UNIVERSITÉ DE LYON

Introduction à l'inférence bayésienne

Cours 1 : axiomatisation

Valentin ROUSSEL

Laboratoire S2HEP, Université de Lyon, France

Octobre 2019

Résumé du cours - Le bayésianisme est une forme d'épistémologie qui connaît un succès croissant dans de nombreux domaines du savoir. En confrontation direct avec le fréquentisme, le bayésianisme suggère de faire usage de l'inférence bayésienne dans le raisonnement scientifique comme d'un critère de démarcation entre la rationalité et l'irrationalité. Tantôt modèle mathématique, tantôt modèle de pensée, le bayésianisme tente de modéliser les formes de croyances en leur attribuant des degrés de crédibilité pouvant prendre des valeurs de 0 (« Je ne crois absolument pas en ... ») à 1 (« J'ai la certitude absolue que ... »). La force du modèle de pensée bayésien est en ceci paradoxale qu'elle est peut-être également sa plus grande faiblesse : son originalité, son anticonformisme à l'inférence fréquentiste – pourtant majoritaire dans les études scientifiques –, et la place prépondérante qu'elle accorde à la subjectivité de l'agent qui s'en réfère. Ce cours en plusieurs parties n'a pas la prétention d'être parfaitement exhaustif ; il propose en revanche une exploration mathématique et épistémologique du concept et des principaux mécanismes à l'œuvre dans la pensée bayésienne.

Mots-clefs - bayésianisme, inférence bayésienne, inférence fréquentiste, probabilités, épistémologie

Introduction

Nous commencerons ce cours par un rappel des axiomes fondamentaux en théorie de probabilités. Cette axiomatisation doit nous permettre de fixer les propriétés que doit vérifier une application $\mathbb{P}$ afin de formaliser l'idée de probabilité. Pour cela, nous commencerons par rappeler les axiomes de Kolmogorov **(Kolmogorov, 1986)** ainsi que les notions d'*espace de probabilité*, de *mesure* et d'*espace mesurable*. Par la suite, nous énoncerons les conséquences des axiomes de Kolmogorov, ce qui nous conduira à la démonstration du théorème de Bayes. Enfin, nous proposerons une courte introduction au théorème de Cox-Jaynes **(Jaynes, 2003)**, et terminerons par la caractérisation du bayésianisme telle qu'énoncée dans le livre d'Isabelle Drouet **(Drouet, 2016)**.

Mesure, espace mesurable et espace de probabilité

Définition - Un *espace mesurable*, ou *espace probabilisable*, est un couple $(\mathcal{X},\mathcal{A})$ où $\mathcal{X}$ est un ensemble et $\mathcal{A}$ une tribu.

Définition - Une tribu sur un ensemble X est un ensemble non vide des parties de X, stable par passage au complémentaire et par union dénombrable.

Définition - Soit $(\mathcal{X},\mathcal{A})$ un espace mesurable, une application μ définie sur $\mathcal{A}$, à valeurs dans $[0,+\infty[$ est appelée *mesure* lorsque les deux propriétés suivantes sont satisfaites :

1. l'ensemble vide a une mesure nulle : $\mu(\varnothing) = 0$
2. l'application μ est $\alpha - additive$: Si E_1, E_2,... est une famille dénombrable de parties de $\mathcal{X}$ appartenant à $\mathcal{A}$, et si ces parties sont deux à deux disjointes, alors la mesure $\mu(E)$ de leur réunion E est égale à la somme des mesures des parties :

$$\mu\left(\bigcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} \mu(E_k)$$

Une partie de X est dite *mesurable* lorsqu'elle appartient à la tribu A.

Définition - Un *espace de probabilités*, ou *espace probabilisé* est un espace construit à partir d'un *espace probabilisable* en le complétant par *une mesure de probabilité*. C'est un triplet $(\Omega, A, \mathbb{P})$ formé d'un ensemble Ω, d'une tribu A sur Ω et d'une mesure $\mathbb{P}$ sur cette tribu telle que $\mathbb{P}(A) = 1$.

