La formule de Bayes est la meilleure... dans un certain sens

par Jean-Baptiste Hiriart-Urruty [1]

Résumé.
On montre que les probabilités conditionnelles exprimées par la formule de Bayes, dans un contexte de probabilités discrètes, minimisent un critère, de fait une fonction convexe du type Kullback-Leibler, sur l'ensemble de toutes les distributions de probabilités possibles.

Introduction

La formule de Bayes dans un contexte de probabilités discrètes, telle qu'on l'apprend en classe de Terminale de lycées ou dans les premières années d'études post-bac, est assurément l'une des formules de mathématiques les plus intéressantes qui existent. Connue sous différentes appellations (formule de probabilités des causes, d'inversion des conditionnements), sa forme déroutante mais simple est d'une grande applicabilité, ce qui fait qu'une multitude d'exercices, d'études de situations, dans des domaines très variés, peuvent être imaginés. Certains sont allés jusqu'au dithyrambe, tel le mathématicien britannique H. Jeffrey en 1973 : *"Le théorème de Bayes est à la théorie des probabilités ce que le théorème de Pythagore est à la géométrie"*. Rien que ça ! D'autres titres, aussi élogieux, sont apparus récemment dans des documentaires écrits comme oraux ([1, 2]). Pour tester sa popularité, nous avons cherché "Formule de Bayes" via Google sur internet : il en est revenu 124 000 références... L'histoire de la formule [2] est bien connue, sa démonstration à la portée des lycéens. Une observation en passant : la prononciation de Bayes à l'anglaise est *bei(z)*, mais j'avoue que je l'ai toujours entendue et utilisée en *baï(z)...* comme Bayonne.

1. La formule, son explicitation, un exemple

1.1 La probabilité conditionnelle "probabilité de l'événement A sachant l'événement B réalisé", notée $P(A|B)$ (ou $P_B(A)$ par certains auteurs), est par définition $\frac{P(A \text{ et } B)}{P(B)}$ [3] (avec, évidemment, $P(B) > 0$). De même, $P(B|A) = \frac{P(B \text{ et } A)}{P(A)}$. Comme $P(A \text{ et } B)$ est aussi $P(B \text{ et } A)$, il en résulte

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}. \tag{1}$$

1. Université Paul Sabatier de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex 09, jbhu@math.univ-toulouse.fr
https ://www.math.univ-toulouse.fr/~jbhu/

2. Formule esquissée par Thomas Bayes (pasteur presbytérien britannique) en 1748 ; publiée dans un mémoire posthume rassemblé et édité par son ami Richard Price en 1763. Elle a trouvé sa forme définitive plus tard sous la plume du mathématicien français Pierre-Simon Laplace (sans que celui-ci ait eu connaissance du travail de Bayes).

3. Nous utilisons la notation (A et B), resp. (A ou B), pour ce qui peut être aussi noté de manière "ensembliste" ($A \cap B$), resp. ($A \cup B$).

Quelques remarques sur la formule (1) : $P(A)$ est la probabilité *a priori* de A, c'est-à-dire les chances qu'avait l'événement A de se réaliser avant que l'on connaisse l'événement B ; $P(B)$ est la probabilité *a priori* de B, un terme qui quantifie les chances que l'événement B se réalise, sans considérer A. Puis, $P(A|B)$ est la probabilité a *posteriori* (de A sachant B réalisé). On comprend à la vue de la formule (1) qu'elle puisse être appelée "formule d'inversion des conditionnements". $P(.|B)$ est à son tour une mesure de probabilité comme l'était $P(.)$.

