Modélisation stochastique, Estimation

Stephan Robert, HEIG-Vd

14 décembre 2009

Introduction

Introduction

But : Estimation de paramètres de lois de probabilités étant donné un certain nombre d'observations.

Exemples : Moyenne, variance, corrélation, densitié spectrale

Applications : Calcul du BER (Bit Error Rate) d'une fibre optique, UWB (Ultra Wide Band), signaux radar.

Notations :

- Si $\theta = E[X] = \mu$, alors son estimateur est noté $\bar{X}$ ou $\hat{\mu}$.
- Si $\theta = Var[X]$ alors son estimateur est noté $\widehat{Var(X)}$ ou $\widehat{\sigma^2}$.
- Si $\theta = R_X(t,s)$ alors son estimateur est noté $\widehat{R_X(t,s)}$.
- $X_1, X_2, \ldots$: variables aléatoires, $x_1, x_2, \ldots$: réalisations.

Exemple

BER d'une fibre optique : Combien de bits n doit-on observer pour obtenir une valeur fiable de p ?

Hypothèse : Le processus des erreurs est Bernoullien :

$$X_n = 1 : n^{\text{ième}} \text{ bit est corrompu}$$

$$X_n = 0 : n^{\text{ième}} \text{ bit n'est pas corrompu}$$

Nous avons :

- $P(X_n = 1) = p$
- $P(X_n = 0) = 1 - p$

Construction d'un estimateur de $\theta = E[X] = p$: $\hat{\theta}$

Exemple (2)

$$\hat{\theta} = \bar{X_n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Par le théorême central limite nous savons que

$$\boxed{\bar{X_n} \rightarrow \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)}$$

ou

$$\left(\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}}\right) \rightarrow \mathcal{N}\left(0, 1\right)$$

Exemple (3)

lorsque n devient grand. Donc

$$P\left(|\bar{X}-p| \leq z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$$

avec

$$\Psi(z_{\alpha/2}) = 1-\alpha/2$$

et

$$\Psi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$$

Exemple (4)

Avec une probabilité égale à $1-\alpha$, la valeur recherchée de l'estimateur de $E[X]=\mu$ est comprise dans l'intervalle

$$[\bar{X}-z_{\alpha/2}\sqrt{p(1-p)/n}, \bar{X}+z_{\alpha/2}\sqrt{p(1-p)/n}]$$

qui constitue un intervalle de confiance de μ au seuil $1-\alpha$.

Remarque : L'intervalle **diminue** avec le **nombre d'échantillons collectés**.

Rappel :

$1-\alpha$	$\alpha/2$	$z_{\alpha/2}$
0.6826	0.1587	1.28
0.90	0.05	1.645
0.95	0.025	1.96
0.99	0.005	2.576
0.999	0.0005	3.291

Exemple (4)

Pour un seuil de confiance de $1-\alpha$, le nombre de mesures à collecter est de

$$n = \frac{1-p}{p}\left(\frac{z_{\alpha/2}}{a}\right)^2$$

a : constante positive qui défini la précision avec laquelle nous aimerions obtenir la valeur du paramètre μ.

Application numérique :

Combien d'échantillons faut-il collectionner pour obtenir l'espérance mathématique avec $p = 0.4$ et une précision de $\pm 5\%$ (a=0.05) et un **seuil de confiance de 99.9%** $(1-\alpha = 0.999)$?

Exemple (5)

Réponse :

$$n = \frac{1-p}{p}\left(\frac{z_{\alpha/2}}{a}\right)^2 = \frac{1-0.4}{0.4}\left(\frac{3.291}{0.05}\right)^2 = 6498$$

Cas de la fibre optique

Combien de temps faut-il attendre avant d'obtenir le résultat ?

- Ligne en fibre optique **fonctionnant à 1 Gb/s**
- BER $\approx 10^{-9}$.
- Estimation de p à $\pm 5\%$, seuil de confiance de 99%.

