Eléments de statistiques pour les data sciences

Cours 7: Estimation - statistique suffisante - Cramér-Rao

Dr. Ph. Müllhaupt

IGM - EPFL

Plan

Le problème de l'estimation

expectative et valeurs asymptotiques

Soit $\hat{\theta}$ un estimateur pour un paramètre θ

propriétés désirées de l'estimateur $\hat{\theta}$

- On aimerait l'absence de biais
$$\mathbb{E}_\theta[\hat{\theta}] = \theta$$
- Une erreur quadratique moyenne (MSE) petite :
$$R(\theta, \hat{\theta}) \triangleq \mathbb{E}_\theta[(\theta - \hat{\theta})^2]$$

Espérance mathématique à partir de l'observation

Dans certain cas, prendre l'espérance mathématique est remplacé par une moyenne d'un grand nombre de réalisations. On remplace l'opérateur $\mathbb{E}$ par une moyenne sur un grand nombre de réalisation.

statistique suffisante

- Soit un vecteur de variables aléatoires Y qui décrit un échantillon. Une expérience conduit à observer y, une valeur particulière du vecteur aléatoire Y.
- La variable aléatoire est paramétrisée par θ. La suffisance est un concept qui permet de caractériser une statistique h(Y) afin de résumer tout ce qui qui est dans Y concernant la déduction de la valeur de θ en laissant de côté ce qui ne donne pas d'information concernant θ.
- La plupart du temps, la statistique h est de dimension plus petite que l'échantillon

$$\dim \mathrm{h} < \dim \mathrm{Y}$$

- La valeur de la statistique h partitionne l'espace des échantillons en catégories.
- Au sein d'une partition il n'y a plus d'information utile concernant le paramètre θ.

définition

Soit un vecteur de variables aléatoires qui décrit un échantillon

$$\mathrm{Y} = (Y_1, \ldots, Y_n)$$

Une statistique

$$\mathrm{U} = \mathrm{h}(\mathrm{Y})$$

est dite suffisante pour θ si la dstribution conditionnelle

$$P(Y|U)$$

ne dépend pas de θ.

statistique suffisante

remarques

- La définition se réfère à la distribution de Y.
- Pour cette raison, on se réfère également à une famille de fonctions de répartition

$$\{F_Y(\cdot;\theta) | \, \theta \in \Theta\}$$

- La définition est équivalente à dire que la fonction de répartition conditionnelle

$$F_{Y|U}(y,u)$$

n'est pas une fonction de θ

- autrement dit la densité de probabilité conditionnelle

$$f_{Y|U}(y,u)$$

ne dépend pas de θ

statistique suffisante

lemme

lemme : conditionnement sur une fonction de l'échantillon

Supposons que Y soit un échantillon d'un vecteur de variables aléatoires discrètes. Soit $U = h(Y)$ une statistique. La distribution de probabilité conditionnelle discrète est donnée par :

$$p_{Y|U}(y, u) = \begin{cases} \frac{p_Y(y)}{p_U(u)} & \text{lorsque } h(y) = u, p_U(u) \neq 0 \\ 0 & \text{sinon} \end{cases}$$

exemple

variables de Bernouilli

Soit $\mathrm{Y} = (Y_1, \ldots, Y_n)$, n expériences de Bernouilli (réussite 1, raté 0). La distribution de probabilité discrète pour chaque composante Y_i est :

$$p_{Y_i}(\mathrm{y_i}) = \begin{cases} p & y_i = 1 \\ (1-p) & y_i = 0 \end{cases}$$

Pour chaque vecteur de variables aléatoires observées $\mathrm{y} = (y_1, y_2, \ldots, y_n)$, on a le nombre total de réussites $u = \sum_{i=1}^{n} y_i$, et le nombre total de ratés $n - u$. Par la propriété d'indépendance, on peut multiplier les probabilité de chaque variable de Bernouilli

$$p_{\mathrm{Y}}(\mathrm{y}) = p^u (1-p)^{n-u} \qquad u = \sum_{i=1}^{n} y_i$$

Considérons la statistique

$$U = \sum_{i=1}^{n} Y_i$$

Elle suit la distribution binomiale $U \sim \mathscr{B}in(n, p)$ donnée par :

