La formule de Bayes, quand les statistiques vous veulent du bien

Denis Maurel

Consultant Data et R&D · onepoint

Septembre 2021

onepoint.

beyond the obvious

SOMMAIRE

Denis Maurel, docteur en intelligence artificielle et chercheur R&D chez onepoint à Bordeaux, nous présente dans cet article le théorème de Bayes. Il nous explique comment ce théorème peut s'appliquer à des problèmes allant de la probabilité du sexe d'un enfant à la fiabilité des tests médicaux.

C'est aussi l'occasion de comprendre comment les équipes de onepoint utilisent ce théorème au quotidien dans leurs projets.

INTRODUCTION - LE PROBLEME DES DEUX FRERES

Au détour d'une balade, vous rencontrez un bon ami à vous que vous n'avez pas vu depuis un moment. Après les prises de nouvelles usuelles, il vous annonce qu'il a deux enfants. Il glisse immédiatement derrière qu'au moins un de ses deux enfants est un garçon. Et parce que vous le connaissez espiègle, vous savez que ses remarques ne sont pas innocentes. Vous sentez arriver la question improbable et celle-ci ne se fait pas attendre :

- À ton avis, sachant ce que je viens de te dire, quelle est la probabilité que mes deux enfants soient des garçons ?

Cette situation, bien qu'incongrue, constitue un des problèmes de base de ce que l'on appelle le **bayésianisme**. Cette discipline consiste en l'application un théorème bien connu des statisticiens au monde qui nous entoure : **le théorème de Bayes**.

Chaque minute, notre cerveau doit traiter des quantités astronomiques d'informations. Chaque décision que nous prenons, même infime, est motivée par **les informations qui sont à notre disposition** ou qui l'ont été dans un passé plus ou moins proche. Connaître l'influence d'une information sur une décision peut se révéler complexe sans l'apport des mathématiques. Ce que propose le théorème de Bayes, c'est de formuler mathématiquement et avec une remarquable simplicité **l'influence d'un événement passé sur un événement à venir**.

L'objectif de cet article n'est pas de vous noyer sous une avalanche de formules qui viserait à vous convaincre de l'utilité et de la pertinence du bayésianisme. Il s'agit plutôt de vous présenter en quoi un des théorèmes fondamentaux des statistiques peut nous permettre de mieux appréhender le monde qui nous entoure, en particulier lorsqu'il s'agit de prendre du recul sur les diverses statistiques que l'on peut rencontrer au quotidien.

Si jamais le fait de faire une multiplication et une division (littéralement) ne vous fait pas peur, je vous invite à venir découvrir le fantastique, que dis-je, le fabuleux théorème de Bayes, et à voir comment il peut nous aider à résoudre notre problème de fraternité.

LE THEOREME DE BAYES

Si je vous demande quelle est la probabilité que vous ayez faim à un moment aléatoire de votre journée, il vous suffira de regarder **la durée des périodes où vous avez faim par rapport à la durée totale d'une journée**. Si vous avez comme moi un amour inconditionnel pour la nourriture, cette probabilité pourra s'approcher de quelque chose comme 80%.

Si maintenant je vous demande quelle est la probabilité que vous ayez faim sachant que vous venez de sortir de table, on peut raisonnablement admettre que cette probabilité sera plutôt de l'ordre de 1%.

La modification de la probabilité d'un événement (le fait que vous ayez faim), à l'aide **d'éléments de contexte** (le fait que vous venez de sortir de table) peut mathématiquement se formuler à l'aide de ce que l'on appellera **les probabilités conditionnelles**. En bon français, si on généralise un peu en parlant de deux événements A et B, on pourra écrire :

- "Pour que A et B surviennent ensemble, il faut que A survienne puis que B survienne ensuite".

Ou, si l'on change l'ordre

- "Pour que A et B surviennent ensemble, il faut que B survienne puis que A survienne ensuite".