Réponse :

$$n = \frac{1-10^{-9}}{10^{-9}}\left(\frac{2.576}{0.05}\right)^2 = 2.65.10^{12} \text{ bits} = 2654 \text{ sec.} \approx 44 \text{ min.}$$

Remarque : Ne pas négliger le temps nécessaire ! (simulations par exemple)

Exercice

Dans le cadre de l'estimation de p de la fibre optique, que doit valoir n si p est dix fois plus petit ? Et si la précision augmente $(\alpha/2 = 0.025)$?

Propriétés des estimateurs

- Paramètre à estimer : θ
- Estimateur : $\hat{\theta}$
- Nous voulons que $E[\hat{\theta}] \to \theta$

Biais d'un estimateur $\hat{\theta}$: $E[\hat{\theta}] - \theta$.
Estimateur **non biaisé** : $E[\hat{\theta}] = \theta$

Exemple 1 : Estimation de la moyenne

$X_1, X_2, ..., X_n$ iid, mêmes μ et σ^2, alors $\hat{\theta} = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$E[\hat{\theta}] = E[\hat{\mu}] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{n\mu}{n} = \mu$$

Exemple 2 : Estimation de la variance

$$\hat{\theta} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)^2$$

$$\begin{aligned}
E[\widehat{\sigma^2}] &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n}X_j\right)^2\right] \\
&= \frac{1}{n-1}\left(n\sigma^2 + n\mu^2 - \sigma^2 - \mu^2 - n\mu^2 + \mu^2\right) \\
&= \frac{1}{n-1}\left((n-1)\sigma^2\right)
\end{aligned}$$

et donc

$$E[\widehat{\sigma^2}] = E[\hat{\theta}] = \frac{n-1}{n-1}\sigma^2 = \sigma^2$$

Propriétés des estimateurs (2)

Consistance d'un estimateur $\hat{\theta}$:

$$\lim_{n\to\infty} P\left(\hat{\theta} - \theta| > \epsilon) = 0\right), \forall \epsilon > 0 \text{ quand } n \to \infty.$$

Exemple 1 : Consistance de l'estimateur de la moyenne
Inégalité de Tchebysheff :

$$P(|\hat{\theta} - \mu| > \epsilon) \leq \frac{Var(\hat{\theta})}{\epsilon^2}$$

Propriétés des estimateurs (3)

$$
\begin{aligned}
Var(\hat{\theta}) &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right]^2 \\
&= E\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j - 2\mu\frac{1}{n}\sum_{i=1}^{n} X_i + \mu^2\right] \\
&= E\left[\frac{1}{n^2}\sum_{i=1}^{n} X_i^2 + 2\sum_{i=1}^{n}\sum_{j>i} X_i X_j - 2\mu\frac{1}{n}\sum_{i=1}^{n} X_i + \mu^2\right] \\
&= \frac{1}{n^2}(n\sigma^2 + n\mu^2) + 2\frac{1}{n^2}\frac{n(n-1)\mu^2}{2} + \mu^2 - 2\mu^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}
$$

Propriétés des estimateurs (4)

Nous avons donc

$$P(|\hat{\theta} - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \to 0$$

lorsque $n \to \infty$, $\forall \epsilon > 0$.

Propriétés des estimateurs (5)

Le meilleur estimateur est celui qui possède la **variance la plus faible** !

Tout estimateur qui atteint la borne de Cramer-Rao est dit **efficace**.

Qu'est-ce qu'est la borne de Cramer-Rao ? Pour un estimateur sans biais et régulier (voir cours), alors :

$$E[\hat{\theta} - \theta]^2 \geq \frac{1}{\mathcal{I}(\theta)}$$

avec

$$\mathcal{I}(\theta) = -E\left[\left(\frac{\partial^2 \log f(x_1, x_2, ..., x_n; \theta)}{\partial \theta^2}\right)\right]$$

Exemple

Soient n variables aléatoires gaussiennes iid $X_1, X_2, ..., X_n$ avec

$$f_{X_i}(x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\left(\frac{x_i - \theta}{\sqrt{2}\sigma}\right)^2\right]$$