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

le lemme donne

$$p_{Y|U}(y, u) = \frac{p_Y(y)}{p_U(u)} = \frac{p^u (1-p)^{n-u}}{\left(\begin{array}{c} n \\ u \end{array} \right) p^u (1-p)^{n-u}} = \frac{1}{\left(\begin{array}{c} n \\ u \end{array} \right)}$$

Intuition : S'il y a u réussites, il y a $\left(\begin{array}{c} n \\ u \end{array} \right)$ façons (arrangements) de le faire, et toutes (tous) équiprobables.

statistique suffisante

démonstration du lemme

$$h(y) \triangleq u$$
$$Y = y \quad \Rightarrow \quad h(Y) = h(y) \Rightarrow U = u \tag{1}$$

Au sens des évènements cela signifie

$$\{Y = y\} \subseteq \{U = u\}$$

on déduit alors

$$p_{Y,U}(y,u) = P(Y = y, U = u) = P(Y = y) = p_Y(y) \tag{2}$$

si $h(u) \neq u$, alors $\{Y = y\} \cap \{U = u\} = \emptyset$ et donc $p_{Y,U}(y,u) = 0$.

principe de suffisance

Si U = h(Y) est une statistique suffisante pour θ, alors toute déduction concernant θ doit pouvoir s'effectuer à travers les réalisations de la statistique (variable aléatoire) U. (L'inférence à partir de la variable aléatoire Y concernant θ doit pouvoir s'effectuer seulement à partir de u.)

En reprenant l'exemple des variables de Bernouilli et de la statistique $U = \sum Y_i$, et en considérant deux réalisations de la statistique U celles donnant $u = 2$ avec $n = 5$ dans les deux cas suivants :

$$y_1 = (0, 1, 0, 1, 0)$$
$$y_2 = (1, 0, 0, 1, 0)$$

Dans ce cas $y_1 \neq y_2$ mais $\sum y_{1,i} = \sum y_{2,i} = 2$ ainsi la réalisation de la statistique U est la même dans les deux cas. L'estimation de p sera bien déterminé entièrement par la statistique suffisante $\sum y_i$. En effet $\hat{p} = 2/5$ dans les deux cas.

théorème de factorisation

Soit $\mathrm{Y} = (Y_1, Y_2, \ldots, Y_n)$ un vecteur de variables aléatoires de distribution jointe $p(\mathrm{Y}, \theta)$. La statistique $\mathrm{U} = \mathrm{h}(\mathrm{Y})$ est suffisante pour θ si, et seulement si, il est possible de trouver deux fonctions b et c telles que

$$p(\mathrm{Y}, \theta) = b(\mathrm{h}(\mathrm{Y}), \theta)\, c(\mathrm{Y})$$

statistique suffisante

exemple : $X_i \sim \mathscr{P}ois(\mu)$

- Soit $X_1, X_2, \ldots X_n$ avec $X_i \sim \mathscr{P}ois(\mu)$.
- On va montrer que $h = \sum X_i$ est une statistique suffisante pour μ

Pour tous les entiers non négatifs x_1, x_2, ..., x_n la distribution de probabilité discrète jointe de $X_1, \ldots, X_n$ est donnée par

$$p(\mathrm{x}, \theta) = \prod_{i=1}^{n} \frac{e^{-\mu}\, \mu^{x_i}}{x_i\,!} = \left(\prod_{i=1}^{n} \frac{1}{x_i\,!} \right) e^{-n\mu} \mu^{y}$$

avec

$$y = \sum_{i=1}^{n} x_i$$

Le résultat suit du théorème de factorisation.

rapport de vraisemblance et suffisance

proposition

L'estimateur selon le maximum de vraisemblance $\hat{\theta}_{\mathrm{MLE}}$ est fonction de toute statistique suffisante pour θ.

