Data Challenge pour les SHS

Analyse de données et introduction aux méthodes d'apprentissage automatique

---

Lundi 18 Janvier 2021

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

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 et introduction au *deep learning* (`Python`)
6. Et des séances dédiées au projet intercallées entre les cours

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

Organisation et intervenants

Intervenants

Les cours seront assurés alternativement par Bénédicte Colnet, Julie Josse, Gaël Varoquaux et les TDs par Bénédicte (doctorante à Inria).

Julie Josse
Advanced Researcher (Inria)

Gaël Varoquaux
Research director (Inria)

Comité de pilotage

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

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

Proposition des différents sujets dans 3 semaines.

> *N.B.: Vous pouvez éventuellement proposer un sujet (en lien avec vos intérêts, ou un autre projet de recherche, ou encore suite à une lecture)! Il vous faudra cependant déjà des données. Dans ce cas contactez nous avec votre proposition.*

Rendu

Présentation orale (10min) et rapport sur les résultats (10 pages)

Informations pratiques

- Language: Français pendant le cours, mais les slides et notebooks seront en anglais.
- Horaires: Lundi 10h-12h, 18 janvier (B. Colnet), 25 janvier (J. Josse), 1 Février (B. Colnet), 8 Février (B. Colnet), 15 Février (G. Varoquaux), 22 Février (G. Varoquaux), [Pas de cours le 1er Mars], 8 Mars (G. Varoquaux), 15 Mars, 22 Mars, 29 Mars (un cours d'économétrie avec Annick Vignes est prévu)
- 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.