Maximum de vraisemblance, méthodes bayesiennes et moindres carrés

M. Bailly-Bechet

Université Nice Sophia Antipolis – France

1 Maximum de vraisemblance

1.1 Principe

Supposons que l'on dispose d'un modèle statistique (par exemple, la loi dirigeant en théorie la distribution de données observées) noté $\mathcal{M}$. Ce modèle peut avoir un ou plusieurs paramètres. Si la loi est une loi normale, par exemple, on a deux paramètres, la moyenne μ et la variance σ^2. Pour la suite, le texte sera rédigé comme si on avait un seul paramètre, noté θ ; tout ceci se généralise au cas où la loi a plusieurs paramètres.

Une question fréquente en inférence statistique est de trouver la meilleure valeur du paramètre θ, quand on dispose de données que l'on cherche à expliquer par ce modèle. Cela implique de définir un "score", qui mesure à quel point une certaine valeur du paramètre correspond mieux aux données observées qu'une autre. Un score classiquement défini, dans le cas des modèles probabilistes dont on parle ici, est la *vraisemblance* du modèle. Cette vraisemblance (*likelihood* en anglais) est la probabilité d'observer les données *obs* dont on dispose avec le modèle $\mathcal{M}$ et la valeur du paramètre θ_0 :

$$\mathcal{L} = P(obs|\theta_0, \mathcal{M}) \tag{1}$$

Si on dispose d'une série de données x_i, et qu'on suppose ces données indépendantes, on pourra réécrire la vraisemblance comme un produit :

$$\mathcal{L} = \prod_{i=1}^{n} P(x_i|\theta_0, \mathcal{M}) \tag{2}$$

Raisonner au *maximum de vraisemblance*, c'est inférer comme valeur d'un paramètre θ la valeur $\theta^\star$ qui rend cette vraisemblance maximale. Cela veut dire que l'on fait l'hypothèse que les données observées sont celles que le modèle avait le plus de chance de générer.

Mathématiquement, on a :

$$\theta^\star = \text{argmax}_\theta P(obs|\theta, \mathcal{M}) \tag{3}$$

1.2 Exemple discret : docteur, maladies et symptômes

On se place tout d'abord dans un cas où les différentes valeurs que peuvent prendre le paramètre sont discrètes. Nous sommes dans un cabinet médical. Les données sont les symptômes du patient, le modèle est que le patient souffre d'une maladie, et les différentes valeurs du paramètre sont les différentes maladies qui pourraient expliquer les symptômes.

Un patient souffre des symptômes suivants [1] :

— Défaillance rénale

— Paralysie partielle

On cherche à trouver un diagnostic pour ces symptômes, mais plusieurs correspondent :

— Un parasite cérébral pourrait expliquer facilement la paralysie et, éventuellement, expliquer la défaillance rénale. On considère que la probabilité d'avoir ces 2 symptômes avec un parasite cérébral est de $\frac{1}{100}$.

— D'autre part, un lupus (maladie auto-immune) expliquerait parfaitement la défaillance rénale, mais assez mal la paralysie. Ici aussi, la probabilité d'avoir ces symptômes suite à un lupus est de $\frac{1}{100}$.

Ici, les deux diagnostics ont la même vraisemblance, qui vaut $\frac{1}{100}$: on ne peut pas avec cette méthode distinguer entre les deux maladies.

On apprend ensuite plusieurs éléments. Après chaque nouveau cas, vérifiez que vous pouvez calculer la vraisemblance de chaque maladie au vu des données qui s'accumulent (on note au passage que la vraisemblance d'un modèle n'a de sens que par rapport à un certain jeu de données, si celui-ci change, la vraisemblance également) :

Vancomycine Le patient a subi un traitement à la vancomycine contre le parasite cérébral. Ce traitement a déclenché chez lui une allergie profonde et a du être arrêté très tôt. La probabilité que ce nouveau symptôme soit causé par un lupus est de $\frac{1}{10}$, alors que la probabilité que ce soit le parasite qui le cause est de $\frac{1}{1000}$.

Stéroïdes Un traitement aux stéroïdes destiné à éradiquer le lupus provoque des crises de démence chez le patient. La probabilité que les crises de démence soient causé par le lupus est quasi nulle, alors que le parasite cérébral les cause dans 1 cas sur 100.