Axiomatique de Kolmogorov

Dans l'axiomatique de Kolmogorov, un espace probabilisé est ainsi défini comme un triplet $(\Omega, A, \mathbb{P})$ où :

1. Ω est un ensemble appelé l'*univers*.
2. A est une tribu des parties de Ω, les éléments de A sont appelés *événements*. A est un sous-ensemble des parties de Ω, tel que :
   (a) Il contient Ω.
   (b) Il est stable par passage au complèmentaire.
   (c) Il est stable par réunion dénombrable.
3. $\mathbb{P}$ est une probabilité sur (Ω, A), c'est-à-dire que $\mathbb{P}$ est une application de A dans $[0, 1]$ vérifiant :
   (a) $\mathbb{P}(\Omega) = 1$
   (b) Si (A_n) est une *famille d'événements* 2 à 2 incompatibles (ou *disjoints*), alors :

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} A_n\right) = \sum_{k=1}^{\infty} \mathbb{P}(A_n)$$

Notes de rappel

Considérons deux événements A et B :

Définition - La *réunion* des événements A, B est un événement noté $A \cup B$. $A \cup B$ est réalisé si A est réalisé OU B est réalisé. La réunion des événements A, B est donc l'événement A OU B : "*Tirer un roi ou un valet*".

Définition - L'*intersection* des événements A, B est un événement noté $A \cap B$. $A \cap B$ est réalisé si A est réalisé ET B est réalisé. La réunion des événements A, B est donc l'événement A ET B : "*Tirer un roi et un valet*".

Définition - Les événements A, B sont dits *disjoints* (ou *incompatibles*) s'ils n'ont aucun résultat en commun, c'est-à-dire $A \cap B = 0$. Deux événements disjoints ne peuvent pas être réalisés simultanément puisqu'ils n'ont aucun résultat en commun : si l'un est réalisé, l'autre ne peut pas l'être : "*Tirer un roi*", "*Tirer un valet*".

Définition - Les événements A et $\bar{A}$ sont dits *contraires*. L'événement contraire de A est réalisé lorsque A n'est pas réalisé.

Quelques conséquences et propriétés

Nous ne reviendrons pas ici sur les démonstrations des propriétés énoncées. Soit $\mathbb{P}$ la fonction de probabilité définie précédement, elle vérifie :

1. $\mathbb{P}(\varnothing) = 0$
2. $$\begin{aligned}\mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \\ &\leq \mathbb{P}(A) + \mathbb{P}(B)\end{aligned}$$
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) \Leftrightarrow A$ et B sont disjoints.
4. si $(A_k)_{1 \leqslant k \leqslant n}$ est une famille d'événements 2 à 2 incompatibles, alors :
$$\mathbb{P}\left(\bigcup_{1 \leq k \leq n} A_k\right) = \sum_{1 \leq k \leq n} \mathbb{P}(A_k)$$
5. Si $A \subset B$ alors $\mathbb{P}(A) \leqslant \mathbb{P}(B)$
6. $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A \cap B)$
   "*Probabilité que B se réalise, mais pas A*"
   En particulier si $A \subset B$ alors $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$
7. $\mathbb{P}(\Omega \setminus A) = 1 - \mathbb{P}(A)$
8. $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$
9. $\mathbb{P}(A \cap \bar{B}) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$
10. Si $\Omega = \bigcup_{i=1}^{n} A_i$ n issues élémentaires équiprobables, alors $\mathbb{P}(A_i) = 1/n$
11. Si A est formé de la réunion de k issues élémentaires équiprobables, alors $\mathbb{P}(A) = k/n$

Théorème de Bayes

Définition - Soit A un événement tel que $\mathbb{P}(A) > 0$ et B un second événement. On appelle probabilité de B *conditionnée* par A (ou probabilité de B sachant A) le réel noté $P(B|A)$ tel que :

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(B \cap A)}{P(A)}$$

Théorème de Bayes - Soit A un événement tel que $\mathbb{P}(A) > 0$ et B un second événement. On retiendra la formule de Bayes suivante :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A).\mathbb{P}(A)}{\mathbb{P}(B)}$$