La démonstration est une conséquence du théorème de factorisation Soit $\mathrm{U} = \mathrm{h}(\mathrm{Y})$ une statistique suffisante pour θ. On a

$$p_{\mathrm{Y}}(\mathrm{y}, \theta) = b(\mathrm{h}(\mathrm{y}), \theta) c(\mathrm{y})$$

$$l(\theta) = \log b(\mathrm{h}(\mathrm{y}, \theta)) + \log c(\mathrm{y})$$

Maximiser $l(\theta)$ revient à maximiser $b(\mathrm{h}, \theta)$. Ainsi le maximum de vraisemblance dépend de θ seulement à travers la dépendance de b par rapport à θ et b est fonction de l'observation y qu'au travers de la statistique $\mathrm{h}(\mathrm{y})$.

suffisance minimale

définition

Une statistique suffisante U est dite **minimale** si elle est une fonction de toute autre statistique suffisante.

- Cela signifie que U extrait l'information sur θ avec une **perte minimale de données**.
- Toute autre statistique suffisante contient au moins autant d'information que U, mais pas moins.

famille exponentielle

définition

Soit Y une variable aléatoire dont la distribution dépend d'un seul paramètre θ et qui est de la forme

$$f(y, \theta) = A(\theta) \cdot B(y) \cdot e^{c(\theta) \cdot d(y)}$$

est appelée une famille exponentielle, avec A, B, c et d des fonctions connues.

variables indépendantes et famille exponentielle

Soit $Y_1, \ldots Y_n$, n variables aléatoires indépendantes qui sont individuellement distribuée par un membre identique de la famille exponentielle, la densité jointe s'écrit :

$$f(y_1, y_2, \ldots, y_n) = A(\theta)^n \left[\prod_{i=1}^{n} B(y_i) \right] \exp \left\{ c(\theta) \cdot \sum_{i=1}^{n} d(y_i) \right\}$$

théorème de Rao-Blackwell

idée principale

On peut améliorer un estimateur en le **conditionnant** par une statistique suffisante.

- Soit $X_1, \ldots, X_n$ un échantillon avec loi dépendant d'un paramètre θ.
- Soit W un estimateur de θ, pas forcément optimal.
- Soit T une statistique suffisante pour θ.
- Alors : $\boxed{\hat{\theta} = \mathbb{E}[W \mid T]}$ est **meilleur ou égal** à W (même espérance mais variance plus petite).

théorème de Rao-Blackwell

L'estimateur $\mathbb{E}[W \mid T]$ est meilleur que W en variance quadratique, et a même espérance.

exemple : Rao-Blackwell sur une Bernoulli

Soit $X_1, \ldots, X_n \sim \text{Bernoulli}(\theta)$, avec $\theta \in [0, 1]$.

- Considérons un estimateur naïf : $W = X_1$.
- On sait que $T = \sum_{i=1}^{n} X_i$ est une statistique suffisante.
- Rao-Blackwell propose d'améliorer W :

$$\hat{\theta} = \mathbb{E}[X_1 \mid T]$$

- En utilisant la symétrie entre les X_i, on obtient :

$$\mathbb{E}[X_1 \mid T] = \frac{T}{n}$$

Conclusion

L'estimateur T/n est meilleur que X_1 : même espérance, variance plus faible.

illustration : variances de X_1 et T/n

$X_1, \ldots, X_n$, $X_i \sim \mathscr{B}er(\theta)$

- $X \sim \mathscr{B}in(n, \theta)$, $X_1 \sim \mathscr{B}er(\theta)$
- X_1 est un estimateur de θ, avec :

$$\mathbb{E}[X_1] = \theta, \quad \mathrm{Var}(X_1) = \theta(1-\theta)$$

- $\hat{\theta} = T/n = \frac{1}{n}\sum X_i$ est aussi sans biais :

$$\mathbb{E}[T/n] = \theta, \quad \mathrm{Var}(T/n) = \frac{\theta(1-\theta)}{n}$$

