Introduction à la modélisation statistique bayésienne

Un cours avec R, Stan, et brms

Ladislas Nalborczyk

23-06-2022

Qu'est-ce que la modélisation statistique bayésienne ?

L'approche bayésienne consiste à traiter chaque entité (e.g., variables observées, paramètres du modèle, données manquantes) comme des variables aléatoires caractérisées par des distributions de probabilité. Dans une analyse bayésienne, chaque entité inconnue se voit assignée une distribution *a priori* qui représente un état de connaissance avant d'observer les données. Une fois les données observées, le théorème de Bayes est utilisé pour mettre à jour la distribution *a priori* en une distribution *a posteriori*. La distribution *a posteriori* est le but d'une analyse bayésienne et peut être résumée par des valeurs ponctuelles ou des intervalles, et interprétée directement dans un cadre probabiliste cohérent.

Cette approche se différencie –à la fois philosophiquement et en pratique– de l'approche traditionnelle fréquentiste, qui constitue actuellement la majorité des formations proposées. Un des avantages de l'approche bayésienne est qu'elle permet à l'analyste de résoudre des problèmes difficiles voire impossibles à résoudre pour l'approche fréquentiste traditionnelle.

Au fil des exemples proposés, nous réaliserons que même dans des situations de modélisation simples, l'approche bayésienne permet un raisonnement probabiliste plus naturel et plus flexible que la machinerie inférentielle de l'approche fréquentiste. La modélisation statistique bayésienne représente une alternative attirante aux approches fréquentistes en ce qu'elle offre un cadre cohérent à la modélisation statistique. L'approche bayésienne permet de construire et de *fitter* des modèles complexes tout en offrant des conclusions relativement intuitives et qui incorporent toute l'incertitude intrinsèque au processus inférentiel.

Objectifs

L'objectif de cette formation est de vous faire découvrir l'approche bayésienne. Les concepts et outils qui seront présentés tout au long de la formation seront illustrés par des cas concrets d'analyse de données. Ce cours est construit autour du langage `R` et du paquet `brms`, une interface au langage probabiliste `Stan`. Par conséquent, il est *indispensable* d'avoir quelques connaissances élementaires

du langage `R`.

La formation est proposée sous une double étiquette Collège doctoral/MaiMoSiNE (Maison de la Modélisation et de la Simulation) avec une priorité d'accès aux étudiant.e.s du collège doctoral de Grenoble.

Pré-requis

Attention, certains pré-requis sont *indispensables* pour participer à cette formation :

- Être familier avec les concepts de base de la statistique inférentielle (e.g., test d'hypothèse, intervalles de confiance, régression linéaire).
- Connaissances élémentaires en manipulation de données en `R`. Objets et calculs élémentaires en `R`.

Contenu de la formation

Cette formation est composée de dix séances de deux heures durant lesquelles seront dispensées connaissances théoriques et travaux pratiques en `R`, dans l'environnement RStudio.

Planning (20h d'enseignement)

1. **Cours n°01 : Mardi 11 octobre 2022 de 14h à 16h**
   - Interprétations probabilistes, rappels de probabilité
   - Hasard et déterminisme, modèles et théories, théorème de Bayes
2. **Cours n°02 : Jeudi 13 octobre 2022 de 14h à 16h**
   - Modèle beta-binomial
   - Prior, posterior, Bayes factor
3. **Cours n°03 : Mardi 18 octobre 2022 de 14h à 16h**
   - Présentation du package `brms`
   - Modèle de régression linéaire, prédicteurs continus
4. **Cours n°04 : Jeudi 20 octobre 2022 de 14h à 16h**
   - Modèle de régression linéaire (suite)
   - Prédicteurs catégoriels, interactions, taille d'effet
5. **Cours n°05 : Mardi 25 octobre 2022 de 14h à 16h**
   - Markov Chain Monte Carlo
   - Algorithmes, outils diagnostiques
6. **Cours n°06 : Jeudi 27 octobre 2022 de 14h à 16h**
   - Modèle linéaire généralisé (GLM)
   - Régression logistique, link function
7. **Cours n°07 : Mardi 8 novembre 2022 de 14h à 16h**

- Comparaison de modèles
- Régularisation, critères d'information

8. **Cours n°08 : Jeudi 10 novembre 2022 de 14h à 16h**
    - Modèles multi-niveaux (MLMs)
    - Shrinkage, constant effect, varying effect, taille d'effet
9. **Cours n°09 : Mardi 15 novembre 2022 de 14h à 16h**
    - Modèles multi-niveaux généralisés (GMLMs)
    - Régression logistique multi-niveaux, méta-analyse
10. **Cours n°10 : Jeudi 17 novembre 2022 de 14h à 16h**
    - Data hackathon
    - Questions, recommandations