Data Sciences for Peace Studies

ECTS : 3

Enseignant responsable : Pavlos DELIAS

Langue du cours : Français

Volume horaire : 18

Description du contenu de l'enseignement :

Introduction (Data Science Basics, Standard Workflow, Roles and Skills, Team Organization Models) **Data 1** (Collection, Sorting, Filtering, Transformation, Tidy data) **Data 2** (Aggregation, Grouping, Summarizing, Relational Data) **Visualization** (Scatterplots, Heatmaps, Maps, Networks, Parameter settings) **Linear Regression 1** (One variable LR, Multiple LR, Understanding the model) **Linear Regression 2** (Correlation and Multicolinearity, Making Predictions) **Classification 1** (Confusion Matrix, ROC curves, Logistic Regression) **Classification 2** (Trees, CART, Random Forests) **Clustering** (Hierarchical clustering, k-means, Recommendation systems) **Text Analytics** (Pre-processing, Bag of Words, Predict Sentiment) Demo: Process Mining – Political Events Analysis Controle Continue

Compétences à acquérir :

- Determine whether the problem is amenable to an analytics solution and reformulate problem statement as an analytics problem
- State the set of assumptions and goals related to the problem
- Understand many different analytics methods, including linear regression, logistic regression, CART, clustering and apply them to find solutions to problems that achieve stated objectives.
- Describe data effectively, predict future outcomes and prescribe decisions using the tools of analytics
- Identify the link between data and models to help make better decisions that add value to individuals, companies and institutions

Pré-requis obligatoires

You will also need a spreadsheet software (MS Excel will make your life easier, but LibreOffice and OpenOffice will do too) and R.

To download R, go to CRAN, the **c**omprehensive **R a**rchive **n**etwork. CRAN is composed of a set of mirror servers distributed around the world and is used to distribute R and R packages. Don't try and pick a mirror that's close to you: instead use the cloud mirror, https://cloud.r-project.org, which automatically figures it out for you[1]

I strongly recommend RStudio. RStudio is an integrated development environment, or IDE, for R programming. Download and install it from http://www.rstudio.com/download. If you don't like installing it on your machine, and you can assure a consistent internet access, you may use the cloud version of it (free too).

[1] Since you download and install R in a location other than the United States, you might encounter some formating issues later in this class due to language differences. To fix this, you will need to type in your R console:

Sys.setlocale("LC_ALL", "C")

Pré-requis recommandés

Basic mathematical knowledge (at a high school level). You should be familiar with concepts like mean, standard deviation, and scatterplots. Mathematical maturity and prior experience with the software will decrease the estimated effort required for this class, but are not necessary to succeed.

Mode de contrôle des connaissances :

50% Controle Continue: It will be an ~1 hr test with closed questions (mostly multiple choice questions). The questions will not demand the use of software, but a basic understanding of how the software tools could contribute will be required. The materials to be included comprise the material of sessions 1-10. **50% Assignment:** Students will have 1 month to prepare a report according to an exercise definition (checklist) that will be delivered during the last session. The definition will point

student to a rich source of data and will outline the basic techniques that should be used to analyze those data. The specific evaluation criteria will be described in the definition, but students should generally expect them to be related to the analytic techniques rather than to the actual solution (analysis) proposed in the report.

Coefficient : 0.50

Bibliographie, lectures recommandées :

The pedagogy of the course is majorly based on the book: Dimitris Bertsimas, Allison O'Hair and Bill Pulleyblank, The Analytics Edge, Dynamic Ideas, 2016. ISBN: 978-0989910897 Another excellent book that describes most of the techniques we will discuss in an intuitive way is: Evans, J. R. (2016). Business analytics. Pearson Higher Ed.[1] A more manager-oriented approach can be found at the (free or donate) book: Caffo, B., Peng, R. D., & Leek, R. H. (2016). Executive data science: A guide to training and managing the best data scientists. Leanpub - https://leanpub.com/eds If you've never programmed before, you might find Hands on Programming with R by Garrett (https://rstudio-education.github.io/hopr/) to be a useful adjunct to this course. If you get stuck in particular with R, start with Google. Typically adding "R" to a query is enough to restrict it to relevant results: if the search isn't useful, it often means that there aren't any R-specific results available. Google is particularly useful for error messages. If you get an error message and you have no idea what it means, try googling it! Chances are that someone else has been confused by it in the past, and there will be help somewhere on the web. If Google doesn't help, try stackoverflow. Start by spending a little time searching for an existing answer, including [R] to restrict your search to questions and answers that use R. [1] Of course, for each technique (Linear Regression, Logistic Regression, Trees, Clustering, etc.) there is a plethora of dedicated textbooks, but their focus is out of scope for this class…