- Donc :

$$\mathrm{Var}(T/n) = \frac{1}{n}\mathrm{Var}(X_1)$$

Exemple numérique pour Rao-Blackwell

On considère $X_1, X_2, X_3 \overset{\text{iid}}{\sim} \mathcal{B}er/(p)$. On veut estimer p, et on définit :

$$T = X_1 + X_2 + X_3 \sim \mathcal{B}in(3, p)$$

Soit $\hat{p}_1 = X_1$ un estimateur naïf. On construit l'estimateur amélioré :

$$\hat{p}_2 = \mathbb{E}[X_1 \mid T]$$

On peut expliciter la loi de $X_1 \mid T = t$ pour $t = 0, 1, 2, 3$. Étant donné la symétrie des X_i, on a :

$T = t$	$P(X_1 = 1 \mid T = t)$	$\mathbb{E}[X_1 \mid T = t]$
0	0	0
1	$\frac{1}{3}$	$\frac{1}{3}$
2	$\frac{2}{3}$	$\frac{2}{3}$
3	1	1

$$\Rightarrow \hat{p}_2 = \mathbb{E}[X_1 \mid T] = \frac{T}{3}$$

Comparaison des variances

On compare la variance des deux estimateurs :

$$\hat{p}_1 = X_1 \quad \text{et} \quad \hat{p}_2 = \mathbb{E}[X_1 \mid T] = \frac{T}{3}$$

Variance de $\hat{p}_1$:

$$\mathrm{Var}(\hat{p}_1) = \mathrm{Var}(X_1) = p(1-p)$$

Variance de $\hat{p}_2$:

$$\mathrm{Var}(\hat{p}_2) = \mathrm{Var}\left(\frac{T}{3}\right) = \frac{1}{9}\mathrm{Var}(T)$$

Or $T = X_1 + X_2 + X_3 \sim \mathrm{Bin}(3, p) \Rightarrow \mathrm{Var}(T) = 3p(1-p)$, donc :

$$\mathrm{Var}(\hat{p}_2) = \frac{1}{9} \cdot 3p(1-p) = \frac{1}{3}p(1-p)$$

Conclusion :

$$\mathrm{Var}(\hat{p}_2) = \frac{1}{3}\mathrm{Var}(\hat{p}_1)$$

$\Rightarrow$ Rao-Blackwell réduit bien la variance.

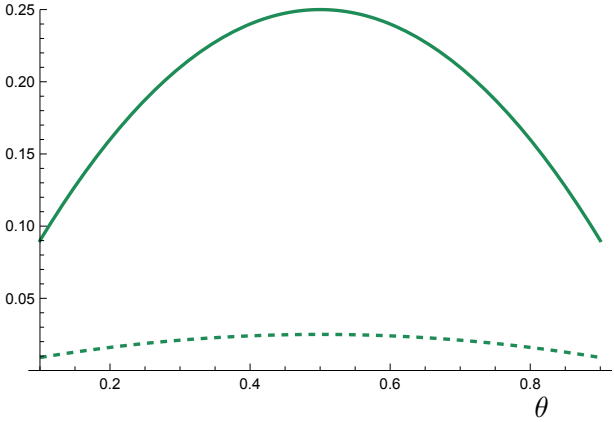

Figure – En trait plein la $\mathrm{Var}(X_1)$ et en traitillé $\mathrm{Var}(T/n)$. Ici $n = 10$.

théorème de Rao-Blackwell

théorème de Rao-Blackwell

Soit une variable aléatoire Y associée à un échantillon. Soit U une statistique et T une statistique suffisante pour θ. Définissons $S = \mathbb{E}(U|T)$. On a :

$$\mathrm{MSE}_\theta(S) \leq \mathrm{MSE}_\theta(U)$$

information de Fisher

- Soit X une variable aléatoire.
- Soit $f(,\theta)$ la densité de probabilité qui dépend d'un paramètre dont la valeur est inconnue mais doit se situer dans un intervalle ouvert de l'espace des paramètres Θ.
- X prend des valeurs dans $\mathcal{X}$ et $f(x,\theta) > 0$ pour chaque valeur de $x \in \mathcal{X}$ et chaque valeur de $\theta \in \Theta$.