1.2 On utilise la formule de Bayes plus souvent sous la forme qui suit. Soit l'ensemble des possibilités Ω "partagé" en n événements $A_1, A_2, ..., A_n$, c'est-à-dire : les A_i sont incompatibles entre eux (l'événement (A_i et A_j) est impossible pour tout $i \neq j$) et les A_i recouvrent toutes les possibilités ((A_1 ou A_2 ou... A_n) $= \Omega$). Considérons un événement B qui résulte de l'une des causes possibles $A_1, A_2, ..., A_n$; alors connaissant les $P(B|A_i)$ et les $P(A_i)$ pour tout $i = 1, 2, ..., n$, il est possible de déterminer la probabilité $P(A_i|B)$ que A_i soit la cause de B. Il suffit de réécrire $P(A_i|B) = \frac{P(B|A_i) \times P(A_i)}{P(B)}$ en intégrant le fait que $P(B) = \sum_{j=1}^{n} P(B \text{ et } A_j)$ (on reconstitue le "morceau" total du puzzle B en ramassant tous les "morceaux" disjoints (B et A_j)) et en décomposant les probabilités $P(B$ et $A_j)$ sous la forme $P(B|A_j) \times P(A_j)$. Cela donne *in fine* :

$$P(A_i|B) = \frac{P(B|A_i) \times P(A_i)}{\sum_{j=1}^{n} P(B|A_j) \times P(A_j)} \text{ pour tout } i = 1, 2, ..., n. \tag{2}$$

Outre le fait qu'elle est aisée à démontrer, cette formule (2) est aussi facile à retenir : on inverse certes les conditionnements mais, surtout, le numérateur est l'un des n éléments constituants du dénominateur. Ainsi, $\sum_{i=1}^{n} P(A_i|B) = 1$.

1.3 Pour illustrer l'utilisation de la formule (2), rien de tel que de l'appliquer à un exemple.

Les stylos à bille d'une école sont fournies par trois fabricants : l'un en fournit 50%, un deuxième 35%, et le troisième 15%. Les stylos, répondant à des contraintes bien précisées à la commande, sont indistinguables à la livraison : en en observant un, il est impossible de savoir de quel fabricant il provient. Les trois fabricants ne sont pas de même qualité : pour le 1^{er} fabricant, il est avéré qu'il fournit 3% de stylos défecteux ; le $2^{ème}$ est un peu moins fiable, en moyenne 6% de ses stylos sont défectueux ; quant au $3^{ème}$, il est heureusement le fournisseur de moindre importance car il est connu que 8% de ses stylos sont défectueux. Le jour de la rentrée, la totalité des stylos est distribuée aux élèves de l'école, et l'un d'entre eux reçoit son stylo et constate immédiatement qu'il ne marche pas... Contre quel fournisseur doit-il se retourner : le premier qui est le plus gros pourvoyeur ? le troisième qui est le moins fiable ? Examinons cela.

Pour $i = 1, 2$ ou 3, soit A_i l'événement "le stylo vient du fournisseur i". Ainsi, $P(A_1) = 0,5$, $P(A_2) = 0,35$, $P(A_3) = 0,15$. Soit B l'événement "le stylo ne marche pas". Les données fournies plus haut nous indiquent que $P(B|A_1) = 0,03$, $P(B|A_2) = 0,06$, $P(B|A_3) = 0,08$. Ce qui nous est utile, maintenant que B est réalisé, est de connaître $P(A_1|B)$, $P(A_2|B)$ et $P(A_3|B)$. La "formule de probabilités des causes" (2), qui porte bien son nom, nous indique :

$$P(A_1|B) = 0,3125 \ , P(A_2|B) = 0,4375 \ , P(A_3|B) = 0,25.$$

Donc, contrairement aux premiers réflexes qu'on pourrait avoir, c'est le $2^{ème}$ fournisseur qu'il faudrait incriminer en premier.

2. Les probabilités a posteriori comme minimiseurs d'un critère de proximité ou de dissimilarité

Les probabilités *a posteriori* exprimées par les formules de (2) sont le résultat de calculs simples, lesquels peuvent apparaître comme un concours de circonstances heureux. Nous montrons ci-dessous qu'elles apparaissent aussi comme minimiseurs d'une fonction convexe sur l'ensemble de toutes les distributions de probabilités possibles. Nous nous sommes inspirés pour cela de l'approche décrite dans [6] pour des distributions à densités [4].

