Ressources : R et Mathématiques pour les data sciences

Laude

18/10/2020

Contents

1 En guise d'introduction

1.1 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"**.

1.2 Compétences requises

Les compétences requises seront :

- do not panic !
- savoir mettre de coté 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)

1.3 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 par que les deux cours sont en forte interaction et se complètent.

1.4 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 bilinaire (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 videos 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 indisensables !)

1.5 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 apprèhender 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 video/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 indisensables !)

1.6 Dossiers à produire par groupe de 2 ou 3

1.6.1 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 résultat (pour la communauté/société civile, pour la Recherche, pour les entreprises)
- l'intérêt de la bibliographie

5) une brève conclusion ouvrant le cas échant 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).

1.6.2 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).

1.6.3 Mentions complémentaires sur le contrôle continu

- Chaque étudiant doit créer son Github
- Les travaux d'un groupe sont livrés sur les Github de chaque étudiant du groupe
- des travaux individuels seront proposés aux étudiants pendant les cours, le résultat sera livré par les étudiants dans leur Github (.Rmd et .pdf + éventuels fichiers)
- de bons travaux intermédiaires permettront de rattraper d'éventuelles défaillances dans le rendu final des travaux de groupes
- les vidéos et les slides des pitchs seront livrés via un email qui comportera un lien de téléchargement (typiquement un lien WeTransfer)

1.7 R vs Python

En fait cela relève d'une problèmatique dépassée, mais regardons quand même L'avis de certains sur la question :

R vs Py en 2020

On trouve 19500 packages sur le CRAN, et le Github MetaCRAN en donne les sources, ce qui ouvre de grandes perspectives aux utilisateurs :

- modifier et adapter les packages
- mieux comprendre ce qu'ils font
- s'inspirer des meilleures pratiques pour son propre code
- annoter les vignettes (i.e. les manuels) avec ses propres remarques pour se créer ses propres documentations

Par ailleurs, il faut noter :

- que l'on peut créer des programmes mixtes R/Python via le package **reticulate**
- que R possède plusieurs modèles *objet* (de simples à très puissants)
- que la vectorisation est native depuis longtemps dans R et qu'il existe des capacités de parallisation de traitement nativement attachées aux enrionnements R
- que les dataframe ou les "tibble", structures de données courantes en R on des comportements relativement semblables à ceux de la librairie Pandas en Python
- que les deux langages peuvent s'appuyer sur Spark, Tensorflow, une base Relationnelle/SQL ou non relationnelle afin de d'augmenter leurs performances opérationnelles
- que les modèles et les structures de données de machine learning classiques (RNN, CNN, DF…) peuvent être créés dans un langage et exploités dans un autre (voir par exemple l'initiative HDM5)
- que les indices dans les matrices et vecteurs R commencent à 1 … comme en maths, et commencent à 0 en Python

2 Ressources documentaires

2.1 Aides mémoire

Aide mémoire R trivial

A avoir sous la main - R

Aide mémoire mathématique ... et autres

A avoir sous la main, wolframalpha- Maths

2.2 Youtube

2.2.1 "fact checker une étude"

sur science étonnnante

2.3 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

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

Très complet, de nombreuses techniques

Markdown

Les bases de markdown

Rmarkdown

simple clair, en français

L'ouvrage de référence Rmarkdown par son auteur

Un cours R de qualité, qui introduit le Tidyverse

Introduction à R et au tidyverse

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

plusqu'une introduction !

Behavior Analysis with Machine Learning and R

Un ouvrage simple sur le ML en R

Tidyverse (une partie dplyr, stringr, tidy)

Manipulations de base en R, dont le Tidyverse

2.4 Un peu plus mathématique ...

Explanatory Model Analysis

Description mathématique à explorer et code R/py

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

Introduction to Econometrics with R

2.5 Tips et cookbook

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

R pour les data sciences

Trouver rapidement une solution à votre problème de syntaxe

R Cookbook

Le biotope R, dont GIT, shiny ...

informatique avec R

Divers sujets R, dont LaTeX

R LaTeX

R Graphic cookbook

graphiques classés

Téléchargement de Cheatsheets

RStudio Cheatsheets

2.6 Thématiques diverses

Unix

The unix workbench, dont GIT

data science et ligne de commande linux

Ce n'est pas du R, mais cela peut servir !

Finances quant ...

