
- ☐ De nombreuses ressources sont fournies à l'étudiant
 - ☐ notes de cours complètes
 - ☐ liens vers des ressources externes
 - ☐ Documents (pdf) de références collectés sur le net en licence opensource
- ☐ Evaluation
 - ☐ RAPPORT = contrôle continu
 - ☐ PARTIEL = Savoir chercher dans son cours, bien connaître son propre rapport (documents autorisés et indispensables !)

OBJECTIF du cours R / Syllabus

- ☐ Acquérir les bases de la programmation en R en se focalisant sur les aspects qui seront utilisés en machine learning. Découvrir le biotope de R et apprendre à appréhender les nombreux outils associés. Savoir appréhender un package R inconnu.
 - ☐ Mots clés : Rstudio, packages, vector, matrix, array/tensor, apply, function, ggplot2
 - ☐ Modalités
 - ☐ Installation de Rstudio, manipulation et calculs de base [présentation non formelle]
 - ☐ Manipulation très sommaire de RMarkdown
 - ☐ Manipulations autour du document "premiers pas vers le machine learning ... avec R"[présentation formelle]
 - ☐ Début de la préparation du rapport en classe (par les étudiants avec des échanges avec le formateur)
 - ☐ Analyse de documents et/ou codes fournis
 - ☐ Sélection des packages ou des codes à analyser et manipuler, qui feront l'objet du rapport
 - ☐ Premières tentatives d'élaboration de code sur ces éléments
 - ☐ création d'une vidéo/pitch de 3 minutes + 1 slide unique, présentés par chaque groupe à la classe en fin de formation (ces vidéos et le slide devront être livrés avec le rapport, avant la dernière séance)
 - ☐ De nombreuses ressources sont fournies à l'étudiant
 - ☐ notes de cours complètes
 - ☐ liens vers des ressources externes
 - ☐ Documents (pdf) de références collectés sur le net en licence opensource
- ☐ Evaluation
 - ☐ RAPPORT = contrôle continu
 - ☐ PARTIEL = Savoir chercher dans son cours, bien connaître son propre rapport, savoir coder des expressions basiques en R (documents autorisés et indispensables !)

Dossiers à produire par groupe de 2 ou 3

Evaluation MATHS

Ecrire un rapport en markdown / LaTeX (ou Rmarkdown / LateX) traitant d'aspects mathématiques liés au cours et s'appuyant sur 3 papiers de recherche.

(on livrera le (R)markdown + son résultat en pdf ou word ou html)

Le rapport comportera :

1. une brève synthèse et un commentaire compact de 3 papiers de recherche, sur des sujets reliés à l'IA ou les data sciences ou le Bigdata (mettre les 3 papiers en annexe, moins d'une page par papier pour la synthèse et le commentaire).
2. un zoom sur une ou plusieurs formulations mathématiques pour chacun des papiers, dont on expliquera avec soin la signification mathématique et l'usage qui en a été fait dans le papier. En cas de doute sur la signification mathématique des formulations, on essaiera d'exprimer objectivement la nature de celui-ci.
3. des liens commentés vers un petit nombre de ressources sélectionnées avec soin (Wikipedia ...) et/ou une bibliographie accessible permettant d'appréhender les notions mathématiques en question.
4. un classement et une comparaison motivés des 3 papiers sur divers critères, typiquement :
 - la qualité ou la pertinence de la méthode employée
 - la reproductibilité de la recherche
 - l'originalité du papier
 - la lisibilité du papier (et/ou son aspect didactique)
 - l'intérêt des résultats (pour la communauté/société civile, pour la Recherche, pour les entreprises)
 - l'intérêt de la bibliographie
5. une brève conclusion ouvrant le cas échéant sur de nouvelles perspectives

NB : vous pouvez utiliser des papiers issus de HAL ou Arxiv.org ou de toute autre source librement accessible.

==> livraison des éléments dans le Github du groupe, accompagné d'un email signalant au formateur que le rapport est disponible (comportant la référence du Github et le nom des membres du groupe).

Evaluation R

Vous devrez rédiger un rapport détaillé, via *Rmarkdown*, sur l'usage d'un ou plusieurs *packages R*, ou faire évoluer, franciser et commenter un *code R* trouvé sur un site comportant du code opensource comme Github ou Kaggle (le lien vers les sources est **obligatoire** afin de pouvoir démontrer votre valeur ajoutée).

On livrera : le *Rmarkdown* + les fichiers associés + le résultat de son exécution en pdf **et** html.

Le rapport visera à démontrer l'intérêt de quelques exemples d'utilisation de fonctions bien choisies d'un ou plusieurs packages R de votre choix en motivant succinctement pourquoi vous avez sélectionné ce ou ces packages et certaines fonctions en particulier.

Le rapport comportera du code R commenté (.Rmd) et les résultats de l'exécution de vos exemples. Il sera accompagné de l'ensemble des fichiers permettant de reproduire l'exécution du code.