Pour éviter les cas triviaux, nous supposons que $P(B) < 1$ et que $P(A_i \text{ et } B) > 0$ pour tout $i = 1, 2, ..., n$.

Les données du problème, décrit au sous-paragraphe 1.2, sont condensées dans le vecteur

$$(P(A_1 \text{ et } B), P(A_2 \text{ et } B), ..., P(A_n \text{ et } B)).$$

Pour mesurer la proximité ou la dissimilarité avec une distribution de probabilités $p = (p_1, p_2, ..., p_n)$, on fait appel à une fonction du type Kullback-Leibler (voir Annexe), ce qui donne la fonction suivante :

$$p = (p_1, p_2, ..., p_n) \mapsto I(p) = \sum\nolimits_{i=1}^{n} p_i \ln\left(\frac{p_i}{P(A_i \text{ et } B)}\right). \tag{3}$$

On étend par continuité la définition de $x\ln(x)$ à 0 en posant $0\ln(0) = 0$.

Le problème d'optimisation posé est celui de *la minimisation de la fonction I sur l'ensemble*

$$\Pi_n = \left\{p = (p_1, p_2, ..., p_n) : p_i \geqslant 0 \text{ pour tout } i, \text{ et } \sum\nolimits_{i=1}^{n} p_i = 1\right\} \tag{4}$$

de toutes les distributions de probabilités possibles. Ce problème d'optimisation doit être traité comme tous les problèmes d'optimisation, c'est-à-dire en se posant les questions de l'existence de solutions, éventuellement de leur unicité, de leurs caractérisations et, *in fine*, en produisant leurs expressions explicites. La synthèse de ce travail est fournie par le résultat suivant.

Théorème. *La fonction I a un et un seul minimiseur sur* Π_n *et celui-ci est* $(p_1^*, p_2^*, ..., p_n^*)$, *où :*

$$p_i^* = P(A_i|B) = \frac{P(B|A_i) \times P(A_i)}{\sum_{j=1}^{n} P(B|A_j) \times P(A_j)} \text{ pour tout } i = 1, 2, ..., n.$$

De plus, la valeur minimale, c'est-à-dire $\min_{p\in\Pi_n} I(p)$, *vaut exactement* $-\ln(P(B))$.

4. sans être convaincus toutefois du raisonnement qui a conduit à leur construction du critère (concaténation de deux autres critères) ni de la démonstration.

En clair, les probabilités *a posteriori* de la formule de BAYES sont les plus proches, ou les moins dissimilaires, des données condensées en $P(A_1 \text{ et } B), P(A_2 \text{ et } B), ..., P(A_n \text{ et } B)$). Bien sûr, ceci dépend du critère de proximité choisi, d'où l'additif "dans un certain sens" du titre de notre note (voir toutefois notre remarque finale).

Démonstration du Théorème.

- *Existence de minimiseurs.* La fonction

$$p = (p_1, p_2, ..., p_n) \mapsto I(p) = \sum\nolimits_{i=1}^{n} p_i \ln(p_i) - \sum\nolimits_{i=1}^{n} p_i \ln(P(A_i \text{ et } B)) \tag{5}$$

est clairement continue sur Π_n. De plus, sa structure "séparée en les composantes", c'est-à-dire une somme de n fonctions qui, chacune, ne dépend que d'une variable p_i, facilite sa minimisation.

L'ensemble-contrainte Π_n est un ensemble compact convexe [5], en fait un polyèdre convexe contenu dans l'hyperplan d'équation $\sum_{i=1}^{n} p_i = 1$, et dont les sommets sont les points $(0, 0, ..., 1, ..., 0)$ (1 comme $i^{ème}$ composante) pour $i = 1, 2, ..., n$.

Nous sommes donc assurés de l'existence de minimiseurs de I sur Π_n.