Ces phrases ne sont ni plus ni moins que la définition des probabilités conditionnelles. Lorsque l'on traduit le français en équations statistiques, il faut s'attarder sur le fait que chaque mot peut avoir son importance. Ainsi, quand on dit "A puis B", on évoque implicitement le fait qu'il va falloir **multiplier deux probabilités**. Pour illustrer, si, par une belle journée de week-end ensoleillé, vous avez 50% de chance d'aller au cinéma, et 50% de chance, une fois que vous y êtes, de voir votre film préféré, vous avez a priori 50% x 50% = 25% de chance d'aller voir votre film préféré (voir Figure 1). Ça vous apprendra à vouloir conjuguer votre amour de Tarantino avec la sortie hebdomadaire du petit dernier.

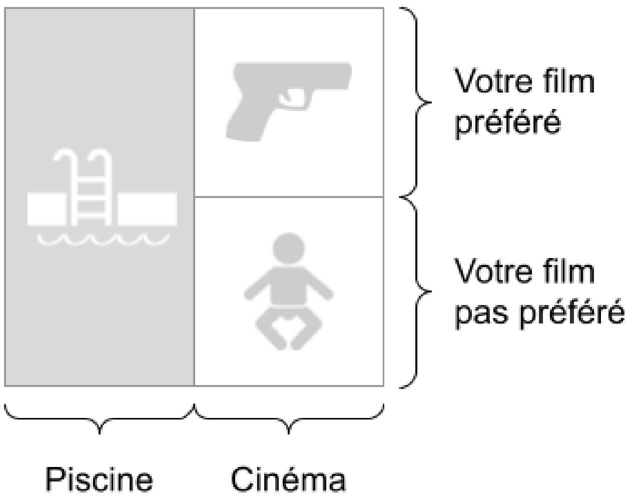


Figure 1. Le carré représente l'ensemble des possibilités, et chaque surface est proportionnelle à la chance que l'événement arrive. On voit graphiquement que vous avez une chance sur quatre d'aller voir votre film préféré

Pour généraliser, si on note ***P(A)*** **la probabilité d'un événement** A et ***P(A sachant B*****) la probabilité de A sachant que l'événement B est connu**, on obtient donc :

$$P(A \text{ et } B) = P(B){\times}P(A \text{ sachant } B) = P(A){\times}P(B \text{ sachant } A)$$

Si maintenant il vous prenait la folle envie de diviser toute cette équation par P(B), vous obtiendriez

$$P(A \text{ sachant } B) = \frac{P(A) \times P(B \text{ sachant } A)}{P(B)}$$

Si en plus de ça, vous aviez été pasteur de l'Église presbytérienne dans le Londres des années 1700, vous auriez découvert ce qui est dorénavant considéré comme une des formules fondamentales des statistiques. Loué soit Thomas Bayes !

Pour l'instant tout peut sembler un peu abstrait. Du coup, pour continuer sur la lancée de début d'article, je vous propose de regarder de nouveau le problème de votre ami fantasque.

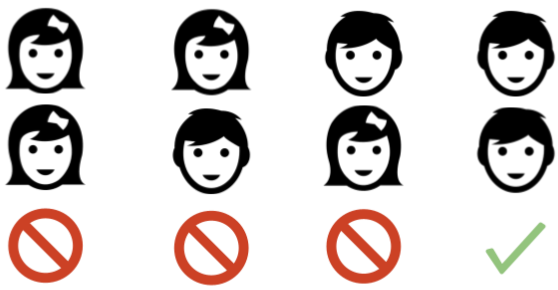
BAYES AU SECOURS DU MYSTERE DES DEUX ENFANTS

Pour rappel, vous savez que votre ami a deux enfants, et qu'**au moins un des deux est un garçon**. La question est, quelle est **la probabilité que l'autre soit aussi un garçon** ?

Dans un premier temps, et pour l'exercice intellectuel, je vous invite à vous poser la question. Prenez juste une minute pour bien visualiser le problème, et demandez-vous quelle réponse vous vient **intuitivement**.

Maintenant que c'est fait, regardons ce que nous apprend la formule de Bayes. On sait que pour calculer *P(les deux enfants sont des garçons sachant qu'un des deux est un garçon)*, nous avons besoin de trois éléments.

