Eléments de statistiques pour les data sciences

Cours 6 : Maximum de vraisemblance

Dr. Ph. Müllhaupt

IGM - EPFL

—

Plan

considérations d'ordre général

Quand un échantillon (petit ensemble) est obtenu d'une population (ensemble de cardinalité grande), une estimée d'un paramètre obtenu à partir de l'échantillon n'est en général pas égal à la valeur inconnue du paramètre de la population.

Même lorsque un dès est lancé 120 fois, on ne s'attend pas à obtenir 20 fois la face 6 si on suppose le dès non biaisé et qu'il soit non biaisé.

Ce qui est d'intérêt est de déterminer l'étendue des possibilités des valeurs des paramètres la plus en adéquation avec les hypothèses et le modèle.

exemple introductif

Prenons comme modèle une distribution binomiale de paramètre p.

$$P(n) = \left(\begin{array}{c} N \\ n \end{array} \right) p^n (1-p)^{N-n}$$

Si on prend une valeur bien spécifique de p, il est possible de calculer la probabilité d'un certain nombre n de réussites.

Il est donc possible de faire un calcul relativement exhaustif d'un grand nombre de cas en variant le paramètre p pour tous les choix de n parmis $1 \ldots N$. Ceci donne la probabilité d'observer un ensemble de résultats de l'expérience.

Il est important d'effectuer cette évaluation avant l'expérience.

exemple introductif

Avec $N = 8$, en prenant p variant par pas de 0.1 de 0.1 à 0.9 :

	paramètre p								
n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.430	0.168	0.058	0.017	0.004	0.000	0.000	0.000	0.000
1	0.383	0.336	0.198	0.090	0.031	0.008	0.001	0.000	0.000
2	0.149	0.294	0.296	0.209	0.109	0.041	0.010	0.001	0.000
3	0.033	0.147	0.254	0.279	0.219	0.124	0.047	0.009	0.000
4	0.005	0.046	0.136	0.232	0.273	0.232	0.136	0.046	0.005
5	0.000	0.009	0.047	0.124	0.219	0.279	0.254	0.147	0.033
6	0.000	0.001	0.010	0.041	0.109	0.209	0.296	0.294	0.149
7	0.000	0.000	0.001	0.008	0.031	0.090	0.198	0.336	0.383
8	0.000	0.000	0.000	0.000	0.004	0.017	0.058	0.168	0.430

exemple introductif

Avec $N = 8$, en prenant p variant par pas de 0.1 de 0.1 à 0.9 :

	paramètre p								
n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.430	0.168	0.058	0.017	0.004	0.000	0.000	0.000	0.000
1	0.383	0.336	0.198	0.090	0.031	0.008	0.001	0.000	0.000
2	0.149	0.294	0.296	0.209	0.109	0.041	0.010	0.001	0.000
3	0.033	0.147	0.254	0.279	0.219	0.124	0.047	0.009	0.000
4	0.005	0.046	0.136	0.232	0.273	0.232	0.136	0.046	0.005
5	0.000	0.009	0.047	0.124	0.219	0.279	0.254	0.147	0.033
6	0.000	0.001	0.010	0.041	0.109	0.209	0.296	0.294	0.149
7	0.000	0.000	0.001	0.008	0.031	0.090	0.198	0.336	0.383
8	0.000	0.000	0.000	0.000	0.004	0.017	0.058	0.168	0.430

maximums

des lignes et des colonnes

on constate :

- un maximum unique dans chaque colonne (probabilité)
- un maximum dans chacune des lignes (vraisemblance)
- le maximum d'une colonne est le maximum d'une ligne unique
- le maximum d'une ligne est le maximum d'une colonne unique

vraisemblance $\neq$ probabilité

le total des colonnes donne toujours 1 (probabilité)

	paramètre p								
n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	0.430	0.168	0.058	0.017	0.004	0.000	0.000	0.000	0.000
1	0.383	0.336	0.198	0.090	0.031	0.008	0.001	0.000	0.000
2	0.149	0.294	0.296	0.209	0.109	0.041	0.010	0.001	0.000
3	0.033	0.147	0.254	0.279	0.219	0.124	0.047	0.009	0.000
4	0.005	0.046	0.136	0.232	0.273	0.232	0.136	0.046	0.005
5	0.000	0.009	0.047	0.124	0.219	0.279	0.254	0.147	0.033
6	0.000	0.001	0.010	0.041	0.109	0.209	0.296	0.294	0.149
7	0.000	0.000	0.001	0.008	0.031	0.090	0.198	0.336	0.383
8	0.000	0.000	0.000	0.000	0.004	0.017	0.058	0.168	0.430
tot	1	1	1	1	1	1	1	1	1

