Projet statistiques bayésiennes

L'objectif de ce projet est de proposer une analyse bayésienne des données de votre choix. Il faut penser cet exercice comme un petit résumé du cours par l'exemple. On attend donc

1. Quelques éléments théoriques
2. Quelques algorithmes faits manuellement
3. Un travail personnel (pas de plagiat de code trouvable sur internet !)
4. Une bonne utilisation des ressources bibliographiques
5. Une bonne utilisation des packages de statistiques bayésiennes (non obligatoire)

1 Évaluation

Le barême suivant est très indicatif, mais donne des précisions sur les attendus.

- **(5 points)** Forme : clarté des explications, structure du projet, orthographe, présentation, finition et pertinence des représentations graphiques...
- **(5 points)** Éléments théoriques : rappels de propriétés vues en cours, présentation d'extensions trouvées dans la bibliographie, explications détaillées d'algorithme, justification des choix effectués...
- **(5 points)** Mise en place de l'analyse : programmation de méthodes vues en cours, structure du code, utilisation de packages, interprétations des résultats, pertinence des commentaires, qualité des résultats...
- **(5 points)** Difficultés et qualité de l'analyse : Choix des questions traitées, difficultés des points théoriques discutés, difficultés des points bibliographiques expliqués, difficultés des algorithmes mis en oeuvre, extensions proposées...

2 Propositions (liste non exhaustive) de points à traiter

Le niveau de difficulté est indiquée par des $(*)$

2.1 Questions pratiques

(*) Considérer seulement une variable discrète, et, après avoir choisi un prior adapté, calculer le posterior de la probabilité d'appartenance à chaque espèce (sous hypothèse d'échantillonage aléatoire).

(*) Considérer une seule des variables continues, et calculer le posterior des paramètres de sa distribution.

(*) Formuler une question dans un cadre de théorie de la décision (test, estimation...) et construire la désicion bayésienne associée

(**) Considérer une variable discrète et une variable continue, et calculer le posterior joint.

(**) Considérer une variable discrète et toutes variables continues, et calculer le posterior joint (après avoir proposé un prior adapté).

(**) Présenter un problème de classification supervisé.

(**) Présenter un problème de régression linéaire.

(***) Considérer seulement les variables continues, et traiter le problème comme un problème de classification non supervisé

(***) Proposer plusieurs modèles, et effectuer un choix bayésien de modèle

(****) Documenter et mettre en place des méthodes plus abouties (RJMCMC, Hamiltonian MC), et utiliser des packages dédiés (Jags, Bugs, Stan...)

(****) **Analyse complète** :

Écrire un modèle hiérarchique complet sur toutes les variables (incluant plusieurs modèles). Formuler différentes questions dans un cadre de théorie de la décision. Mettre en place plusieurs algorithmes pour traiter le cas où toutes les variables sont observées, ainsi que les cas où certaines variables ne sont pas observées. Proposer différents tests, estimateurs, classifieurs, et modèles.

2.2 Points théoriques

(*) On pourra montrer la propriété de conjugaison d'une famille de priors, pour la vraisemblance choisie.

(*) On pourra calculer le prior de Jeffreys pour la vraisemblance choisie.

(*) On pourra rappeler les méthodes de construction de tests, ou de choix de modèle, explications pédagogiques et détaillées à l'appui.

(**) On pourra formuler un problème dans un cadre de théorie de la décision, et calculer la décision bayésienne associée.

(**) On pourra documenter d'autres choix de priors non-informatifs, et les calculer pour la vraisemblance choisie.

(***) On pourra documenter (et éventuellement prouver) des théorèmes ou propriétés remarquables non vues en cours (ex : comportement asymptotique du posterior, comparaison entre région de crédibilité et intervalles de confiance). On s'efforcera de les illustrer sur des exemples.

(***) On pourra prouver la convergence des algorithmes utilisés.

3 Remarques générales

Faîtes très **attention au plagiat** qui sera pénalisé par un 0, quelque soit la quantité plagiée. Est considérée comme plagiat :

- Une ou plusieurs phrases copiées collées d'une source, sans guillemets et la source immédiatement citée
- Un ou plusieurs paragraphes reformulés d'une ou plusieurs sources, sans citer dûment les sources dans cette partie en expliquant que l'on s'en inspire
- Un code réutilisé sans citer la source

De façon générale, on vous demande donc de **citer toutes les sources utilisées**, et de le dire explicitement à chaque passage repris littérallement, ou reformulé rapidement.

Attention, la forme compte aussi, donc une relecture attentive est indispensable afin d'éviter des fautes de syntaxe ou d'orthographe.

L'objectif de ce projet est vraiment de mobiliser les connaissances vues en cours afin de mettre en place une analyse de données réelles, en étant capable de justifier théoriquement tout ce qui est fait.

Toute ou une partie de ce travail peut-être mutualisé, mais merci de **préciser clairement** quelles parties ont été faites avec qui. Si tout le projet est fait à deux, vous pouvez rendre un seul rapport. Sinon, même pour une partie mutualisée, merci de rédiger des commentaires individuels, afin de vous approprier le travail.

Au delà de l'évaluation, l'objectif principal de ce projet reste l'apprentissage, et la clarification des notions abordées en cours. Vous pouvez donc commencer par le plus facile et monter en difficulté.

4 Bibliographie (non exhaustive) proposée

[1] R. Christensen, W. Johnson, A. Branscum, and T.E. Hanson. *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians.* Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2011.

[2] Allen B. Downey. *Think Bayes.* O'Reilly Media, Sebastopol, California, 2013.

[3] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, Third Edition.* Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

[4] Kevin P Murphy. *Machine learning: a probabilistic perspective.* Cambridge, MA, 2012.