Data Challenge pour les SHS

Analyse de données et introduction aux méthodes d'apprentissage automatique

---

Lundi 17 Janvier 2022

Data Challenge pour les SHS

Objectifs du cours: Se doter d'outils statistiques et de visualisation pour analyser un jeu de données. Aborder les méthodes modernes d'apprentissage automatique par l'application, et en percevoir les forces et les limites.

Data Challenge pour les SHS

Objectifs du cours: Se doter d'outils statistiques et de visualisation pour analyser un jeu de données. Aborder les méthodes modernes d'apprentissage automatique par l'application, et en percevoir les forces et les limites.

Programme 10 cours + 1 séance d'examen

1. Cours 1: Visualisation de données et statistiques descriptives (`R`)
2. Cours 2 & 3: Réduction de dimension, ACP, et clustering (`R`)
3. Cours 4: Régression (`R`)
4. Cours 5 & 6: Modèles prédictifs (`Python` et `scikit-learn`)
5. Cours 7: Modèles prédictifs avancés dont analyse de texte (`Python`)
6. Puis 3 séances dédiées au projet

Langages utilisés

Nous utiliserons `R` et `Python` pendant ce cours

`R`

- Environnement: RStudio
- Outils utilisés:
  - Visualisation (`ggplot2`)
  - Régression (`glm`)
  - Analyse en composantes principales (`FactoMineR`)

Cours 1, 2, 3, et 4

`Python`

- Environnement: Jupyter
- Outils utilisés:
  - *Machine learning* avec `scikit-learn`)
  - Analyse de texte avec `FastText`

Cours 5, 6 et 7

Intervenants

Julie Josse

Advanced Researcher (Inria)

Gaël Varoquaux

Research director (Inria)

Assistés de Bénédicte Colnet (doctorante) et Lorenzo Gasparollo (ingénieur de recherche).

Comité de pilotage

Julie Josse (Ecole Polytechnique, Inria), Jean-Pierre Nadal (CAMS, CNRS & EHESS), et Gaël Varoquaux (Inria)

Évaluation par projet

Objectifs

Appliquer les méthodes vues en cours, et commencer par des étapes d'exploration, visualisation des données, puis des modélisations en utilisant les méthodes/algorithmes nécessaires pour répondre à la question (en insistant sur les compromis pouvoir prédictif/interprétabilité, la nécessité de toujours se comparer à des méthodes simples, etc).

Sujet

Un jeu de données sera proposé, avec des questions d'exploration de données comme de la visualisation, d'apprentissage statistique, ainsi que des questions d'interprétation.

Rendu

Note de synthèse de une page ainsi qu'une présentation orale (10min).

Informations pratiques

- Language: ??
- Horaires: Lundi 10h-12h. Vacances ??
- Contact: benedicte.colnet@inria.fr
  Ne pas hésiter pour toute question. Nous pourrons aussi mettre en place un slack ou bien une permanence selon les besoins.

- L'évaluation se fait 100% sur le projet.
- Prérequis: Il est fortement conseillé d'effectuer en amont l'installation et la prise en main des outils `R` et `Python` que nous allons utiliser pour les étudiants n'ayant jamais utilisé ces outils. Nous veillerons néanmoins à ce que tous les débutants puissent suivre le cours et en sortir avec les objectifs énoncés ci-avant.