- **La probabilité que les deux enfants soient des garçons**. On peut facilement trouver qu'il y a une chance sur quatre pour que les deux enfants d'une personne soient du même sexe, et donc deux garçons. Il y a une chance sur deux pour le premier enfant, suivi d'une chance sur deux pour le second. En multipliant les deux probabilités, on arrive à une chance sur quatre. On a donc *P(les deux enfants sont des garçons)* = 25%.



- **La probabilité que notre contexte survienne**. Autrement dit, quelle est la probabilité qu'au moins un des deux enfants soit un garçon, toute autre information mise de côté. Pour la trouver, il suffit de compter toutes les possibilités où, ayant deux enfants, au moins un des deux est un garçon. Si vous avez deux enfants, il y a quatre possibilités : deux filles, deux garçons, un garçon puis une fille, une fille puis un garçon. Sur ces quatre possibilités, trois contiennent au moins un garçon. On peut donc dire que la probabilité que l'on cherche est égale à 75%.

- **La probabilité que notre contexte survienne sachant que l'élément est vérifié**. Ici, ça revient à nous demander quelle est la probabilité qu'un des deux enfants soit un garçon sachant que les deux sont des garçons. Sauf grosse erreur, on peut raisonnablement dire que cette probabilité est égale à 100%.

Si maintenant on injecte ces informations dans notre formule, on trouve

$$P(\text{deux garçons sachant au moins un garçon}) = \frac{25\% \times 100\%}{75\%} \approx 33.33\%$$

La bonne réponse est donc qu'il y a une chance sur trois pour que votre ami ait deux garçons, sachant qu'au moins un des deux est un garçon. Si vous aviez la bonne réponse, toutes mes félicitations, vous raisonnez comme un vrai bayésien. Si, comme la très vaste majorité d'entre nous, vous vous êtes intuitivement dit que l'autre enfant avait une chance sur deux d'être aussi un garçon, rassurez-vous, **c'est tout à fait normal**. On touche du doigt ce qui fait que les statistiques ne sont pas toujours la matière préférée des collégiens : les résultats, bien que justes, peuvent parfois se révéler complètement contre-intuitifs.

Le caractère contre-intuitif du problème vient ici de sa formulation. Si votre ami vous avez dit

- "**Mon premier enfant est un garçon**, quelle est la probabilité pour que le second soit aussi un garçon".

Alors vous auriez eu la bonne réponse. En effet, le sexe du premier enfant n'a pas d'influence sur le sexe du second. Cependant, dans sa formulation originale, votre fallacieux ami s'est bien gardé de préciser quel enfant (le premier ou le second) était un garçon.

Si vous avez besoin de vous convaincre de ce résultat, il suffit de faire un dessin.

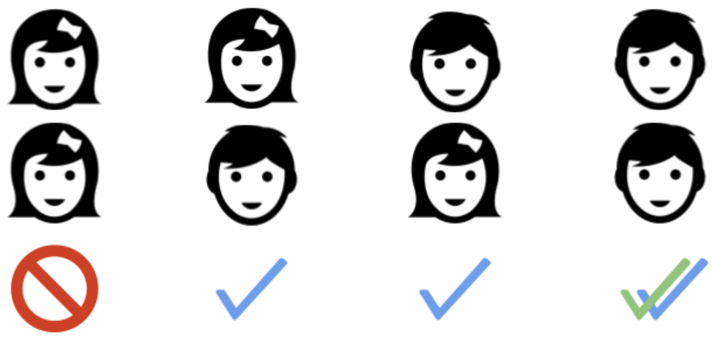


Figure 2. *Lorsque l'on regarde les possibilités en termes de duo d'enfants, sur les quatre duos possibles, le duo fille-fille est écarté d'office car il ne respecte pas le contexte "un des deux enfants est un garçon", représenté par les marques bleues. La contrainte que l'autre enfant soit aussi un garçon, représentée par une marque verte, n'est respectée que le par quatrième duo, le duo garçon-garçon. Il y a donc une chance sur trois d'avoir deux garçons.*

On sait qu'au moins un des enfants est un garçon, il nous faut donc exclure l'ensemble des possibilités qui ne comportent pas au moins un garçon. Seul le duo avec deux filles est concerné. Une fois le contexte pris en compte, il nous reste trois branches : garçon-garçon, garçon-fille et fille-garçon. Sur ces trois duos, combien répondent à notre demande que l'autre enfant soit aussi un garçon ? Il n'y en a effectivement qu'une seule, le duo garçon-garçon. Pour résumer, il y a une possibilité sur trois qui répond à notre demande, la probabilité finale est donc d'une chance sur trois.