Maladie de Wilson Une étude montre que la maladie de Wilson, une maladie génétique très rare, pourrait causer dans certains cas l'ensemble des symptômes observés, avec une probabilité de 1 pour un million.

Quel diagnostic est le plus vraisemblable à chaque étape du traitement ? Est ce que le fait que la maladie génétique soit rare influe sur votre analyse ? Que pensez-vous du diagnostic "Le patient a un parasite cérébral *et* un lupus" ?

Ne regardez pas la page suivante avant d'avoir répondu...

1. Inspiré de la série Dr. House, saison 2, épisode "Le Rasoir d'Occam" – données médicales non vérifiées

Vancomycine La vraisemblance de l'ensemble des symptômes si le patient a un parasite est $\frac{1}{100} \times \frac{1}{1000} = \frac{1}{100000}$. La vraisemblance du lupus est $\frac{1}{100} \times \frac{1}{10} = \frac{1}{1000}$, c'est dont le lupus qui est le diagnostic ele plus vraisemblable ici.

Stéroïdes Le lupus n'ayant aucune chance de provoquer les crises de démence, sa vraisemblance est nulle ; la vraisemblance du parasite cérébral est de $\frac{1}{100} \times \frac{1}{1000} times \frac{1}{100} = \frac{1}{10000000}$, faible mais non nulle : c'est le parasite qui est le diagnotic préféré.

Malaide de Wilson La vraisemblance de la maladie de Wilson pour expliquer les symptômes est de $\frac{1}{1000000}$. Ceci est plus élevé que le vraisemblance du parasite ($\frac{1}{10000000}$) ou du lupus (0), c'est donc le diagnostic conservé au maximum de vraisemblance. Le fait que cette maladie géńetique soit très rare *a priori* n'influence pas sur le raisonnement au maximum de vraisemblance ; la prise en compte des *a priori* se fait lors de calculs bayesiens (voir ci-dessous).

Parasite et lupus ? Le double diagnostic est tentant : il permet d'expliquer chacun des symptômes par la maladie qui a le plus de chances d'expliquer une partie des symptômes. Par contre, cela revient à ne plus respecter le modèle de départ, qui est que tous les symptômes sont la conséquence d'une seule maladie, et ce n'est donc pas comparable aux cas précédents. L'équivalent statistique serait de vouloir expliquer une partie des données avec une loi, et une partie avec une autre : on aura toujours dans ce cas une meilleure vraisemblance, le cas extrême étant celui ou chaque valeur unique est expliquée (parfaitement) par une loi différente.

1.3 Exemple continu : pile ou face

Prenons une pièce que l'on jette n fois. On obtient k piles. Quelle est l'estimation au maximum de vraisemblance de la fréquence p avec laquelle la pièce fait pile ? Pour la calculer, il faut l'expression de la vraisemblance, *i.e* la probabilité d'obtenir ce résultat en fonction de p. Si on suppose les tirages indépendants, on peut l'exprimer à l'aide d'une loi binomiale :

$$\mathcal{L} = \binom{n}{k} p^k (1-p)^{n-k} \tag{4}$$

Pour trouver la valeur de p qui maximise, à k et n donnés, cette expression, on calcule sa dérivée par rapport à p et on l'égalise à 0 :

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial p} &= \frac{\partial}{\partial p}\binom{n}{k} p^k (1-p)^{n-k} && (5)\\
&= \frac{\partial}{\partial p}\left(p^k (1-p)^{n-k}\right) && (6)\\
&= \left(\frac{\partial}{\partial p} p^k\right)(1-p)^{n-k} + p^k \left(\frac{\partial}{\partial p}(1-p)^{n-k}\right) && (7)\\
&= kp^{k-1}(1-p)^{n-k} - (n-k)p^k (1-p)^{n-k-1} && (8)\\
&= p^{k-1}(1-p)^{n-k-1}\left[k(1-p) - (n-k)p\right] = 0 && (9)
\end{aligned}$$

Si $p \neq 0$, la dernière ligne implique $k(1-p) - (n-k)p = 0$, c-à-d. $k - np = 0$ et $p = \frac{k}{n}$. Au maximum de vraisemblance, on estime que le paramètre p est égal à la fréquence observée. On

pourrait de la même manière retrouver que le meilleur estimateur de la moyenne d'une population est la moyennes observée, et que la variance observée sous-estime la variance de la population.