Démonstration - La probabilité de deux évènements A et B simultanés, notée $\mathbb{P}(A \cap B)$, est la probabilité de A fois la probabilité de B sachant A, notée $\mathbb{P}(B|A)$:

$$\mathbb{P}(B \cap A) = \mathbb{P}(A)\mathbb{P}(B|A) \quad (1)$$

De même, la probabilité de deux évènements A et B simultanés, est aussi égale à la probabilité de B fois la probabilité de A sachant B.

$$\mathbb{P}(B \cap A) = \mathbb{P}(B)\mathbb{P}(A|B) \quad (2)$$

Il vient donc :

$$\mathbb{P}(A)\mathbb{P}(B|A) = \mathbb{P}(B)\mathbb{P}(A|B) \quad (3)$$

Et ainsi :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A).\mathbb{P}(A)}{\mathbb{P}(B)} \quad (4)$$

Formules des probabilités totales

Soit un espace probabilisé $(\Omega, A, \mathbb{P})$. Si $(A_i)_{i \in I}$ est un système exhaustif (fini ou dénombrable) d'évènements, et si quel que soit $i \in I$, $\mathbb{P}(A_i) \neq 0$, alors pour tout évènement B :

$$\mathbb{P}(B) = \sum_{i \in I} \mathbb{P}(B|A_i)\mathbb{P}(A_i) = \sum_{i \in I} \mathbb{P}(B \cap A_i) \quad (1)$$

Réécriture et généralisation du Théorème de Bayes

On remarque que :

$$\begin{aligned} \mathbb{P}(B) &= \mathbb{P}(A \cap B) + \mathbb{P}(\bar{A} \cap B) && (2) \\ &= \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|\bar{A})\mathbb{P}(\bar{A}) && (3) \end{aligned}$$

On a ainsi :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|\bar{A})\mathbb{P}(\bar{A})} \quad (4)$$

de (1) et (4) on peut alors exprimer la forme générale du théorème de Bayes pour des variables discrètes :

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_j \mathbb{P}(B|A_j)\mathbb{P}(A_j)} \quad (5)$$

Caractérisation du bayésianisme

Dans cette partie, nous reprendrons point par point les éléments de caractérisation « en trois étages » de la pensée bayésienne, tels que présentés dans l'ouvrage de référence d'Isabelle Drouet **(Drouet, 2016)**. Drouet précise ainsi qu'il est possible d'identifier trois étages de thèses dans « l'édifice bayésien » : les croyances considérées de manière synchroniques, la dynamique des croyances et le raisonnement scientifique.

Croyances considérées de manière synchroniques

1. Les croyances viennent pas degré.
2. Les degrés de croyance d'un agent rationnel sont des probabilités.
3. Par définition, les degrés de croyance d'un agent rationnel satisfont donc les axiomes du calcul des probabilités de Kolmogorov.
4. Un agent rationnel croit les tautologies au degré 1, et conséquemment, les contradictions au degré 0.
5. Certaines probabilités sont des degrés de croyance rationnelle, c'est-à-dire les degrés de croyance d'un individu rationnel. De fait, elles portent également sur l'évaluation et l'expression de l'incertitude.
6. Le point (5) est compatible avec l'existence de probabilités qui ne sont pas des degrés de croyance rationnelle, en particulier avec l'existence de probabilités qui sont des fréquences relatives ou des limites de fréquences relatives.
7. La point (5) est classiquement étendue par les bayésiens de la façon suivante : toutes les probabilités sont des degrés de croyance rationnelle.

Drouet **(Drouet, 2016)** suggère en ce sens que « le bayésianisme classique est un monisme[1] concernant les probabilités ».

Dynamique des croyances

1. Les degrés de croyance rationnelle sont révisés par conditionnalisation bayésienne. Cela signifie qu'un agent rationnel croyant en A au degré $\mathbb{P}(A)$ et qui viendrait à apprendre B, c'est-à-dire ici à croire B au degré 1, devrait en conséquence croire en A au degré $\mathbb{P}(A|B)$. Ceci décrit une révision épistémique globale qui affecte en même temps les degrés de toutes les croyances entretenues par un agent rationnel.

1. Système qui considère l'ensemble des choses comme réductible à un seul principe

2. **Généralisation** - Un agent rationnel dont les degrés de croyance initiaux sont représentés par la distribution de probabilité $\mathbb{P}(\Delta)$ doit, s'il apprend B, réviser ses degrés de croyance de sorte qu'ils soient finalement représentés par la distribution de probabilité $\mathbb{P}(\Delta|B)$.

