Projet du cours d'apprentissage statistique ISUP 3

Benjamin Guedj

28 septembre 2016

Ce projet nécessite **30 heures de travail personnel**.

Dates importantes

Date limite pour proposer un groupe : 26 septembre 2016, 23h59. **Passée !**

Date limite d'envoi des projets : 28 novembre 2016, 23h59, à envoyer à l'adresse benjamin.guedj@inria.fr

Le non-respect de ces dates limites pénalisera votre note.

Instructions

Ce projet est à réaliser en groupe de **5 personnes**.

Vous allez travailler sur des données publiques de votre choix, et explorer à partir de ces données la (ou les) question(s) de votre choix. Votre travail consiste donc à :

- Choisir un jeu de données et une ou des questions à explorer (voir les conseils ci-dessous).
- À l'aide du logiciel R [1], analyser ces données pour essayer de répondre à votre question. Vous êtes encouragés à développer vos propres modèles et méthodes statistiques si nécessaire.
- Rédiger un rapport décrivant votre travail (voir les instructions ci-dessous).

Chaque projet est donc constitué de **deux fichiers** à me rendre :

1. le rapport (sous forme de **fichier pdf**) : je vous conseille fortement de le rédiger en LaTeX [2],

1. https://cran.r-project.org
2. https://www.latex-project.org

2. et votre code R "opérationel", c'est-à-dire qui fonctionne sans intervention de ma part et fournit des résultats complets. Votre code sera annoté de façon à être lisible facilement (description des fonctions, entrées / sorties, etc.).

Choix des données

Vous allez devoir choisir un jeu de données (*dataset*). Si vous avez déjà une question en tête, vous pouvez chercher un dataset sur le web. Par exemple, si vous souhaitez explorer une question liée à la qualité gustative de bières, vous pouvez taper "beer quality public data set" dans votre moteur de recherche préféré.

Vous pouvez également consulter des sites répertoriant des jeux de données libres. Voici une liste non exhaustive :

— https://www.kaggle.com
— https://archive.ics.uci.edu/ml/datasets.html
— https://www.data.gov
— https://data.gov.uk
— https://www.data.gouv.fr/fr/
— http://www.census.gov/data.html
— http://data.europa.eu/euodp/en/data/
— http://www.healthdata.gov
— https://aws.amazon.com/fr/datasets/
— https://www.gapminder.org/data/
— https://www.google.com/trends/explore
— https://www.google.com/finance

Prenez garde à la taille des jeux de données : certains sont très massifs et vous n'arriverez pas à les traiter sur votre laptop (besoin de machines puissantes et d'une programmation appropriée). Si un dataset est trop gros pour que vous l'exploitiez dans son ensemble, vous pouvez n'en conserver qu'une partie (par exemple en ne gardant que 500 films et 10000 users pour le Netflix dataset).

Le rapport

Le rapport doit démontrer votre pleine compréhension des problèmes que vous aurez posés, et proposer des résolutions **rédigées**. Le code R que vous aurez utilisé sera fourni en annexe. Voici une **proposition** de plan :

1. Partie décrivant la problématique que vous avez décidé d'explorer (en la motivant) et le jeu de données sur lequel vous avez choisi d'explorer cette question (en le motivant également). **Important :** n'oubliez pas d'indiquer le lien pour que je puisse récupérer les données.

2. Partie décrivant l'exploration du jeu de données (en justifiant vos choix et interprétant vos résultats)
3. Partie réalisant une synthèse sur votre travail et les réponses / résultats que vous avez obtenus.
4. Annexe contenant le code informatique produit.

Attention : le rapport doit très clairement faire figurer les noms des membres du groupe.

Évaluation

La note finale prendra en compte les éléments suivants :
— L'originalité et la pertinence de la question abordée / du jeu de données choisi.
— La qualité de l'analyse statistique effectuée. Toute approche originale sera considérée positivement.
— La qualité du rapport et du code : explications, commentaires, profondeur d'analyse, mise en perspective, qualité des sorties graphiques, pertinence du choix des résultats montrés, analyse critique des résultats.

Fiches de secours en R et LaTeX

https://cran.r-project.org/doc/contrib/Short-refcard.pdf
https://wch.github.io/latexsheet/latexsheet-a4.pdf