vraisemblance $\neq$ probabilité

le total des lignes change de ligne en ligne et ne donne pas 1 (vraisemblance)

	paramètre p								
tot	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.677	0.430	0.168	0.058	0.017	0.004	0.000	0.000	0.000	0.000
1.047	0.383	0.336	0.198	0.090	0.031	0.008	0.001	0.000	0.000
1.109	0.149	0.294	0.296	0.209	0.109	0.041	0.010	0.001	0.000
1.112	0.033	0.147	0.254	0.279	0.219	0.124	0.047	0.009	0.000
1.111	0.005	0.046	0.136	0.232	0.273	0.232	0.136	0.046	0.005
1.112	0.000	0.009	0.047	0.124	0.219	0.279	0.254	0.147	0.033
1.109	0.000	0.001	0.010	0.041	0.109	0.209	0.296	0.294	0.149
1.047	0.000	0.000	0.001	0.008	0.031	0.090	0.198	0.336	0.383
0.677	0.000	0.000	0.000	0.000	0.004	0.017	0.058	0.168	0.430

expérience et vraisemblance

Supposons une expérience effectuée donnant un résultat E. Supposons que le modèle probabiliste pour l'expérience comporte un paramètre θ. On aimerait, au vu de E, déterminer une valeur de θ. En utilisant les lois des probabilités, on peut déterminer la probabilité de E. Cette probabilité sera fonction de θ

$$P(E;\theta)$$

Parmis toutes les valeurs de θ on choisira celle qui donne la plus grande probabilité.

vraisemblance et logarithme de vraisemblance

définitions

définition

Une fonction de vraisemblance de θ est une fonction définie par

$$L(\theta) \triangleq k \cdot P(E;\theta)$$

avec k un nombre positif ou une fonction positive des données mais pas fonction de θ

logarithme de vraisemblance :

$$l(\theta) \triangleq \log L(\theta)$$

espace des paramètres :

On dénote par Ω l'ensemble des paramètres possibles pour θ

estimateur du maxium de vraisemblance

MLE — Maximum Likelihood Estimator

$$\hat{\theta}_{\text{MLE}} \triangleq \arg\max_{\theta\in\Omega} L(\theta) = \arg\max_{\theta\in\Omega} l(\theta)$$

c'est l'estimée selon le maximum de vraisemblance (Maximum Likelihood Estimator abbréviation MLE).

prendre le logarithme permet d'additionner les logarithmes lorsqu'il y a des facteurs multiplicatifs (par exemple, lors d'évènements indépendants, on multiplie les probabilités et donc on additionne les logarithmes des vraisemblances).

score
$S \triangleq l'(\theta)$

score :
Le score $S(\theta)$ est la dérivée du logarithme de vraisemblance

$$S(\theta) \triangleq \frac{d}{d\theta} \log(L(\theta)) = \frac{dl(\theta)}{d\theta} = l'(\theta)$$

condition nécessaire pour ML (cas régulier) :
mais pas suffisante pour le maximum de vraisemblance (ML)

$$S(\theta) = l' = 0$$

information :
l'information est l'opposé de la dérivée seconde

$$\mathscr{I} \triangleq -\frac{d^2}{d\theta^2}\log(L(\theta)) = -\frac{d^2 l(\theta)}{d\theta^2} = -l''(\theta)$$

condition nécessaire et suffisante (cas régulier) pour ML local :

$$\begin{array}{ll} l'(\theta) = 0 & l''(\theta) < 0 \\ S(\theta) = 0 & \mathscr{I}(\theta) > 0 \end{array}$$

si solution unique alors ML

exemple

loi binomiale

Soit un grand nombre de pièces mécaniques de même type (donc population homogène) pour lesquels on aimerait déterminer le pourcentage de pièces défectueuses. On choisit au hasard n pièces et on détecte x pièces défectueuses. La probabilité de ce que l'on vient d'observer (E est x parmis n sont des pièces défectueuses) suit une loi binomiale

$$P(E;\theta) = \left(\begin{array}{c} n \\ x \end{array} \right) \theta^x (1-\theta)^{n-x}$$