Notez un détail : si $n = 1$ et $k = 1$, le maximum de vraisemblance nous indique que le meilleur modèle est $p = 1$, c-à-d. que la pièce fait toujours pile. Ceci vient contredire notre intuition – on présuppose que la pièce n'est pas truquée et que les deux résultats sont possibles ; ici aussi, cet *a priori* ne peut être traité que par des méthodes bayesiennes. Par contre, cela permet de préciser un point : la technique du maximum de vraisemblance ne permet de retrouver la vraie valeur d'un paramètre que si la taille de l'échantillon est grande (mathématiquement, si elle tend vers l'infini).

1.4 Et si on a des *a priori* ?

Si on veut inclure dans l'analyse des *a priori* – par exemple le fait qu'une maladie rare est rare, et que donc elle a moins de chance d'être la cause de symptômes grippaux que la grippe – on va employer les statistiques bayesiennes. Elles se basent sur le théorème de Bayes, une loi sur les probabilités conditionnelles :

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{10}$$

d'où :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{11}$$

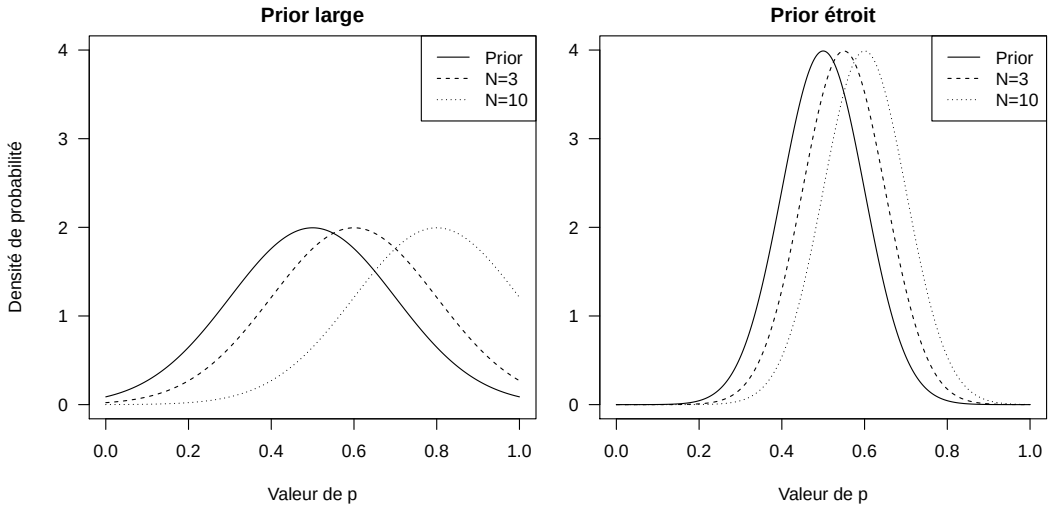
Dans un contexte d'inférence, on peut écrire cette dernière formule :

$$P(\theta|obs) = \frac{P(obs|\theta)P(\theta)}{P(obs)} = \frac{P(obs|\theta)P(\theta)}{\sum_\theta P(obs|\theta)P(\theta)}, \tag{12}$$

la dernière transformation au dénominateur servant à calculer $P(obs)$ en combinant toutes les probabilités d'observer les données avec toutes les valeurs de paramètres possibles dans le modèle.

On reconnait au numérateur le terme $P(obs|\theta)$, la vraisemblance. Ce terme est multiplié par $P(\theta)$, qui représente un *a priori* sur le paramètre θ que l'on cherche à estimer. En pratique, on va chercher non pas la "meilleure valeur" d'un paramètre, comme dans le maximum de vraisemblance, mais on va estimer la distribution de probabilité de ce paramètre : les valeurs de $P(\theta)$ pour différentes valeurs de θ donnent l'*a priori* sur les différentes valeurs possibles, et le terme $P(obs|\theta)$, la vraisemblance, vient modifier la distribution finale $P(\theta|obs)$ (nommée *a posteriori*) en y incluant l'influence des données observées dans l'expérience.