Analyse technique

Open Quant

Séries temporelles

Forecasting

Cartographie

Geoprocessing

Text Mining with R

Pour traiter du texte, une approche "tidy"

string et Regex

Stats utilisées en psycho socio edu

R for the social scientist

R data science education

Stats utilisées dans l'agriculture

Statistical Analysis of Agricultural Experiments, R

3 Outils

Editeur LaTeX

A essayer absolument

Déterminer la meilleure représentation des données

A explorer en profondeur

Exemples de graphes ggplot2

R graph gallery

façonner une dataviz ggplot2 avec esquisse

Esquisse

créer des données factices pour tester vos programmes

mockaro

4 TIPS

4.1 Installation du contexte de travail

4.1.1 Sur sa propre machine :

Installer R à partir du CRAN, puis Rstudio (le cas échéant Pandoc). Pour les machines Windows, installer Rtools qui peut être très utile et permet d'accéder à des commandes Linux utiles lors de certaines opérations particulières.

Certains packages (rares), nécessitent d'avoir installé une version Java de développement et de connaître le PATH de celui-ci.

Une interface Github et un Git local peuvent être utiles.

4.1.2 Sur le cloud

Le site Rstudio propose un Cloud avec une version gratuite très opérationnelle … Cela permet d'être opérationnel rapidement (attention certain knir de Rmarkdown ne fonctinneront pas et les performances ne sont pas garanties, mais c'est quand même très pratique).

4.2 Editeur Vim ou Vi oujours présent sous Linux

```
petit mode d'empoi 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!"
```

4.3 Tester son Rstudio

Créer un fichier test1.py

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

```
1
```

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

4.3.1 linux niveau 0-

```
ls
whoami
ls -lta
```

4.3.2 essayer vim

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

4.3.3 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
```

4.3.4 se préparer à installer des packages python

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

4.3.5 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
```

4.4 A essayer si l'on a des difficultés avec le knit des pdf

Pour pouvoir utiliser la génération de pdf/LaTeX il est probable que vous soyez amené à installer Pandoc sur votre machine.

Pour autant certaines installations posent problème.

Il est parfois judicieux d'essayer l'installation du package **tinytex** puis de procéder comme suit :

- Sortir de Rstudio !
- Entrez à nouveau dans Rstudio
- Passer la commande suivante pour finir l'installation : tinytex::install_tinytex()

Enfin, dans la partie entête de votre Rmd (en YAML), partie output introduire l'appel xelatex, par exemple comme ceci:

```
  pdf_document:
    latex_engine: xelatex
    toc: yes
    toc_depth: 3
    number_sections: true
    fig_height: 3
    fig_width: 5
    fig_caption: true
    df_print: kable
    highlight: tango
    keep_tex: true
```

4.5 Comment créer un commentaire dans un texte Rmarkdown

La syntaxe a privilégier est la suivante :

```
[IMPORTANT]: # (Ce commentaire génial
                ne sera pas présent dans
                le document généré)
```

4.6 Créer une bibliographie

Créer un fichier bibliographie **ma_biblio.bib**, puis l'invoquer dans le YAML du Rmd:

```
bibliography: ma_biblio.bib
```

Un fichier bibliographie ressemble à cela :

```
@book{Laude2018,
abstract = {2e édition. La couv. porte en plus : "Informatique technique" ; "Fichiers complémentaires à
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
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, n
}
```

Dans le texte (la source .Rmd) on introduit une référence de la façon suivante :

```
Henri Laude a exprimé une opinion intéressante [@Laude2018] au sujet du thème traité
par Munier [@Munier2006] sur comment l'esprit vient aux machine.
```

Ce qui donne le rendu de texte suivant, avec les références entre parenthèses :

Henri Laude a exprimé une opinion intéressante (Laude and Laude 2018) au sujet du thème traité par Munier (Munier-Temime 2006) sur comment l'esprit vient aux machine.

… et en fin de document une bibliographie, qui ne comprend **que** les documents référencés dans le corps du document !

La Bibliographie se trouve alors en fin de document

Laude, Henri., and Eva. Laude. 2018. *Data scientist et langage R : guide d'autoformation à l'exploitation intelligente des big data*. 2e édition. Editions ENI.

Munier-Temime, Brigitte. 2006. "Comment L'esprit Vient Aux Machines. L'imaginaire de L'objet et de La Machine Aux Débuts de La Modernité." https://doi.org/10.3406/colan.2006.5363.