Module 3
Estimation paramétrique

Sommaire

Dernière mise à jour le 3 janvier 2019

3.1 Introduction

Nous avons étudié dans le module 2 la théorie de décision de Bayes dont l'objectif était de déterminer la règle de décision optimale en utilisant des modèles probabilistes. L'objectif de ce module est d'estimer ces probabilités d'une manière paramétrique afin de pouvoir appliquer ces règles de décision.

Les méthodes **d'estimations paramétriques** supposent que nous connaissons la forme de la densité de probabilité, c'est-à-dire que $\mathbf{x}$ provient de la classe ω_i dont la densité est notée $p(\mathbf{x}|\omega_i)$. L'objectif consiste, donc, à estimer les paramètres de $p(\mathbf{x}|\omega_i)$ à partir d'un ensemble d'échantillons d'apprentissage.

Il existe plusieurs méthodes pour l'estimation des paramètres décrivant les densités paramétriques. Dans ce module, nous allons détailler deux méthodes couramment utilisées, à savoir :

1. L'estimateur du maximum de vraisemblance.
2. L'estimateur bayésien.

3.2 Estimateur du maximum de vraisemblance

L'estimateur du maximum de vraisemblance est une méthode d'estimation attrayante, d'une part parce qu'elle présente de bonnes propriétés de convergence lorsque le nombre d'échantillons augmente, et d'autre part, parce qu'elle est simple à implémenter.

3.2.1 Principe général

Soit $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_c$, c ensemble d'échantillons provenant de c classes de formes $\omega_1, \omega_2, ..., \omega_c$, respectivement. Supposons que les échantillons de la classe ω_i suivent une loi de probabilité

$p(\mathbf{x}|\omega_i)$, $i = 1, ..., c$ et que la forme paramétrique de $p(\mathbf{x}|\omega_i)$ est déterminée par le vecteur de paramètre $\boldsymbol{\theta}_i = (\theta_1, \theta_2, ..., \theta_p)$ avec p la dimension du vecteur de paramètre.

▶ **Exemple 4.1** Supposons que les échantillons provenant d'une classe de formes ω suivent une densité normale multivariée $\mathcal{N}$ de moyenne $\boldsymbol{\mu}$ et de variance $\boldsymbol{\Sigma}$. Dans ce cas, le vecteur de paramètre $\boldsymbol{\theta}$ de la classe ω est donné par :

$$\text{Si} \quad p(\mathbf{x}|\omega) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{Alors} \quad \boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\Sigma} \end{pmatrix}.$$

Pour rendre explicite la dépendance de la densité $p(\mathbf{x}|\omega_i)$ de $\boldsymbol{\theta}_i$, nous noterons cette densité $p(\mathbf{x}|\omega_i, \boldsymbol{\theta}_i)$.

Pour simplifier le problème, on suppose que l'ensemble des échantillons $\mathcal{D}_i$ de la classe ω_i ne fournissent aucune information sur $\boldsymbol{\theta}_j$ si $j \neq i$ [2]. Nous pouvons donc traiter les classes séparément et s'affranchir de la notation des symboles de classes. Le problème consiste, alors, à utiliser un ensemble de n échantillons $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ de densité $p(\mathbf{x}|\boldsymbol{\theta})$ pour estimer $\boldsymbol{\theta}$.

Nous supposons que ces échantillons sont indépendants, nous obtenons alors :

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta}). \tag{3.1}$$

avec $p(\mathcal{D}|\boldsymbol{\theta})$ la vraisemblance de θ par rapport à $\mathcal{D}$.

L'estimateur du maximum de vraisemblance est celui qui permet de déterminer $\hat{\boldsymbol{\theta}}$ qui maximise $p(\mathcal{D}|\boldsymbol{\theta})$. Pour des raisons analytiques, il est préférable de considérer le logarithme de la vraisemblance aussi appelé *log-vraisemblance* ou *log likelihood*.

$$l(\boldsymbol{\theta}) = \ln p(\mathcal{D}|\boldsymbol{\theta}). \tag{3.2}$$

Le logarithme est une fonction monotone croissante, alors la valeur de $\hat{\boldsymbol{\theta}}$ qui maximise le log-vraisemblance maximise aussi la vraisemblance.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}). \tag{3.3}$$

Soit le vecteur de paramètre $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^t$ de dimension p et $\nabla_{\boldsymbol{\theta}}$ l'opérateur gradient,

$$\nabla_{\boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial}{\partial\theta_1} \\ \frac{\partial}{\partial\theta_2} \\ . \\ . \\ \frac{\partial}{\partial\theta_p} \end{bmatrix}$$