information de Fisher

rappel

$$
\begin{aligned}
L(\theta, x) &\triangleq k \cdot f(x, \theta) \\
l(\theta, x) &\triangleq \log L(\theta, x) \\
S(\theta, x) &\triangleq \frac{\partial l(\theta, x)}{\partial \theta} \\
I(\theta, x) &\triangleq -\frac{\partial^2 l(\theta, x)}{\partial \theta^2}
\end{aligned}
$$

dans la suite on considère le cas simple $k = 1$

information de Fisher

définition

l'information attendue de Fisher est définie par :

$$\boxed{\mathcal{I}(\theta, x) \triangleq \mathbb{E}_\theta[I(\theta, x)]}$$

information de Fisher

exemple : distribution de Bernouilli $\mathscr{B}er(p)$

$$\theta = p$$

$$l(x,p) = \log f(x,p) = x \log p + (1-x)\log(1-p)$$

$$l''(x,p) = -\left(\frac{x}{p^2} + \frac{1-x}{(1-p)^2}\right)$$

Comme $\mathbb{E}[X] = p$ on a

$$\mathcal{I}(p) = -\mathbb{E}_\theta[l''(X,p)] = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$$

information de Fisher

exemple : distribution normale $\mathcal{N}(\mu, \sigma^2)$

Soit la variable aléatoire $X \sim \mathcal{N}(\mu, \sigma^2)$ de moyenne $-\infty < \mu < \infty$ inconnue et de variance σ^2 connue.
Déterminons l'information $\mathcal{I}(\mu)$ dans X. Pour $-\infty < x < +\infty$,

$$I(x, \mu) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

de telle sorte que

$$I'(x, \mu) = \frac{x-\mu}{\sigma^2}$$

et

$$I''(x, \mu) = -\frac{1}{\sigma^2}$$

et l'information de Fisher s'écrit donc

$$\mathcal{I}(\mu) = -I''(x, \mu) = \frac{1}{\sigma^2}$$

information de Fisher

d'un échantillon

Supposons un ensemble de variables aléatoires X_1, X_2, ..., X_n dont une réalisation de chacune des variables donne un échantillon de taille n. Toutes les variables aléatoires sont identiquement distribuées et indépendantes.
On peut donc multiplier les probabilités

$$\begin{aligned} f_n(x,\theta) &= f_1(x,\theta)\, f_2(x,\theta) \cdots f_n(x,\theta) \\ &= f(x,\theta)\, f(x,\theta) \cdots f(x,\theta) \end{aligned}$$

et ainsi additionner les informations.

$$\mathcal{I}_n(\theta) = n\,\mathcal{I}(\theta)$$

borne de Cramér-Rao

inégalité concernant l'information

une définition équivalente l'information consiste à poser

$$\mathcal{I}(\theta) \triangleq \mathbb{E}_\theta[(l'(x,\theta))^2]$$

si $f(x,\theta)$ est une densité de probabilité, on a

$$\mathcal{I}(\theta) = \int_S (l'(x,\theta))^2 \, f(x,\theta) \, dx$$

On sait que

$$\int_s f(x,\theta)dx = 1$$

pour chaque valeur de $\theta \in \theta$. On suppose que l'on peut permuter l'opération de dérivation par rapport à θ et d'intégration par rapport à x, et donc

$$\int_S f'(x,\theta)\, dx = 0 \qquad \theta \in \Theta \tag{3}$$

En prenant une dérivée supplémentaire

$$\int_S f''(x,\theta) = 0 \qquad \theta \in \Theta$$

Comme $l'(x,\theta) = \frac{f'(x,\theta)}{f(x,\theta)}$

$$\mathbb{E}_\theta[l'(x,\theta)] = \int_S l'(x,\theta)\, f(x,\theta) = \int_S f'(x,\theta), dx$$