- *Unicité de minimiseurs.* Comme la fonction $x\ln(x)$ est convexe, et même strictement convexe, sur $[0, 1]$, la fonction $p = (p_1, p_2, ..., p_n) \mapsto I(p)$ (voir (5) pour son expression décomposée) est strictement convexe sur Π_n. Il n'y a donc qu'un seul minimiseur $(p_1^*, p_2^*, ..., p_n^*)$ de I sur Π_n. Il s'agit maintenant de le trouver.

- *Caractérisation du minimiseur.* Nous sommes en présence d'une fonction convexe différentiable (du moins aux points $p = (p_1, p_2, ..., p_n)$ où $p_i > 0$ pour tout i) à minimiser sur un ensemble-contrainte définie par des égalités (ou inégalités) linéaires. Supposant qu'aucune des composantes p_i^* ne s'annule (ce qui sera vérifié *a posteriori*), nous utilisons la condition nécessaire et suffisante d'optimalité de LAGRANGE applicable dans ce contexte ([4, pages $54 - 57$]) : le vecteur gradient $\nabla I(p^*)$ de I en $p^* = (p_1^*, p_2^*, ..., p_n^*)$ est colinéaire au vecteur $\overrightarrow{\nu} = (1, 1, ..., 1)$ qui est normal (ou orthogonal ici) à l'hyperplan d'équation $\sum_{i=1}^{n} p_i = 1$ (ce vecteur normal étant indépendant du point d'appui p^*). Traduit en termes mathématiques, cela donne : il existe un scalaire λ^* (appelé multiplicateur de LAGRANGE) tel que $\nabla I(p^*) = \lambda^* \overrightarrow{\nu}$, soit, sous forme détaillée,

$$\begin{aligned}
\ln(p_1^*) + 1 - \ln(P(A_i \text{ et } B)) &= \lambda^*, \\
\ln(p_2^*) + 1 - \ln(P(A_i \text{ et } B)) &= \lambda^*, \\
&.... \\
&... \\
\ln(p_n^*) + 1 - \ln(P(A_i \text{ et } B)) &= \lambda^*.
\end{aligned}$$

Sachant que $\sum_{i=1}^{n} p_i^* = 1$ et $\sum_{i=1}^{n} P(A_i \text{ et } B) = P(B)$, des calculs simples, que nous

5. Un petit exercice intéressant consiste à visualiser Π_3 dans $\mathbb{R}^3$. Si les S_i sont les trois points $(1, 0, 0), (0, 1, 0), (0, 0, 1)$ sur les axes, le fait que l'enveloppe convexe est l'ensemble des combinaisons convexes (c'est-à-dire des barycentres à coefficients poistifs ou nuls) joint à l'écriture $(p_1, p_2, p_3) = p_1 S_1 + p_2 S_2 + p_3 S_3$ montre que Π_3 est la partie limitée par le triangle de sommets les S_i.

épargnons au lecteur, conduisent à :

$$\begin{aligned} \lambda^* &= 1 - \ln(P(B)), \\ p_i^* &= \frac{P(A_i \text{ et } B)}{P(B)} = P(A_i|B) \text{ pour tout } i = 1, 2, ...n. \end{aligned}$$

La relation $\nabla I(p^*) - \lambda^* \overrightarrow{\nu} = 0$ exprime que p^* mimimise (partout) la fonction convexe $p \mapsto I(p) - \lambda^*(\sum_{i=1}^n p_i - 1)$. Donc, $I(p) \geqslant I(p^*)$ pour tout $p = (p_1, p_2, ..., p_n)$ vérifiant $\sum_{i=1}^n p_i = 1$. Le point $p^* = (p_1^*, p_2^*, ..., p_n^*)$ ainsi trouvé est bien *le* minimiseur de I sur Π_n.