Comme $\left(\begin{array}{c} n \\ x \end{array} \right) > 0$, on peut prendre pour fonction de vraisemblance

$$L(\theta) = \theta^x \, (1-\theta)^{n-x}$$

ce qui donne pour le logarithme de la vraisemblance

$$l(\theta) = x \log(\theta) + (n-x) \log(1-\theta)$$

exemple

loi binomiale

Calcul du maximum en annulant la première dérivée

$$S(\theta) = l'(\theta) = \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0 \qquad \Rightarrow \qquad \boxed{\hat{\theta}_{\text{MLE}} = \frac{x}{n}}$$

Il s'agit bien d'un maximum relatif en vérifiant

$$\mathscr{I}(\theta) = -l''(\theta) = \frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2} > 0 \qquad \text{lorsque } \theta = x/n$$

exemple introductif (suite)

exemple lorsque $n = 2$ et $N = 8$

Supposons que l'expérience conduit à obtenir $n = 2$ avec $N = 8$ expériences en tout.

$$\hat{p} = \frac{n}{N} = 2/8 = 1/4 = 0.25$$

et il n'était pas dans le tableau initial

$$p(2) = \begin{pmatrix} 8 \\ 2 \end{pmatrix} \hat{p}^2(1-\hat{p})^6 = \boxed{0.3114}$$

et c'est entre les colonne 0.2 et 0.3 du tableau

	paramètre p								
n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
1	0.383	0.336	0.198	0.090	0.031	0.008	0.001	0.000	0.000
2	0.149	0.294	0.296	0.209	0.109	0.041	0.010	0.001	0.000
3	0.033	0.147	0.254	0.279	0.219	0.124	0.047	0.009	0.000

exemple : bactéries dans une eau de rivière

distribution de Poisson

Lorsqu'il est possible de déterminer le nombre de bactéries par unité de volume d'eau prélevée, il est possible de déterminer le paramètre μ de la distribution de Poisson en prélevant n échantillons.
la probabilité d'avoir x bactéries par unité de volume est supposée obéir à une loi de Poisson de moyenne μ

$$f(x) = (\mu)^x \frac{e^{-\mu}}{x!} \qquad x = 0, 1, 2, 3, \ldots$$

Comme les volumes d'eau sont disjoints, les probabilités sont indépendantes et on a

$$P(E; \mu) = f(x_1)\, f(x_2) \ldots, f(x_n) = \Pi_{i=1}^{n} (\mu)_i^x \frac{e^{-\mu}}{x_i!} = \frac{\mu^{\sum x_i}\, e^{-n\,\mu}}{x_1!\, x_2! \ldots x_n!}$$

on peut prendre comme fonction de vraisemblance

$$L(\mu) = \mu^{\sum x_i}\, e^{-n\,\mu} \qquad 0 \leq \mu < +\infty$$

exemple : bactéries dans une eau de rivière

distribution de Poisson

Dans certains pas, il n'est pas possible de déterminer le nombre de bactéries par unité de volume d'eau. On peut seulement déterminer s'il y a présence ou absence de bactérie dans un prélèvement donné.

On préléve n tubes d'eau de même volume et on teste la présence de bactéries.

exemple : bactéries dans une eau de rivière

distribution de Poisson

La probabilité qu'il y ait x bactéries dans un volume v d'eau donné suit une distribution de Poisson de moyenne $\mu\, v$

$$f(x) = (\mu\, v)^x \, \frac{e^{-\mu\, v}}{x!} \qquad x = 0, 1, 2, 3, \ldots$$

- probabilité de ne pas tester de bactérie : $f(0) = e^{-\mu\, v}$
- probabilité de tester la présence de bactéries : $1 - p = 1 - e^{-\mu\, v}$

exemple : bactéries dans une eau de rivière

distribution de Poisson

Le test de n tubes sont indépendants, on peut donc représenter le résultat de tester y tubes parmi n tubes par une loi binomiale

$$P(E;\mu) = \left(\begin{array}{c} n \\ y \end{array} \right) p^y (1-p)^{n-y} \qquad p = e^{-\mu\, v} \qquad 0 \leq \mu < +\infty$$

on peut prendre comme fonction de vraisemblance

$$L(\mu) = p^y (1-p)^{n-y}$$

$$p = e^{-\mu v} \Rightarrow \mu = -\frac{1}{v} \log p$$

Résultat de l'exemple précédent, $p^y(1-p)^{n-y}$ est maximisé pour

$$\hat{p}_{\text{MLE}} = \frac{y}{n} \Rightarrow \hat{\mu}_{\text{MLE}} = -\frac{1}{v} \log \hat{p}_{\text{MLE}} = -\frac{1}{v} \log \frac{y}{n} = \frac{\log n - \log y}{v}$$