Le *log-vraisemblance* est donné par :

$$\begin{aligned} l(\boldsymbol{\theta}) &= \ln p(\mathcal{D}|\boldsymbol{\theta}) \\ &= \ln \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta}) \\ &= \sum_{j=1}^{n} \ln p(\mathbf{x}_j|\boldsymbol{\theta}) \end{aligned} \tag{3.4}$$

D'où :

$$\nabla_{\boldsymbol{\theta}} l = \sum_{j=1}^{n} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_j|\boldsymbol{\theta}) \tag{3.5}$$

Les équations d'estimation de $\hat{\boldsymbol{\theta}}$, peuvent alors être obtenues à partir de :

$$\nabla_{\boldsymbol{\theta}} l = 0 \tag{3.6}$$

c'est-à-dire en annulant les dérivées partielles de la fonction de vraisemblance par rapport à chaque composante θ_i du vecteur du paramètre $\boldsymbol{\theta}$.

L'estimateur de maximum de vraisemblance s'obtient donc en annulant les dérivées partielles de la fonction de vraisemblance par rapport à chaque composante θ_i.

Opérateur gradient

$$\nabla f(x, y, z) = \begin{bmatrix} \frac{\partial}{\partial x} \\ \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} \end{bmatrix}$$

La fonction logarithme népérien

$$\ln(x \times y) = \ln x + \ln y$$
$$\ln(\tfrac{x}{y}) = \ln x - \ln y$$
$$\ln x^a = a \ln x$$

3.2.2 Cas de densité normale : $\boldsymbol{\mu}$ inconnu

On considère une distribution normale multivariée $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})]$$

Supposons, dans un premier temps qu'uniquement la moyenne $\boldsymbol{\mu}$ est inconnue,

$$\ln p(\mathbf{x}_k|\boldsymbol{\mu}) = -\frac{1}{2}\ln[(2\pi)^d|\boldsymbol{\Sigma}|] - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{3.7}$$

et

$$\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k|\boldsymbol{\mu}) = \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \boldsymbol{\mu})$$

Nous pouvons identifier $\boldsymbol{\theta}$ par $\boldsymbol{\mu}$ puisque c'est la seule inconnue. Nous aurons alors :

$$\sum_{k=1}^{n} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}) = 0$$

En multipliant par $\boldsymbol{\Sigma}$ et en arrangeant les termes, on obtient :

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n} x_k$$

Ce résultat est très satisfaisant parce qu'il peut être interprété par : l'estimateur du maximum de vraisemblance pour une densité de probabilité normale de moyenne inconnue est la moyenne arithmétique des échantillons observés.

3.2.3 Cas de densité normale : $\boldsymbol{\mu}$ et $\boldsymbol{\Sigma}$ unconnus

Dans le cas d'une densité normale multivariée, généralement, les valeurs de $\boldsymbol{\mu}$ et $\boldsymbol{\Sigma}$ sont tous les deux inconnues. Ces deux paramètres constituent les deux composantes du vecteur

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\Sigma} \end{pmatrix}.$$

Cas de densité uni variée

Considérons le cas d'une densité uni variée avec $\theta_1 = \mu$ et $\theta_2 = \sigma^2$. L'équation 3.7 devient :

$$\ln p(\mathbf{x}_k|\boldsymbol{\theta}) = -\frac{1}{2}\ln(2\pi)\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$$

et sa dérivée devient :

$$\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k|\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2} \end{bmatrix}.$$

En appliquant la relation de l'équation 3.6, on obtient :

$$\begin{cases} \sum_{j=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 \\ \\ \sum_{j=1}^{n}(-\frac{1}{2\hat{\theta}_2} + \frac{(x_k-\hat{\theta}_1)^2}{2\hat{\theta}_2^2}) = 0 \end{cases}$$

Avec $\hat{\theta}_1$ et $\hat{\theta}_2$ les estimateurs du maximum de vraisemblance de θ_1 et θ_2 respectivement. En remplacant $\hat{\theta}_1$ par $\hat{\mu}$ et $\hat{\theta}_2$ par $\hat{\sigma}^2$, nous obtenons les estimateurs de maximum de vraisemblance de μ et de σ.

$$\begin{cases} \hat{\mu} = \frac{1}{n}\sum_{j=1}^{n} x_j \\ \\ \hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \hat{\mu})^2 \end{cases}$$

Le détail des calculs mathématiques qui permettent d'aboutir aux résultats précédents sera présenté dans les exercices du module 3.

