Data analysis and visualisation

Presentation

Learning objectives

To study any type of information, you need to have an overall idea of the distribution of the information and its characteristics. In this context, data visualisation is the starting point for data analysis.

Data analysis and visualisation require concrete tools to make explicit the information on which future decisions are based. This analysis is based on the intuition of the data scientist and is guided throughout the analysis process by visualisation methods that reflect the information present in the data.

There are also many cases where, because of their size, the data cannot be manipulated. In these cases, mastering data reduction is also essential.

In the first part of the course, we propose to study data visualisation and characterisation methods, while defining the conditions for using these techniques. The course begins with an introduction to probability, random variables and the tools provided by descriptive statistics. A good understanding of these concepts is essential and forms the basis of this course. We then present the usual algorithms for dimensionality reduction.

Secondly, we will look at data visualisation, which is the graphical representation of information and data. Using visual elements such as graphs and maps, data visualisation tools make it easy to see and understand trends or unusual values in data.

Increasingly popular with organisations in all sectors, data visualisation is gradually becoming the norm. Its advantages make it the ideal tool to help in decision-making and contribute to the right direction of actions taken. This course aims not only to introduce visualisation techniques, but also to provide guidelines for presenting information correctly.

Description of the programme

Descriptive statistics

PCA (Principal Component Analysis)

MDS (MultiDimensional Scaling) and Isomap

T-SNE (t-distributed Stochastic Neighbor Embedding)

Lle (Locally Linear Embeddings)

Pandas library

Seaborn bookshop

Plotly graphics library

Dash Framework

Heroku Cloud Server

Visualisation and communication

Databases

Generic central skills and knowledge targeted in the discipline

Throughout the course, various applications programmed in python will be designed to enable students to:

* Manipulate descriptive statistical methods and evaluate the hypothesis of interest.
* Handle high-dimensional data by applying dimension reduction methods (in application of the theory proposed in the Mathematics for AI course): PCA, MDS, isomap, t-SNE and lle.
* Understand the theoretical concepts and their implementation in sklearn.
* Select and apply the appropriate statistical tools.
* Design systematic data analysis processes.
* Obtain results from a data set and contextualise them.
* Explain results and conclusions accurately and effectively.
* Implement data analysis pipelines in Python.
* Retrieving and cleaning data
* Processing and analysing data
* Correct communication of data
* Setting up a Dashboard
* Deploying a web analytics application on a server.

How knowledge is tested

Reports at the end of each session.

Bibliography

- Chapitres 2 et 3 de VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data (1st ed.). O'Reilly Media. Acceso gratuito: https://jakevdp.github.io/PythonDataScienceHandbook/

- Chapitre 2 de Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems (2nd ed.). O'Reilly Media.

- Bishop, C. M. (2006), Pattern Recognition and Machine Learning, Information science and statistics, 1st ed. 2006. corr. 2nd printing edn, Springer

- Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. W. McKinney. O'Reilly Media, 1 edition, (Feb 5, 2013).

- Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.

- Plotly Technologies Inc. Title: Collaborative data science Publisher: Plotly Technologies Inc. Place of publication: Montréal, QC Date of publication: 2015 URL: https://plot.ly

- McKinney, W., & others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56). URL: https://pandas.pydata.org/

- Chapitres 2 et 3 de VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data* (1st ed.). O'Reilly Media. URL: https://jakevdp.github.io/PythonDataScienceHandbook/

- Silberschatz, Korth y Sudarshan. Database System Concepts. Mc Graw Hill, Sixth edition (2011).

- Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. doi:10.21105/joss.03021

- https://dash-gallery.plotly.host/Portal/

- https://dash.plotly.com/urls

- https://www.heroku.com/

Teaching team

* Benjamin OCAMPO

Total des heures	**22h**
MN	22h