Simulation stochastique et méthodes bayésiennes pour le traitement du signal

TP1 : Introduction aux Statistiques Bayésiennes
Yann Traonmilin

Ce TP en R consiste en un code à compléter (remplacer la mention `!XXX!` par ce qu'il faut), des paramètres à interpréter et à faire varier et des graphes à étudier.

Télécharger dans le dossier courant le fichier source et les données :

https://drive.google.com/open?id=1XOHOWGpjCotMvI7vI5aZ-IpYLSg3RDPX

https://drive.google.com/open?id=1SVoZT7qVqZgWwLzI4ilzWPquKQ_VVQNA

1 Enquête sur les machines à sous d'un casino

Une machine à sous disposant d'un bouton (le joueur doit payer pour l'utiliser) donne 1EUR avec une probabilité θ et 0 EUR sinon. Tous les casinos annoncent une probabilité de succès de 0.5. Si le casino est honnête on a réellement $\theta = 0.5$, si $\theta < 0.5$ la machine est biaisée et le casino est malhonnête. Un informateur nous prévient que 30% des casinos utilisent en fait des machines à sous avec $\theta = \theta_1 := 0.3$ et 70% avec $\theta = \theta_2 := 0.5$. En testant à plusieurs reprises une machine à sous on cherche à déterminer si le casino est honnête ou pas. Pour cela, on va simuler ce scénario.

1. On modélise les expériences par une suite de v.a. discrète $X = (X_i)_{i=1,n} \in \{0,1\}^n$ i.i.d avec $\mathbb{P}(X_i = 1|\theta) = \theta$ et $\mathbb{P}(X_i = 0|\theta) = 1-\theta$. Proposer un **modèle bayésien** pour inclure l'information a priori donnée par l'informateur.

2. La loi a posteriori $\mathbb{P}(\theta|X = x)$ est uniquement définie par $\mathbb{P}(\theta = \theta_1|X = x)$ (on a forcément $\mathbb{P}(\theta = \theta_2|X = x) = 1 - \mathbb{P}(\theta = \theta_1|X = x)$). Rappeler l'expression de la loi a posteriori $\mathbb{P}(\theta|X)$ (on note k la somme gagnée sur l'ensemble des expériences).

3. Générer une série de 1000 expériences dans le cas d'une machine à sous malhonnête (θ_1) et dans le cas honnête (θ_2).

4. Représenter l'estimateur du maximum de vraisemblance de θ (rappel: la moyenne empirique) en fonction du nombre d'observation (on utilisera un sous-ensemble des expériences générées à la question précédente). Interpréter le résultat.

5. Représenter la loi a posteriori (empirique) en fonction du nombre d'observations dans le cas honnête. Interpréter le résultat.

6. Représenter la loi a posteriori (empirique) en fonction du nombre d'observations dans le cas malhonnête. Interpréter le résultat.

7. Comparer l'approche fréquentiste et bayésienne.

8. On étudie le cas malhonnête. représenter sur un même graphique l'influence de la loi a priori. Choisir une échelle des axes adéquate pour visualiser l'influence de l'a priori. Interpréter le résultat.

9. On recommence l'étude bayésiennes dans une nouvelle ville. Cette fois ci $\theta_1 := 0.4$, $\theta_2 := 0.5$. Un informateur nous préviens que 10% des casinos utilisent en fait des machines à sous avec $\theta = \theta_1 := 0.4$ et 90% avec $\theta = \theta_2 := 0.5$. On dispose de 3 jeux de données X, Y, Z. En faire l'analyse bayésienne précédente. Dans chaque cas, peut-on conclure sur la machine étudiée.

10. Dans l'un des jeux de données précédent, la machine testée avait pour paramètre $\theta = 0.35$, pouvez-vous déterminer lequel à partir de l'analyse bayésienne précédente? En déduire les limites de cette modélisation. Proposer une méthode pour déterminer quelle machine à un paramètre $\theta = 0.35$.

2 Pour aller plus loin

On se donne le modèle bayésien suivant :

$$\begin{aligned} X &\sim \mathcal{N}(\theta, \sigma) \\ \theta &\sim \pi \\ \pi : \mathbb{P}(\theta = \theta_1) &= \mu \\ \mathbb{P}(\theta = \theta_2) &= 1 - \mu \end{aligned} \tag{1}$$

1. Refaire les questions 1 à 8 de la section précédente en choisissant des valeurs pour θ_1, θ_2, μ.

2. Etudier l'influence de σ