Raisonnement scientifique

1. Les hypothèses scientifiques peuvent faire l'object de croyances graduées.
2. L'évaluation rationnelle d'une hypothèse scientifique consiste essentiellement à quantifier la croyance dont elle fait l'objet une fois que toutes les informations pertinentes disponibles au moment de l'évaluation ont été prises en compte.
3. L'existence d'une relation de confirmation entre une hypothèse scientifique H et les données empiriques décrites par E dépend essentiellement de $\mathbb{P}(H|E)$.

Théorème de Cox-Jaynes

Le théorème de Cox-Jaynes **(Cox, 1946)** propose une codification et une quantification originale de la démarche d'apprentissage en se fondant sur cinq postulats - ou desiderata - simples.

Ce système induit une interprétation logique des *probabilités*, indépendamment de la notion de *fréquence* et fournit une base rationnelle au mécanisme d'induction logique. Sous les conditions imposées par les postulats, ce théorème postule que toute autre forme de prise en compte des informations dans le cadre de cette représentation particulière de la connaissance serait en fait biaisée.

Les résultats de ce théorème fûrent redécouverts par Jaynes **(Jaynes, 2003)** qui formulat alors une série d'implications pour les méthodes bayésiennes. Nous revenons ici sur les postulats et les règles du théorèmes de Cox. Nous approfondirons les résultats et les conséquences de ce théorèmes dans une partie ultérieure de ce cours.

Axiomes de Cox

1. **Consistance** - S'il existe plusieurs façons de trouver un résultat, elles doivent aboutir au même résultat.
2. **Continuité méthodique** - Le changement de la valeur d'un paramètre ne doit pas contraindre le changement de la méthode de calcul.
3. **Universalité** - Les méthodes de calcul doivent être généralisables et non destinées à un usage particulier.
4. **Spécifications non ambiguës** - Une proposition doit pouvoir être comprise d'une façon et d'une seule
5. **Rétention d'information** - Aucune donnée pertinente ne doit être omise.

Règles

1. **Plausibilité** - Les nombres peuvent représenter des degrés de plausibilité. Il faut pouvoir exprimer quantitativement l'expression de deux plausibilités en fonction de l'une et de l'autre : des plausibilités plus grandes seront représentées par des nombres plus grands ; l'ensemble des nombres réels est adopté pour permettre cette quantification.
2. **Inférence** - Les règles d'inférence ne doivent pas contredire les règles d'inférence communes. Autrement dit, la logique ne doit pas être contredite par le modèle.
3. **Cohérence** - Si une conclusion peut être obtenue par plus d'un moyen, alors tous ces moyens doivent donner le même résultat.
4. **Honnêteté** - L'agent rationnel ne doit pas délibérément ignorer une partie des informations dont il a connaissance et fonder ses conclusions sur le reste : l'agent rationnel est non-idéologique et neutre de point de vue.
5. **Reproductibilité** - L'agent rationnel doit représenter des états de connaissance équivalents par des plausibilités équivalentes.
6. **Somme** - Lorsque deux plausibilités du même état se composent, la plausibilité composée est nécessairement égale ou supérieure à la plus grande des deux.
7. **Produit** - Lorsque deux plausibilités doivent toutes deux être vérifiées pour qu'un état puisse exister, cet état ne peut avoir de plausibilité plus grande que la plus petite des deux précédentes.

Références

Cox, R. T. (1946). Probability, frequency, and reasonable expectation. *American journal of physics*, 14(1) :1–13.

Drouet, I. (2016). *Le bayésianisme aujourd'hui. Fondements et pratiques*, volume 1. Editions Matériologiques, Paris, 1 edition. An optional note.

Jaynes, E. T. (2003). *Probability Theory : The Logic of Science*, volume 1. Cambridge University Press, Cambridge. p. Chapitres 1 to 3.

Kolmogorov, A. N. (1986). *Selected Works of A. N. Kolmogorov. Probability Theory and Mathematical Statistics*, volume 2. Kluwer, Dordrecht.