Voila deux cas avec des *a priori* différents, et les résultats de deux expériences où on lance N fois une pièce, qui fait dans cet exemple toujours pile. Plus on la lance, plus on a la certitude que sa vraie probabilité de faire pile, notée p, est élevée.



Estimation bayesienne de la distribution de probabilité a posteriori *de la probabilité* p *pour une pièce de faire pile après* N *tirages à partir d'un* a priori *large (à gauche), donc incertain, ou d'un* a priori *étroit (à droite), marquant une certitude plus forte au départ.*

Quelques éléments de conclusion et d'ouverture sur ces méthodes bayesiennes qui ne sont qu'effleurées ici :

- Aucune expérience ne pourra modifier des certitudes absolues, comme un *a priori* $P(\theta = \theta_0) = 0$.
- Le résultat sera toujours un mélange entre la distribution du paramètre *a priori* et les résultats expérimentaux : il faut peser soigneusement les deux parties en choisissant le degré de certitude que l'on a dans l'*a priori*. Le choix de la distribution *a priori* est d'ailleurs un problème en soi dans les méthodes bayesiennes, mais il existe des techniques permettant en particulier de choisir un *a priori* le moins informatif possible – et donc, de travailler dans le contexte bayesien, avec ces avantages (on travaille sur des distributions de probabiilté et pas sur une estimation ponctuelle des paramètres) même sans inclure d'information pouvant biaiser le résultat final.
- La manière de travailler bayesienne impose souvent de travailler numériquement, sauf dans quelques cas bien particuliers (i.e. un *a priori* normal donne un *a posteriori* normal).
- La bonne nouvelle est que, si la quantité de données est grande, les analyses bayesiennes et les analyses au maximum de vraisemblance convergent sur les mêmes estimateurs ! En effet, pour de très grands échantillons, la distribution *a priori* n'a quasiment plus aucune importance, les données ayant beaucoup plus de poids dans la résultat que l'*a priori*.

2 Lien avec les moindres carrés

Dans un modèle de régression linéaire simple, on a des données appariées (x_i, y_i) et on suppose un modèle du type :

$$y_i = ax_i + b + \epsilon_i, \tag{13}$$

avec les ϵ_i distribués selon une loi normale de moyenne nulle et de variance σ^2 indépendante de i. La densité de probabilité d'une loi normale de moyenne μ et de variance σ^2 vaut :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(\frac{-(x-\mu)^2}{2\sigma^2})$$

.

Autrement dit, cela veut dire que pour chaque couple (x_i, y_i), on peut écrire la probabilité de y_i en fonction de x_i ainsi :

$$\begin{aligned}
\mathcal{P}(y_i|x_i) =& \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i - (ax_i + b))^2}{2\sigma^2}\right) && (14)\\
=& \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-\epsilon_i^2}{2\sigma^2}\right) && (15)
\end{aligned}$$

Dans cette équation, les valeurs de a et b sont cachées, mais ce sont d'elles que dépendent les ϵ_i ; c-à-d. qu'optimiser la vraisemblance exprimée en fonction de ϵ_i revient bien chercher les valeurs de a et b qui la maximisent.

La vraisemblance de l'ensemble des n données (x_i, y_i) est donc :

$$\begin{aligned}
\mathcal{L} =& \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-\epsilon_i^2}{2\sigma^2}\right) && (16)\\
=& \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \prod_{i=1}^{n} \exp\left(\frac{-\epsilon_i^2}{2\sigma^2}\right) && (17)\\
=& \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2\right) && (18)
\end{aligned}$$

Cette vraisemblance est maximale quand $\exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2\right)$ est maximal. On peut facilement constater que cette fonction est monotone décroissante sur $[0, +\infty[$ en fonction de $\sum_{i=1}^{n} \epsilon_i^2$, c-à-d. que maximiser la vraisemblance revient exactement à minimiser $\sum_{i=1}^{n} \epsilon_i^2$: le critère des moindres carrés et le critère du maximum de vraisemblance donnent les mêmes estimateurs pour la régression linéaire simple. Ceci est du à la nature particulière de la loi normale, et n'est pas vrai dans le cas général ; mais l'approche de maximum de vraisemblance peut toujours etre employée si on est capable d'écrire le modèle probabiliste générant les données, ce qui est le cas par exemple quand on travaille sur le modèle linéaire généralisé.