et donc

$$\mathbb{E}_\theta(l'(X,\theta)] = 0$$

Comme la moyenne de $l'(X,\theta)$ est 0 (le raisonnement est le suivant : le logarithme de vraisemblance est maximisé et en moyenne la pente du logarithme de vraisemblance est alors nulle), et comme $\mathcal{I} = \mathbb{E}_\theta[(l'(X,\theta))^2]$ on a à cause de (3) on a

$$\mathcal{I}(\theta) = \mathrm{Var}_\theta[l'(x,\theta)]$$

de plus

$$l''(x,\theta) = \frac{f(x,\theta)\, f''(x,\theta) - (f'(x,\theta))^2}{(f(x,\theta))^2} = \frac{f''(x,\theta)}{f(x,\theta)} - (l'(x,\theta))^2$$

et finalement

$$\mathbb{E}[l''(X,\theta)] = \int_S f''(x,\theta)dx - \mathcal{I}(\theta)$$

borne de Cramér-Rao

inégalité concernant l'information

Soit un ensemble de variables aléatoires indépendantes X_1, X_2, ..., X_n identiquement distribuées de densité de probabilité $f(x, \theta)$ avec θ un paramètre appartenant à un intervalle ouvert.
On notera $f_n(\mathrm{x}, \theta)$ la densité de probabilité jointe de X_1, X_2, ..., X_n.

borne de Cramér-Rao

inégalité concernant l'information

Soit

$$T = r(X_1, \ldots, X_n) = r(\mathrm{X})$$

un estimateur arbitraire de θ pour lequel la variance est finie.
Considérons la covariance entre T et la variable aléatoire $l_n'(\mathrm{X}, \theta)$. Comme $l_n'(\mathrm{x}, \theta) = f_n'(\mathrm{x}, \theta)/f_n(\mathrm{x}, \theta)$ il s'ensuit que pour une observation unique

$$\mathbb{E}_\theta[l'(\mathrm{X}, \theta)] = \int_S \ldots \int_S f_n'(\mathrm{x}, \theta) dx_1 \ldots dx_n = 0$$

En conséquence

$$\begin{aligned} \mathrm{Cov}_\theta[T, l_n'(\mathrm{X}, \theta)] &= \mathbb{E}_\theta[T\, l'(\mathrm{X}, \theta)] \\ &= \int_S \cdots \int_S r(\mathrm{x})\, l'(\mathrm{x}, \theta)\, f_n(\mathrm{x}, \theta)\, dx_1 \ldots dx_n \\ &= \int_S \cdots \int_S r(\mathrm{x})\, f_n'(\mathrm{x}, \theta)\, dx_1 \ldots dx_n \end{aligned}$$

borne de Cramér-Rao

Maintenant $\mathbb{E}_\theta(T) = m(\theta)$ pour $\theta \in \Theta$ et donc

$$\int_S \dots \int_S r(\mathrm{x})\, f_n(\mathrm{x}, \theta)\, dx_1 \dots dx_n = m(\theta) \qquad \text{pour} \theta \in \Theta$$

En dérivant par rapport à θ cette dernière équation en prenant la dérivée dans l'intégrale

$$\int_S \dots \int_S r(\mathrm{x})\, f_n^{'}(\mathrm{x}, \theta)\, dx_1 \dots dx_n = m^{'}(\theta) \qquad \text{pour} \theta \in \Theta$$

et donc

$$\mathrm{Cov}_\theta[T, l^{'}(\mathrm{X}, \theta)] = m^{'}(\theta) \qquad \theta \in \Theta$$

par les propriétés de la covariance (cf. leçon suivante, Cauchy-Schwarz))

borne de Cramér-Rao

inégalité de l'information

$$\{\mathrm{Cov}[T\, l'(\mathrm{X},\theta)]\}^2 \quad \leq \quad \mathrm{Var}_\theta(T)\,\mathrm{Var}_\theta[l'(\mathrm{X},\theta)]$$

$$\boxed{\mathrm{Var}_\theta(T) \geq \frac{[m'(\theta)]^2}{n\mathcal{I}(\theta)}}$$