Donc

$$f_{X_1, X_2, ... X_n}(x_1, x_2, ..., x_n; \theta) =$$

$$\frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - \theta}{\sigma}\right)^2\right]$$

Exemple

En prenant le logarithme naturel il vient :

$\log f(x_1, x_2, ..., x_n; \theta) =$

$$= \log\left[\frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\theta}{\sigma}\right)^2\right]\right]$$

$$= -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2$$

En dérivant cette fonction par rapport à θ :

$\frac{\partial}{\partial\theta}\log f(x_1, x_2, ..., x_n; \theta) =$

$$= \frac{\partial}{\partial\theta}\left(-\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right)$$

Exemple

$$\frac{\partial}{\partial \theta} \log f(x_1, x_2, ..., x_n; \theta) =$$

$$\begin{aligned} &= \frac{2}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta) \\ &= \frac{n}{\sigma^2} \left(\frac{\sum_{i=1}^{n} x_i}{n} - \theta \right) \end{aligned}$$

et en dérivant encore une fois il vient

$$\frac{\partial^2}{\partial \theta^2} \log f(x_1, x_2, ..., x_n; \theta) = \frac{-n}{\sigma^2}$$

La borne de Cramer-Rao est donc égale à σ^2/n

Exemple

Nous savons que la moyenne de l'estimateur de la moyenne

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

est θ et que la variance de l'estimateur de la moyenne est égale à σ^2/n (cours p.199-200). L'estimateur $\hat{\theta}$ est donc un **estimateur efficace** puisque sa variance est égale à la borne de Cramer-Rao.

Exercice 1

Estimateur de Maximum de Vraissemblance

Introduction

En anglais : **Maximum Likelihood Estimator (MLE)**

Définition : Vraissemblance de θ :

$$\boxed{\mathcal{L}(x_1, ..., x_n; \theta) = f(x_1; \theta) ... f(x_n; \theta)}$$

ou

$$\log \mathcal{L}(x_1, ..., x_n; \theta) = \log(f(x_1; \theta) ... f(x_n; \theta)) = \sum_{i=1}^{n} \log(f(x_i; \theta))$$

Maximisation de la fonction de vraissemblance :

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \, \mathcal{L}(\theta)$$

Introduction (2)

Recherche du **maximum** :

$$\frac{\partial \log \mathcal{L}(x_1, ..., x_n; \theta)}{\partial \theta} = 0$$

Condition pour que ce soit effectivement un maximum (et non un “minimum”) :

$$\frac{\partial^2 \log \mathcal{L}(x_1, ..., x_n; \theta)}{\partial \theta^2} \leq 0$$

Exemple : Loi Normale

Souhait : Estimation de **l'espérance mathématique** et la **variance** de la loi de Normale à partir d'un échantillon $x_1, ..., x_n$.

Fonction de densité de probabilité :

$$f_X(x;\theta_1,\theta_2) = f_X(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

Fonction de vraissemblance :

$$f_{X_1,\ldots X_n}(x_1,\ldots,x_n;\mu,\sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left[-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

ou

$$\log\mathcal{L}(x_1,\ldots,x_n;\mu,\sigma) \quad = \quad \frac{n}{2}\log\left(\frac{1}{2\pi\sigma^2}\right) - \sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}$$

Exemple : Loi Normale (2)

Dérivée de la fonction de vraissemblance par rapport au premier paramètre :

$$\frac{\partial \log \mathcal{L}(x_1, ..., x_n; \mu, \sigma)}{\partial \mu} = \sum_{i=1}^{n} \frac{(x_i - \mu)}{\sigma^2} = 0$$

ce qui donne un estimateur sans biais :

$$\hat{\mu} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Exemple : Loi Normale (3)

Dérivée de la fonction de vraissemblance par rapport au deuxième paramètre :

$$\frac{\partial \log \mathcal{L}(x_1, ..., x_n; \mu, \hat{\sigma})}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i - \hat{\mu})^2}{\sigma^3} = 0$$

ce qui donne

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (X_i - \hat{\mu})^2}{n}$$

Remarque : Cet estimateur est constistant mais **biaisé** !