En espérant vous avoir convaincu, je vous propose pour finir de nous attarder sur une situation plus proche de ce que l'on peut vivre au quotidien et où les statistiques peuvent se révéler très utiles.

LE BAYESIANISME ET LES TESTS MEDICAUX

Il s'agit là encore d'un exemple récurrent lorsque l'on explore le milieu du bayésianisme. Imaginez que vous rentrez d'un pays où se propage une maladie rare. Pour vous rassurer, vous choisissez d'aller passer un test de dépistage afin de ne prendre aucun risque. Vous savez qu'il y a environ une personne sur dix mille qui contracte cette maladie dans le pays d'où vous revenez, donc les chances pour que vous l'ayez attrapé sont plus que faibles, de l'ordre de 0.01%, mais sait-on jamais.

Vous vous rendez dans votre centre de soin préféré pour passer ledit test, et là, stupeur et damnation, **le test revient positif**. En bon bayésien, vous préférez vous assurer des données du problème, aussi vous demandez au médecin le niveau de confiance que l'on peut raisonnablement avoir dans ce test. Le médecin vous répond que, lorsqu'un un individu est malade, le test est fiable à 100%. Pas démoralisé pour un sou, vous lui retorquez que vous seriez plutôt intéressé par la fiabilité du test pour les personnes non malades, personnes dont vous espérez secrètement faire partie. Cette fois-ci, le médecin vous dit que, pour qui n'a pas la maladie, le test est fiable à 99%.

Bigre, vous voilà donc avec des données peu encourageantes et une solide boule au ventre en arrivance. C'était bien votre veine d'attraper une maladie rare en vacances. Histoire de pouvoir vous plaindre chiffres à l'appui, vous choisissez de coucher rapidement les calculs sur un coin de napperon afin de vérifier ce que vous dit la toute-puissante formule de Bayes.

Ce qui vous intéresse ici, c'est **le fait d'être ou non malade**. Le contexte que vous avez, c'est que l**e test est revenu positif**. Est-ce qu'il est possible de faire quelque chose de tout ça ? En appliquant la formule, on obtient :

$$P(\text{malade sachant test positif}) = \frac{P(\text{test positif sachant malade}) \times P(\text{malade})}{P(\text{positif})}$$

Pour ce qui est des deux premiers termes, vous avez facilement la réponse : le test est fiable à 100% pour les personnes malades, d'où P(test positif sachant malade) = 100%, et vous savez qu'environ une personne sur dix milles attrape cette maladie dans le pays d'où vous revenez, d'où *P(malade)* = 1 / 10 000 = 0.01%. Reste maintenant à déterminer la probabilité pour un individu aléatoire d'avoir un test positif.

Histoire de faire durer un tout petit peu le suspense, on aura ici besoin d'une petite étape supplémentaire pour obtenir ce que l'on veut. Plutôt que de se casser les dents à essayer de calculer ce terme directement, on va préférer **diviser les calculs** en se servant des informations à notre disposition. Si on prend notre cas comme exemple, on cherche à calculer *P(test positif)* et on sait que pour l'événement “être malade”, il n'y a que deux possibilités : “je suis malade” et “je ne suis pas malade”. On appelle cette méthode **la formule des probabilités totale**, et cette dernière nous dit que :

$$P(\text{test positif}) = P(\text{test positif et malade}) + P(\text{test positif et pas malade})$$

Pour illustrer grossièrement cette méthode, et histoire de revenir aux exemples improbables que l'on a tous connu à l'école, imaginez que vous ayez des carottes à compter. Rien ne vous empêche dans un premier temps de trier les carottes par taille, petite ou grande. Une fois ce tri fait, vous pouvez compter combien il y a de carottes dans chaque tas, puis vous additionnez les deux nombres pour obtenir le nombre total de carottes. C'est exactement la même chose qu'il se passe ici, sauf que l'on préfère compter suivant le fait d'être ou non malade plutôt que par taille. Notez que l'on aurait aussi pu le faire par taille, mais ça ne nous aurait pas servi à grand-chose.