Quant à la valeur minimale de I sur Π_n, on a immédiatement :

$$I(p^*) = \sum\nolimits_{i=1}^n p_i^* \ln\left(\frac{1}{P(B)}\right) = -\ln(P(B)).$$

Annexe

Les notions d'entropie et d'information (*cf.* §6.6 dans [3] : *Notions d'entropie et d'information*) font la part belle aux fonctions convexes de la variable réelle et aux résultats relevant de l'Optimisation. Nous en rappelons quelques éléments ici.

- Tout d'abord, les fonctions logarithme, exponentielle, ainsi que la fonction particulière $x\ln(x)$ interviennent à tout bout de champ. Certes les fonctions logarithme et exponentielle sont réciproques l'une de l'autre, $x\ln(x)$ est une primitive de $\ln(x)+1$, mais il n'y a pas que cela : les fonctions convexes $-\ln(x)$ et $x\ln(x)$ sont "associées" au sens que nous décrivons ci-dessous.

Soit $\varphi :]0, +\infty[\to \mathbb{R}$ une fonction *convexe* ; on lui associe une nouvelle fonction $\varphi^\diamond :]0, +\infty[\to \mathbb{R}$ définie par $(\varphi^\diamond)(x) = x\varphi(\frac{1}{x})$. Cette transformation $(.)^\diamond$ est une involution, au sens qu'elle est égale à son inverse, ou bien qu'en l'appliquant deux fois de suite on retombe sur ses pieds. Il n'est pas difficile - sans être tout à fait immédiat - de démontrer que $\varphi^\diamond$ *est convexe si et seulement si* φ *est convexe.* Nous dirons que $\varphi^\diamond$ est *l'associée* de φ. Il est clair que $\varphi^\diamond(1) = 0$ dès que $\varphi(1) = 0$. Un premier exemple en est : si $\varphi(x) = -\ln(x)$, alors $(\varphi^\diamond)(x) = x\ln(x)$.

A l'aide de $\varphi :]0, +\infty[\to \mathbb{R}$ convexe et vérifiant $\varphi(1) = 0$, on définit une sorte de mesure de proximité ou de dissimilarité de deux vecteurs $p = (p_1, p_2, ..., p_n)$ et $q = (q_1, q_2, ..., q_n)$ à coordonnées strictement positives par :

$$I_\varphi(p, q) = \sum_{i=1}^n q_i \varphi\left(\frac{p_i}{q_i}\right). \tag{6}$$

D'une manière plus précise, l'utilisation courante de $I_\varphi(p, q)$ est pour deux distributions de probabilités $p = (p_1, p_2, ..., p_n)$ et $q = (q_1, q_2, ..., q_n)$. La fonction I_φ ne définit pas une distance sur $(]0, +\infty[)^n$, elle n'est même pas symétrique. Toutefois, nous avons ([5, Exercice 6.28]) :

$$I_\varphi(q, p) = I_{\varphi^\diamond}(p, q) \ ;$$

$I_\varphi(p,q) \geqslant 0$ si p et q sont des distributions de probabilités.

Donnons quelques exemples de tels φ et I_φ :

- Avec $\varphi(x) = -\ln(x)$, $I_\varphi(p,q) = \sum_{i=1}^n q_i \ln\left(\frac{q_i}{p_i}\right)$.

- Avec $\varphi(x) = x\ln(x)$. Alors, comme déjà noté, $\varphi^\diamond(x) = -\ln(x)$ et $I_\varphi(p,q) = \sum_{i=1}^n p_i \ln\left(\frac{p_i}{q_i}\right)$.

En calcul des probabilités ou théorie de l'information, c'est cet $I_\varphi(p,q)$ qu'on appelle mesure de proximité (ou écart, ou divergence, ou entropie relative) au sens de KULLBACK-LEIBLER [6] de $p = (p_1, p_2, ..., p_n)$ à $q = (q_1, q_2, ..., q_n)$. Dans le cas de distributions de probabilités p et q, $I_\varphi(p,q) = \sum_{i=1}^n p_i \left[\ln(p_i) - \ln(q_i)\right]$ est l'espérance de la différence des logarithmes de p et q , en prenant la distribution de probabilités p pour calculer cette espérance.