Exercice 2

Détection, décisions et tests d'hypothèses

Introduction

Un processus de décision, généralement basé sur un certain nombre d'observations, consiste à décider si une **hypothèse est valide ou non**. Le choix fait est binaire : soit l'hypothèse est **acceptée**, soit elle est **rejetée**. Par exemple :

1. Un bit observé sur une ligne numérique est soit "1", soit "0"
2. Un avion est soit présent soit absent dans une région étant donné un signal radar
3. Un voleur est présent ou non dans la maison suivant les capteurs d'alarmes
4. Un patient a une maladie ou non suivant certains symptômes

Introduction (2)

Truc : Diviser l'espace d'observation pour minimiser le risque d'erreurs

Convention :

- H_0 : *hypothèse nulle*
- H_1 : *hypothèse alternative*

Procédure de test : Déterminer si $\theta \in H_0$ ou si $\theta \notin H_0$

Notation (abusive) : $P(H_i) = P(\theta \in H_i)$, $i = 0, 1$

Introduction (3)

Définitions

- *Erreur de première espèce* : rejeter H_0 alors qu'elle est vraie.
- *Erreur de deuxième espèce* : accepter H_0 alors qu'elle est fausse.

But des tests : Minimiser ces erreurs -> compromis à faire ! ! !

Puissance d'un test ($\phi(x)$) : Probabilité de rejeter l'hypothèse nulle à raison. La puissance du test est donc le complément de l'erreur de deuxième espèce.

Introduction (4)

Critère de Bayes :

$$\sum_{i=0}^{1} \sum_{j=0}^{1} c_{ji} P(H_i) \int_{d(x)=j} f(x|H_i) dx$$

Connu : $P(H_0)$ et $P(H_1)$ avec $P(H_0) + P(H_1) = 1$

A trouver : $d(x)$ tel que le coût total soit **minimisé**.

Introduction (5)

Définition Rapport de vraissemblance (*Likelihood ratio*).

$$\boxed{\Lambda(x) = \frac{f(x|H_1)}{f(x|H_0)}}$$

ou

$$\log \Lambda(x) = \log \frac{f(x|H_1)}{f(x|H_0)}$$

Introduction (6)

Théorème

La règle de décision est donnée par

$$d(x) = \mathbf{1}[\Lambda(x) > \tau]$$

avec

$$\tau = \frac{P(H_0)}{P(H_1)} \frac{c_{10}}{c_{01}}$$

et $\mathbf{1}$ étant la fonction indicatrice

Remarque : La preuve est donnée dans le cours.

Exemple de calcul du rapport de vraissemblance

Soit une variable aléatoire distribuée selon l'une ou l'autre des lois de probabilité connues (gaussienne et de Cauchy) ayant les densités de probabilité suivantes :

$$f(x|H_0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{x^2}{2\sigma^2}\right]$$

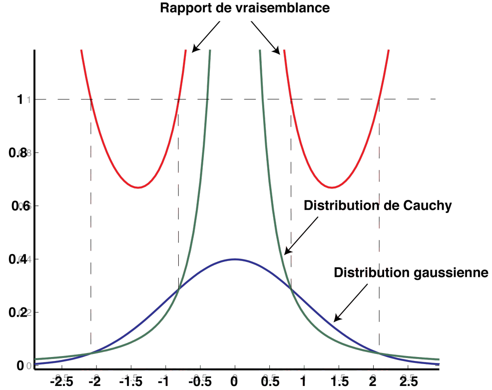
et

$$f(x|H_1) = \frac{a}{a^2 + x^2}$$

avec $a > 0$. Le rapport de vraisemblance est donné par

$$\Lambda(x) = \frac{\sqrt{2\pi\sigma^2 a^2}}{a^2 + x^2} \exp\left[\frac{x^2}{2\sigma^2}\right]$$

Rapport de vraisemblance entre les distributions gaussienne et de Cauchy, avec $\tau = 1$.



Détection d'un signal bruité

Soit une source générant un signal aléatoire binaire $x(t)$ prenant soit la valeur A avec une probabilité de $P(H_1)$ soit la valeur 0 avec une probabilité $P(H_0)$.