vraisemblance et tableau de fréquences

classe	A_1	A_2	$\cdots$	A_n	total
fréquence observée	f_1	f_2	$\cdots$	f_n	n
fréquence théorique	$n\,p_1$	$n\,p_2$	$\cdots$	$n\,p_n$	n

$$P(E;\theta) = \left(\begin{array}{c} n \\ f_1\, f_2\, \cdots\, f_n \end{array} \right) p_1^{f_1} p_2^{f_2} \cdots p_n^{f_n}$$

$$L(\theta) = p_1^{f_1} p_2^{f_2} \cdots p_n^{f_n}$$

$$l(\theta) = f_1 \log p_1 + f_2 \log p_2 + \ldots + f_n \log p_n$$

vraisemblance et tableaux de fréquences

nombre de pièces défectueuses	0	1	2	3	≥ 4	total
fréquence observée	133	52	12	3	0	200

hypothèse : le nombre de pièces défectueuses parmi 10 est supposée suivre une distribution binomiale

$$p_j = \begin{pmatrix} 10 \\ j \end{pmatrix} \theta^j (1-\theta)^{10-j} \qquad j = 0, 1, 2, \ldots, 10$$

$$p_{4+} = 1 - p_0 - p_1 - p_2 - p_3$$

$$L(\theta) = p_0^{133} p_1^{52} p_2^{12} p_3^3 p_{4+}^0 \qquad 0 \leq \theta \leq 1$$

$$\begin{aligned} L(\theta) &= [(1-\theta)^{10}]^{133} [\theta(1-\theta)^9]^{52} [\theta^2(1-\theta)^8]^{12} [\theta^3(1-\theta)^7]^3 \\ &= \theta^{85}(1-\theta)^{1915} \end{aligned}$$

$$n = 2000 \qquad x = 85$$

$$\hat{\theta}_{\text{MLE}} = \frac{x}{n} = \frac{85}{2000} = 0.0425$$

$$\begin{aligned}
e_0 &= p_0 \cdot 200 = (1 - 0.0425)^{10} \cdot 200 = 129.54 \\
e_1 &= p_1 \cdot 200 = C_1^{10} \cdot 0.0425 \cdot (1 - 0.0425)^9 \cdot 200 = 57.50 \\
e_2 &= p_2 \cdot 200 = C_2^{10} \cdot (0.0425)^2 \cdot (1 - 0.0425)^8 \cdot 200 = 11.48 \\
e_3 &= p_3 \cdot 200 = C_3^{10} \cdot (0.0425)^3 \cdot (1 - 0.0425)^7 \cdot 200 = 1.36 \\
e_4 &= 200 - e_0 - e_1 - e_2 - e_3 = 0.12
\end{aligned}$$

nombre de pièces défectueuses	0	1	2	3	≥ 4	total
fréquence observée	133	52	12	3	0	200
fréquence théorique	129.54	57.50	11.48	1.36	0.12	200

$$d = 2.74 \qquad P\{\chi^2_{(4)} \geq d\} \approx 0.6$$

propriétés de la vraisemblance

et logarithme de vraisemblance

- lorsque les évènements sont indépendants les probabilités se multiplient et les vraisemblances se multiplient (les logarithmes de vraisemblance s'additionent).

vraisemblance relative

et logarithme de vraisemblance relative

Supposons que 40 tubes soient testés qui contiennent 10 [ml] d'eau chacun. Si 26 sont négatifs et 14 positifs, alors on aura

$$\hat{\mu}_{\mathsf{MLE}} = \frac{\log 40 - \log 26}{10} = 0.0431$$

bactéries par ml. Plus grande est la concentration en bactéries, le plus probable sera que les n tubes soient testés positifs. Ainsi le plus μ devient grand le plus $y = 0$ sera probable comme résulat observé. Si on observe $y = 0$, l'estimée du maximum de vraisemblance pour μ sera $+\infty$. En pratique, il est plus judicieux de ne pas donner uniquement une valeur de μ mais plutôt une fourchette de valeurs pour μ. Ceci est possible en introduisant la vraisemblance relative.

vraisemblance relative

et logarithme de vraisemblance relative

$$R(\theta) \triangleq \frac{L(\theta)}{L(\hat{\theta}_{\mathrm{MLE}})}$$

$$r(\theta) \triangleq \log L(\theta) - \log L(\hat{\theta}_{\mathrm{MLE}})$$

Propriétés :

$$0 \leq R(\theta) \leq 1 \qquad -\infty \leq r(\theta) \leq 0 \qquad \forall \theta$$

vraisemblance relative et intervalles de vraisemblances

définition

L'ensemble des paramètres pour lesquels

$$\boxed{R(\theta) \geq \alpha}$$

est appelé une à région de $100 - \alpha$ % de vraisemblance pour le paramètre θ.