En appliquant **la définition des probabilités conditionnelles** vu en début d'article, on peut continuer notre calcul comme suit :

$$P(\text{test positif}) = P(\text{malade}) \times P(\text{test positif sachant malade}) \\ + P(\text{pas malade}) \times P(\text{test positif sachant que pas malade})$$

Tout est bon ! Ça a été fastidieux, mais maintenant nous sommes heureux, car nous connaissons tous les termes de notre formule :

- *P(test positif sachant malade) = 100%*
- *P(malade) = 0.01%* (1 pour 10 000)
- *P(test positif sachant pas malade) = 1%* (d'après le médecin, le test est négatif pour les gens saints dans 99% des cas, il reste donc 1% de cas où les personnes saines ont un faux positif)
- *P(pas malade) = 99.99%* (les 9 999 autres cas sur 10 000)

On applique le sacro-saint théorème bayésien, et, stupéfaction, nous arrivons à la conclusion que

$$P(\text{malade sachant test positif}) = \frac{100\% \times 0.01\%}{100\% \times 0.01\% + 1\% \times 99.99\%} \approx 1\%$$

Même avec un test positif, vous n'avez finalement **qu'un petit pourcent de chance d'être malade**, quand bien même le test est efficace à 99%. C'est le déraisonnable pouvoir du bayésianisme : la formule est sans appel, les calculs sont justes, mais **le résultat vient parfois violemment contredire notre intuition**. Ce résultat explique aussi pourquoi est-ce que l'on pratique souvent plusieurs tests pour confirmer un résultat : le premier n'est pas toujours suffisant pour avoir une quasi-certitude sur un diagnostic. Pour information, si jamais vous faisiez un second test **complètement indépendant du premier**, et que ce test se révélait aussi positif, vos chances d'être malade grimperaient d'un coup à **98%**. On a rarement vu meilleure application du proverbe "plutôt deux fois qu'une".

LE THEOREME DE BAYES CHEZ ONEPOINT: LE PROJET MUTATRON

Si vous êtes arrivés à cette partie de l'article, il y a de bonnes chances pour que vous commenciez à entrapercevoir les possibilités offertes par notre très cher théorème. Ceci étant dit, n'ont été présentés que des cas de la vie courante sans lien avec **les nouvelles technologies**. Pour vous montrer que ce théorème ne sert pas qu'à poser des questions étranges sur les enfants de vos proches et sur les tests médicaux, je vais vous parler d'un projet de recherche développé en interne au sein de l'équipe R&D de onepoint. Il s'agit du projet baptisé **Mutatron**.

Avant de rentrer dans le vif de la théorie, un peu de contexte. Afin de garantir la stabilité du code d'un projet produit sur le long terme, les équipes de qualification et test se basent sur toutes sortes de **tests** afin de valider que le code produit à chaque instant **n'introduit pas de régression** par rapport à l'existant.

Le souci, c'est que pour des projets d'importance industrielle, ces campagnes de tests peuvent mettre **un certain temps** à s'exécuter. On parle ici de plusieurs heures, voire plusieurs jours, pour vérifier que tout va bien.

Lorsque le développement s'articule suivant le **cycle de l'intégration continue**, les campagnes de tests doivent être exécutées souvent. Dans le meilleur des cas, elles sont exécutées **à chaque fois que chacun des développeurs envoie du code sur le dépôt central du projet**. Le souci, c'est que la taille d'un projet est très souvent corrélée à la taille de l'équipe qui le produit, et il n'est pas rare que du code soit prêt à l'envoi **plusieurs fois par jour**. Un goulot d'étranglement va alors se créer, les tests n'ayant pas le temps de finir de s'exécuter entre deux envois.