On portera une attention toute particulière à l'apport didactique du rapport (il faudra soigner la pertinence de l'introduction du rapport et sa lisibilité).

Evidemment, le code ne devra pas plagier les exemples fournis avec la documentation des packages ou paraphraser l'auteur, mais fournir un éclairage didactique personnel illustrant un ou plusieurs aspects de votre choix.

=> livraison des éléments dans le Github du groupe, accompagné d'un email signalant au formateur que le rapport est disponible (comportant la référence du Github et le nom des membres du groupe).

R vs Python

En fait cela relève d'une problématique dépassée, mais regardons quand même les stats d'utilisation des deux langages.

Jul 2020	Jul 2019	Change	Programming Language	Ratings	Change
1	2	▲	C	16.45%	+2.24%
2	1	▼	Java	15.10%	+0.04%
3	3		Python	9.09%	-0.17%
4	4		C++	6.21%	-0.49%
5	5		C#	5.25%	+0.88%
6	6		Visual Basic	5.23%	+1.03%
7	7		JavaScript	2.48%	+0.18%
8	20	▲	R	2.41%	+1.57%
9	8	▼	PHP	1.90%	-0.27%
10	13	▲	Swift	1.43%	+0.31%
11	9	▼	SQL	1.40%	-0.58%
12	16	▲	Go	1.21%	+0.19%
13	12	▼	Assembly language	0.94%	-0.45%
14	19	▲	Perl	0.87%	-0.04%
15	14	▼	MATLAB	0.84%	-0.24%
16	11	▼	Ruby	0.81%	-0.83%
17	30	▲	Scratch	0.72%	+0.35%
18	33	▲	Rust	0.70%	+0.36%

Utilisation générale des langages

L'avis de certains sur la question :

R vs Py en 2020 (<https://mc.ai/battle-of-beasts%E2%80%8A-%E2%80%8Apython-vs-r-in-data-science-2020/>)

Ressources documentaires

Aides mémoire

Aide mémoire R trivial

A avoir sous la main - R (<https://www.duclert.org/>)

Aide mémoire mathématique ... et autres

A avoir sous la main, wolframalpha- Maths (<https://www.wolframalpha.com/>)

Youtube

“fact checker une étude”

sur science étonnante (<https://www.youtube.com/watch?v=NkdczX1Sq-U>)

Ouvrages en ligne : programmation R et un peu de maths ou de Machine Learning

Exploration de données avec R

Les basiques de R pour la BI, simple et en français (<https://bookdown.org/ael/rexplor/>)

Un cours très lisible en français avec des tutoriaux d'étudiants pour s'inspirer

Très complet, de nombreuses techniques (<https://stt4230.rbind.io/>)

Markdown

Les bases de markdown (<https://guides.github.com/features/mastering-markdown/>)

Rmarkdown

simple clair, en français (<http://larmarange.github.io/analyse-R/rmarkdown-les-rapports-automatisees.html>)

L'ouvrage de référence Rmarkdown par son auteur (<https://bookdown.org/yihui/rmarkdown/>)

Un cours R de qualité, qui introduit le Tidyverse

Introduction à R et au tidyverse (<https://juba.github.io/tidyverse/index.html>)

Etudier les séries temporelle, le temps, y compris les Multi TS

plusqu'une introduction ! (https://bookdown.org/singh_pratap_tejendra/intro_time_series_r/)

Behavior Analysis with Machine Learning and R

Un ouvrage simple sur le ML en R (<https://enriquegit.github.io/behavior-free/index.html>)

Tidyverse (une partie dplyr, stringr, tidy)

Manipulations de base en R, dont le Tidyverse (<https://juba.github.io/tidyverse/index.html>)

Un peu plus mathématique ...

Explanatory Model Analysis

Description mathématique à explorer et code R/py (<https://pbiecek.github.io/ema/>)

l'économétrie, les stats ... et leur maths (exemples R)

Introduction to Econometrics with R (<https://www.econometrics-with-r.org/index.html>)

Tips et cookbook

le cours R de référence de monsieur Peng, pour aller directement au but

R pour les data sciences (<https://bookdown.org/rdpeng/rprogdatascience/>)

Trouver rapidement une solution à votre problème de syntaxe

R Cookbook (<https://rc2e.com/index.html>)

Le biotope R, dont GIT, shiny ...

informatique avec R (<https://info201.github.io/index.html>)

Divers sujets R, dont LaTeX

R LaTeX (<https://bookdown.org/Yuleng/polimethods/>)

R Graphic cookbook

graphiques classés (<https://r-graphics.org/>)

Téléchargement de Cheatsheets

RStudio Cheatsheets (<https://rstudio.com/resources/cheatsheets/>)

Thématiques diverses

Unix

The unix workbench, dont GIT (<https://bookdown.org/sean/the-unix-workbench/>)

data science et ligne de commande linux

Ce n'est pas du R, mais cela peut servir ! (<https://www.datascienceatthecommandline.com/1e/>)

Finances quant ...