Cas de densité multivariée

Dans le cas d'une densité multivariée, les analyses sont similaires, nous vous épargnions des calculs mathématiques un peu plus complexes pour ce cas plus général. Mais sachez

que les résultats obtenus pour une densité multivariée sont similaires au cas uni varié et les estimateurs de maximum de vraisemblance de μ et de σ sont :

$$\begin{cases} \hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_j \\ \\ \hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{j=1}^{n} (\mathbf{x}_j - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t \end{cases}$$

3.3 Estimateurs bayésiens

3.3.1 Principe général

L'objectif général des méthodes paramétriques consiste, rappelons-le, à utiliser un ensemble de n échantillons $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ de densité $p(\mathbf{x}|\boldsymbol{\theta})$ pour estimer le paramètre $\boldsymbol{\theta}$. À la différence de l'estimateur par maximum de vraisemblance, les estimateurs bayésiens supposent que nous disposons d'une information à priori sur le vecteur $\boldsymbol{\theta}$ résumé par sa densité à priori $p(\boldsymbol{\theta})$.

La règle de décision de Bayes permet d'écrire la densité a posteriori du paramètre $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{x})} \tag{3.8}$$

avec

$$p(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta}).$$

et

$$p(\mathbf{x}) = \int p(x|\boldsymbol{\theta})p(\boldsymbol{\theta})d(\boldsymbol{\theta})$$

Les estimateurs bayésiens les plus connus sont l'estimateur de la moyenne a posteriori et l'estimateur du maximum a posteriori.

3.3.2 Estimateur de la moyenne a posteriori (MMSE)

L'estimateur MMSE (*Minimum mean square error*) de $\boldsymbol{\theta}$ est une fonction de $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}^t$. Cet estimateur noté $\hat{\boldsymbol{\theta}}_{MMSE}(\mathbf{x})$ se base sur la minimisation de la fonction coût suivante :

$$c\left(\hat{\boldsymbol{\theta}}(\mathbf{x})\right) = E\left[\left(\hat{\boldsymbol{\theta}}_{MMSE}(\mathbf{x}) - \hat{\boldsymbol{\theta}}\right)^2\right] \tag{3.9}$$

Sans rentrer dans les détails du calcul mathématique, nous obtenons :

$$\hat{\boldsymbol{\theta}}_{MMSE}(\mathbf{x}) = E\left[\boldsymbol{\theta}|\mathbf{x}\right] \tag{3.10}$$

L'estimateur MMSE est donc la moyenne de la densité a posteriori du paramètre $\boldsymbol{\theta}$.
Nous invitons les lecteurs qui désirent se documenter sur le calcul mathématique de l'équation (3.10) à consulter la référence [1].

3.3.3 Estimateur du maximum a posteriori (MAP)

L'estimateur MAP de $\boldsymbol{\theta}$ est une fonction de $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}^t$. Cet estimateur noté $\hat{\boldsymbol{\theta}}_{MAP}(\mathbf{x})$ se base sur la minimisation de la fonction coût suivante [1] :

$$c\left[\left(\hat{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta}\right)\right] = \begin{cases} 1 & \text{Si } \left|\hat{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta}\right| > \frac{\Delta}{2} \\ 0 & \text{Sinon} \end{cases} \tag{3.11}$$

Où Δ est un paramètre choisi arbitrairement.
Sans rentrer dans les détails de calcul mathématique, nous obtenons :

$$\hat{\boldsymbol{\theta}}_{MAP}(\mathbf{x}) = \arg\max_{\boldsymbol{\theta}} \quad p(\boldsymbol{\theta}|\mathbf{x}) \tag{3.12}$$

L'estimateur MMSE est donc obtenu en cherchant le mode de la densité a posteriori du paramètre $\boldsymbol{\theta}$.

3.4 Annexes

Dans cette section, nous citons des fonctions R utiles pour l'estimation de densité de probabilité en utilisant les méthodes paramétriques.

— `mle` (*maximum likelihood estimation*) : est la fonction qui permet d'estimer les paramètres en utilisant l'estimateur du maximum de vraisemblance.
https://stat.ethz.ch/R-manual/R-devel/library/stats4/html/mle.html

— `dnorm` : Permet de générer une densité normale
https://stat.ethz.ch/R-manual/R-devel/library/stats/html/Normal.html

Références

[1] Estimation bayésienne. tourneret.perso.enseeiht.fr/MODAP/Estimation_bayesienne.pdf. Consulté : 1 janvier 2019.

[2] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2000.