En général, on considère les valeurs α de 50 %, 10 %, et 1 % de vraisemblance (ou régions de vraisemblance).

En introduisant la fonction du logarithme de vraisemblance relative $r(\theta)$

$$\boxed{r(\theta) \geq -0.69, -2.3, \text{ et } -4.61 \qquad \text{respectivement}}$$

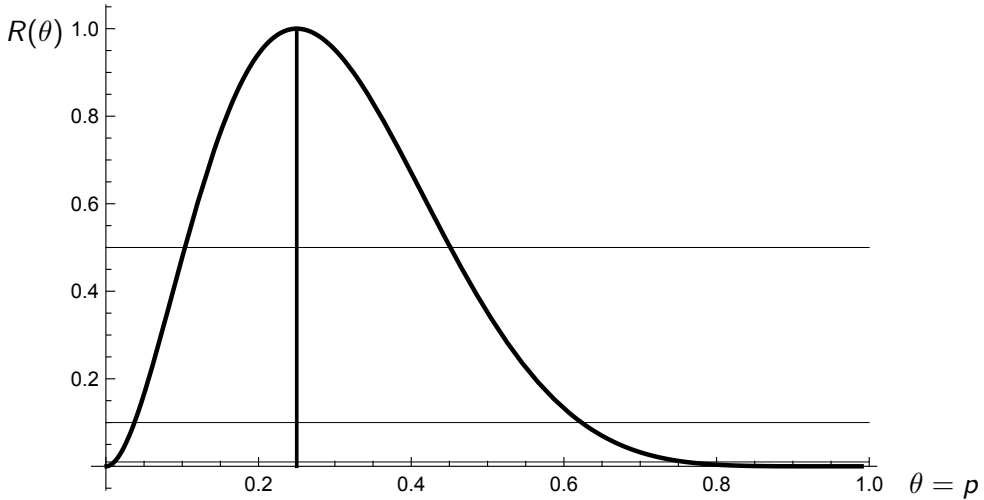

Figure – Vraisemblance relative $R(\theta)$ en fonction de θ. Les intervalles associés à 1 %, 10 %, et 50 % sont obtenus en coupant les niveau correspondant par le graphique. Cas binomial $N = 8$ avec $n = 2$.

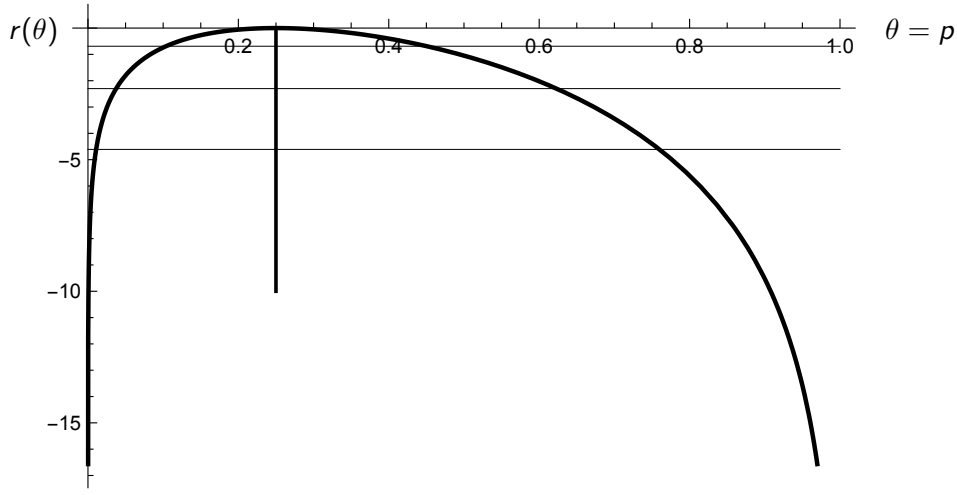

Figure – Logarithme de vraisemblance relative $r(\theta)$ en fonction de θ. Les intervalles associés à 1 %, 10 %, et 50 % sont obtenus en coupant les niveau correspondant par le graphique. Cas binomial $N = 8$ avec $n = 2$.