Analyse technique (<https://bookdown.org/kochiuyu/technical-analysis-with-r-second-edition/>)

Open Quant (<https://bookdown.org/souzatharsis/open-quant-live-book/>)

Séries temporelles

Forecasting (<https://otexts.com/fpp2/>)

Cartographie

Geoprocessing (https://bakaniko.github.io/FOSS4G2019_Geoprocessing_with_R_workshop/)

Text Mining with R

Pour traiter du texte, une approche "tidy" (<https://www.tidyttextmining.com/>)

string et Regex (<https://www.gastonsanchez.com/r4strings/>)

Stats utilisées en psycho socio edu

R for the social scientist (<https://bookdown.org/burak2358/SARP-EN/>)

R data science education (<https://datascienceineducation.com/index.html>)

Stats utilisées dans l'agriculture

Statistical Analysis of Agricultural Experiments, R (<https://rstats4ag.org/>)

Outils

Editeur LaTeX

A essayer absolument (<https://arachnoid.com/latex/index.html>)

Déterminer la meilleure représentation des données

A explorer en profondeur (<https://www.data-to-viz.com/>)

Exemples de graphes ggplot2

R graph gallery (<https://www.r-graph-gallery.com/>)

façonner une dataviz ggplot2 avec esquisse

Esquisse (<https://github.com/dreamRs/esquisse>)

créer des données factices pour tester vos programmes

mockaro (<https://www.mockaroo.com/>)

TIPS

Editeur Vim ou Vi toujours présent sous Linux

petit mode d'emploi des éditeurs vi vim view de linux à mémoriser absolument

pour commencer tapez sur le "i", cela vous met en mode insert
quand vous avez fini appuyez sur "esc" deux fois par précaution, puis sur le ":"
vous pouvez alors sauver en tapant "w" ou "w nom_de_fichier.extension"
puis quitter sans sauver en tapant "q!"

Tester son Rstudio

Créer un fichier test1.py

```
a = 1  
print(a)
```

```
1
```

Stocker ses commandes "console" dans un fichier shell.sh et les tester par copier-coller

linux niveau 0-

```
ls  
whoami  
ls -lta
```

essayer vim

```
vim unfichier.txt  
cat unfichier.txt
```

installer git sur son projet local

```
git init  
git config user.email "henri.laude@ar-p.com"  
git config user.name "henri laude"  
git config --list
```

se préparer à installer des packages python

```
pip3 install --upgrade pip
```

tester l'appel de Python en R

```
library(reticulate)
py_available(initialize = FALSE)
```

```
[1] TRUE
```

```
py_numpy_available(initialize = FALSE)
```

```
[1] TRUE
```

```
a <- 0
reticulate::source_python("test1.py")
print(a)
```

```
[1] 1
```

Contenu d'un fichier de biblio

ma_biblio.bib


```
@book{Laude2018,
abstract = {2e édition. La couv. porte en plus : "Informatique technique" ; "Fichiers complémentaires à télécharger" "Avec cette nouvelle {\e}dition, le livre s'enrichit de nouveaux sujets comme le d{\e}veloppement full-stack avec R (bases de données, processus parall{\e}les, programmation fonctionnelle, API), le partage de r{\e}sultats d'analyse avec R Markdown et les dashboard Shiny, l'{\e}tude des repr{\e}sentations cartographiques et l'impl{\e}mentation de graphes Deep learning avec TensorFlow."--Page 4 de la couverture.},
author = {Laude, Henri. and Laude, Eva.},
edition = {2e {\e}dition },
isbn = {240901397X},
pages = {811},
publisher = {Editions ENI},
title = {{Data scientist et langage R : guide d'autoformation à l'exploitation intelligente des big data}},
year = {2018}
}
```

```
@article{Munier2006,
title = {Comment l'esprit vient aux machines. L'imaginaire de l'objet et de la machine aux débuts de la modernité},
author = {Munier-Temime, Brigitte},
booktitle = {Communication et langages, n°150, 2006. La «valeur» de la médiation littéraire.},
year = {2006},
ISSN = {0336-1500},
url = {https://www.persee.fr/doc/colan_0336-1500_2006_num_150_1_5363},
doi = {10.3406/colan.2006.5363},
language = {fre},
publisher = {Armand Colin},
abstract = {Des objets techniques d'une complexité croissante informent et modifient notre quotidien, nos pratiques et nos usages. Notre téléphone portable, par exemple, aux fonctions sans cesse plus nombreuses, n'acquiert-il pas des allures de couteau suisse ? Sous l'impulsion des nouvelles technologies de l'information et de la communication, l'imaginaire, qui a toujours investi les machines, s'éloigne de l'usine pour envahir le bureau, le sac à main ou la maison. Brigitte Munier relève ainsi qu'à côté d'une conception rationnelle de nos objets, les figures séculaires du génie familial ou de l'apprenti sorcier sont communément mobilisées. Afin de souligner ce rôle de l'imaginaire dans la compréhension des objets simples et techniques, elle interroge la poétisation de l'objet au tout début de la modernité dont nous procédons.}
}
```