Analyse et visualisation de données

En bref

Langue de cours: Anglais

Présentation

Objectifs d'apprentissage

Pour étudier tout type d'information, il faut avoir une idée globale de la distribution de l'information et de ses caractéristiques. Dans ce contexte, la visualisation des données est le point de départ pour les analyser.

L'analyse et la visualisation des données nécessitent des outils concrets afin d'expliciter les informations permettant d'asseoir la prise de décisions futures. Cette analyse est basée sur l'intuition du scientifique des données et est guidée tout au long du processus d'analyse, au moyen de méthodes de visualisation qui reflètent les informations présentes dans les données.

Il existe également de nombreux cas pour lesquels, en raison de leur taille, les données ne peuvent pas être manipulées. Dans ces cas, maîtriser la réduction des données est également essentielle.

Dans un premier temps, nous proposons d'étudier les méthodes de visualisation et de caractérisation des données, tout en définissant les conditions d'utilisation de ces techniques. Le cours débute par une introduction aux probabilités, aux variables aléatoires et aux outils fournis par les statistiques descriptives. Une bonne compréhension de ces concepts est essentielle et constitue le fondement de ce cours. Dans un second temps, nous présenterons les algorithmes usuels de réduction de dimensionnalité.

Dans un second temps nous nous intéresserons à la visualisation de données, qui correspond à la représentation graphique d'informations et de données. À l'aide d'éléments visuels comme les graphiques et les cartes, les outils de visualisation de données permettent de voir et de comprendre facilement des tendances ou des valeurs inhabituelles dans les données.

De plus en plus populaire auprès des organisations de tous secteurs, elle se généralise peu à peu. Ses avantages en font l'outil idéal pour aider à la prise de décision et contribuer à la bonne orientation des actions menées. Ce cours vise non seulement à introduire des techniques de visualisation, mais également à fournir des lignes directrices pour présenter correctement les informations.

Description du programme

Statistiques descriptives

ACP(Analyse en composantes principales)

MDS (MultiDimensional Scaling) et Isomap

T-SNE (t-distributed Stochastic Neighbor Embedding)

Lle (Locally Linear Embeddings)

Librairie Pandas

Libraire Seaborn

Librairie graphique Plotly

Framework Dash

Serveur Cloud Heroku

Visualisation et communication

Bases de données

Compétences et connaissances scientifiques et techniques visées dans la discipline

A travers le cours, différentes applications programmées en python seront conçues afin de rendre les étudiants capables de :

- Manipuler des méthodes statistiques descriptives et d'évaluer l'hypothèse d'intérêt.
- Manipuler des données de grande dimension en appliquant des méthodes de réduction de dimension (en application de la théorie proposée dans le cours de mathématiques pour l'IA) : ACP, MDS, isomap, t-SNE et lle.
- Comprendre les concepts théoriques et leur mise en œuvre dans sklearn.
- Sélectionner et appliquer les outils statistiques appropriés.
- Concevoir des processus d'analyse systématique des données.
- Obtenir des résultats à partir d'un ensemble de données et les contextualiser.
- Expliquez les résultats et les conclusions avec justesse et efficacité.
- Implémenter des pipelines d'analyse de données en Python.
- Récupération et nettoyage des données
- Traitement et analyse des données
- Communication correcte des données
- Mise en place d'un dashboard Dash
- Déploiement sur un serveur, d'une application web analytique.

Modalité de contrôle des connaissances

Compte rendu à la fin de chaque séance.

Bibliographie

- 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/

Equipe pédagogique

- Benjamin OCAMPO

Total des heures		**22h**
CM	Cours Magistral	15h
TD	Travaux Dirigés	7h