C'est ici qu'intervient Mutatron. Afin de maximiser l'efficacité des campagnes de tests, nous allons chercher à déterminer **quels sont les tests qui sont le plus susceptibles d'échouer**, étant donnés les modifications de code à envoyer. L'idée est simple : lorsque vous modifiez le circuit d'eau dans la cuisine, il y a peu de chances pour que le circuit électrique de la salle de bain ait besoin d'être vérifié. Pour faire ça, Mutatron se base sur des statistiques bayésiennes pour réussir à déterminer quels sont les tests qui ont le plus de chance d'échouer.

Grossièrement, conjointement avec le temps d'exécution de chaque test, Mutatron va étudier pour chaque test *t* la probabilité

$$P(\text{test } t \text{ échoue sachant que les éléments de code } E \text{ ont été modifiés})$$

Il va ensuite se servir de ces éléments pour générer l'ordonnancement des tests qui aura en moyenne **le plus faible temps entre l'exécution et l'apparition du premier test échouant**. Là où le paradigme statistique gagne encore en intérêt, c'est qu'il permet de s'intéresser à d'autres probabilités telles que

$$P(\text{il reste un test en échec alors que les } n \text{ premiers sont passés})$$

Si l'on souhaite aller vite, et se contenter d'un test rapide avant d'envoyer du code sur le dépôt, on pourra tout à fait regarder quelle valeur de *n* nous permet d'avoir la probabilité précédente au-dessus de 95%. Si les modifications ne touchent pas l'intégralité du projet, il y a de bonnes chances pour que *n* ne représente qu'une **fraction de l'ensemble total des tests**. Cette méthode ne se substitue cependant pas à une exécution exhaustive des tests. Elle permet simplement d'identifier **localement** et **de façon précoce** les problèmes qui risqueraient de s'accumuler globalement.

Je ne m'étendrais pas plus sur les possibilités offertes par le paradigme bayésien dans le cadre de ce projet. J'espère seulement vous avoir convaincu que les possibilités offertes par les probabilités sont immenses, et ce, **quel que soit le domaine d'application**.

CONCLUSION

Le bayésianisme permet la prise en compte d'un ensemble **d'éléments de contexte** afin de mettre à jour **la probabilité d'apparition d'un événement**. Cette formule, que certains considèrent comme **le fondement de la rationalité**, peut être appliquée à toute prise de décision, indépendamment du domaine d'application et de la nature de l'événement à analyser.

Nous avons pu voir au travers de quelques exemples que l'utilisation de cette formule n'était cependant **pas toujours évidente**, et il vous faudra parfois faire attention où vous mettez les pieds lorsqu'il s'agira de l'appliquer à des cas concrets pouvant mener à une prise de décision véritable, typiquement dans un cadre professionnel.

La formule de Bayes, bien que théoriquement parfaite, présente un inconvénient pratique majeur. Lorsque les éléments de contexte à prendre en compte sont trop nombreux, **le temps de calcul nécessaire à la définition exacte des probabilités explose**. Il faut alors choisir les éléments de contexte **les plus pertinents**, ce qui peut ne pas être évident. Après tout, pourquoi ne pas considérer ce que vous avez mangé le mois dernier dans le calcul de vos chances d'avoir attrapé une maladie rare ?

Ceci dit, cela ne doit pas vous décourager, car les possibilités offertes par le théorème de Bayes restent plus qu'intéressantes **dans de très nombreux cas**. À titre d'exemple chez onepoint, nous avons ainsi pu voir comment le projet de recherche Mutatron se servait de l'analyse bayésienne afin d'optimiser les campagnes de tests, souvent chronophages, lancées en cours de développement.

Ce qu'il faut retenir, c'est que **la définition exacte des probabilités n'est possible en pratique que dans de très rares cas**. Pour les cas plus complexes, il vous faudra vous adapter, et utiliser judicieusement les possibilités offertes par le théorème de Bayes. Pour finir, je vous laisserai méditer sur une citation du statisticien George Box : **"Tous les modèles sont faux, certains sont utiles"**.