- Avec $\varphi(x) = (1-\sqrt{x})^2$. Ici $\varphi^\diamond = \varphi$ et $I_\varphi(p,q) = \sum_{i=1}^n \left(\sqrt{p_i} - \sqrt{q_i}\right)^2$.

- Avec $\varphi(x) = (1-x)^2$. Alors $\varphi^\diamond(x) = x + \frac{1}{x} - 2$ et $I_\varphi(p,q) = \sum_{i=1}^n \frac{(p_i - q_i)^2}{q_i}$.

- Avec $\varphi(x) = |x-1|$. Ici $\varphi^\diamond = \varphi$ et $I_\varphi(p,q) = \sum_{i=1}^n |p_i - q_i|$.

Remarque finale. Le travail de minimisation que nous avons conduit avec l'une des fonctions φ peut l'être également avec les quatre autres fonctions présentées dans l'Annexe ci-dessus. A nouveau, comme dans le Théorème, le seul minimiseur de I_φ est $(P(A_1|B), P(A_2|B), ... P(A_n|B))$; seules les valeurs optimales $\min_{p\in\Pi_n} I_\varphi(p)$ diffèrent.

Remerciements. Je remercie AGNÈS LAGNOUX de l'université JEAN JAURÈS (Toulouse II) des discussions sur le sujet, montrant jusqu'où on pouvait aller ou ne pas aller.

Références

1. *Une formule mathématique universelle existe-t-elle ?* Forum Science publique du 9 novembre 2012, émission radiophonique de 57 mn sur France Culture.

2. *La formule qui décrypte le monde.* Dossier de Science et Vie n°1142 (2012).

3. Y. CAUMEL, *Probabilités et processus stochastiques.* Collection Statistiques et probabilités appliquées. Springer-Verlag France (2011).

4. J.-B. HIRIART-URRUTY, *Les mathématiques du mieux faire ; vol. 1 : Premiers pas en optimisation.* Collection Opuscules, Edition Ellipses (2008).

5. J.-B. HIRIART-URRUTY, *Optimisation et Analyse convexe. Exercices corrigés.* Collection Enseignement Sup-Mathématiques, EDP Sciences (2009).

6. TAN BUI-THANH, *The optimality of Bayes'* theorem. Newsjournal of the Society for Industrial and Applied Mathematics, Vol. 54, Issue 6 (July/August 2021).

6. SALOMON KULLBACK (1907−1994) et RICHARD LEIBLER (1914−2003) sont deux mathématiciens et cryptologues américains. Leur définition d'entropie relative est publiée pour la première fois dans l'article suivant :

S. Kullback, R. Leibler, *On information and sufficiency,* Annals of Mathematical Statistics $22, 79-86$ (1951).

*Cette note est dédiée à la mémoire d'*Yves Caumel (1947 – 2018)*, professeur de mathématiques appliquées à l'ISAE-ENSICA de Toulouse, un bon ami avec qui nous ne discutions pas seulement de mathématiques.*

"Yves Caumel était docteur en mathématiques, diplômé en histoire et philosophie des sciences. Après une expérience industrielle dans les domaines de la recherche et de la formation, il avait intégré l'ENSICA à Toulouse en 1992 en tant que professeur, responsable du département de mathématiques. Après avoir enseigné également à l'ISAE-Supaéro, il était parti à la retraite en septembre 2012. Auteur de plusieurs ouvrages dédiés aux mathématiques (dont *Cours d'analyse fonctionnelle et complexe, Probabilités et processus stocahstiques*), Yves Caumel a profondément marqué ses élèves et collègues par ses qualités humaines au premier rang desquelles son ouverture aux autres, sa générosité mais aussi son enthousiasme et ses remarquables qualités pédagogiques" (Source : *In memoriam*, Isae-Supaéro).