Avant de parvenir à destination ce signal est additionné d'un bruit gaussien $n(t)$ de valeur moyenne nulle et de variance σ^2.

A la réception, au temps t, on doit décider à quelle valeur du signal émis correspond la valeur du signal reçu.

Détection d'un signal bruité (2)

Nous avons

$$f(x|H_0) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{x^2}{2\sigma^2}\right]$$

et

$$f(x|H_1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-A)^2}{2\sigma^2}\right]$$

Rapport de vraissemblance :

$$\Lambda(x) = \exp\left[\frac{2xA - A^2}{2\sigma^2}\right]$$

ou

$$\log \Lambda(x) = \frac{2xA - A^2}{2\sigma^2}.$$

Détection d'un signal bruité (2)

Comme nous l'avons vu, la règle de décision est donnée par

$$d(x) \quad = \quad \mathbf{1}[\Lambda(x) > \lambda]$$

Ici il s'agit de chercher le seuil de décision, à savoir x_p lorsque $\Lambda(x) = \lambda$. En cherchant x_p avec

$$\exp\left[\frac{2x_pA - A^2}{2\sigma^2}\right] = \frac{P(H_0)}{P(H_1)}\frac{c_{10}}{c_{01}}$$

il vient

$$x_p = \frac{A}{2} + \frac{\sigma^2}{A}\log\left(\frac{P(H_0)}{P(H_1)}\frac{c_{10}}{c_{01}}\right)$$

et appliquant la règle de décision :

$$\begin{aligned} d(x) \quad &= \quad \mathbf{1}[x > x_p] \\ &= \quad \mathbf{1}\left[x > \frac{A}{2} + \frac{\sigma^2}{A}\log\left(\frac{P(H_0)}{P(H_1)}\frac{c_{10}}{c_{01}}\right)\right] \end{aligned}$$

Détection d'un signal bruité (3)

Application numérique :

$$A = 10,\ \sigma = 3,\ c_{01} = 20,\ c_{10} = 1,$$

$$P(H_0) = 0.8,\ P(H_1) = 0.2$$

Nous trouvons la valeur de $x_p = 3.55$

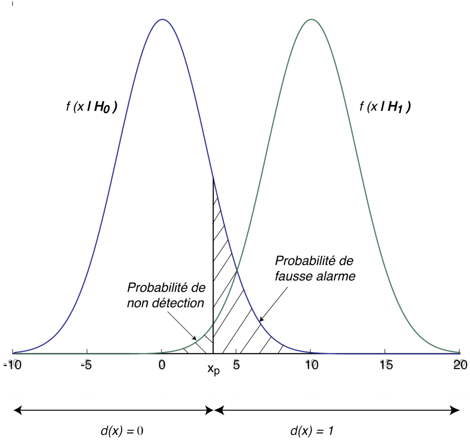
Probabilité de non detection :

$$P(d(x) = 0|H_1) = \int_{-\infty}^{x_p} f(x|H_1) = 0.0158$$

Probabilité de fausse alarme :

$$P(d(x) = 1|H_0) = \int_{x_p}^{\infty} f(x|H_0) = 0.118$$

Détection d'un signal bruité (4)



Test du Minimax

Problème : Si les probabilités $P(H_0)$ et $P(H_1)$ (approche Bayesienne) sont inconnues, que fait-on ?

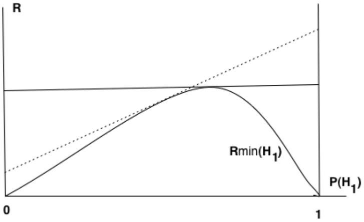
Cherchons le risque moyen, avec $c_{00} = c_{11} = 0$:

$$\mathcal{R} = \sum_{i=0}^{1} \sum_{j=0}^{1} c_{ji} P(H_i) \int_{d(x)=j} f(x|H_i) dx$$

$$= c_{01} P(H_1) \int_{d(x)=0} f(x|H_1) dx + c_{10} P(H_0) \int_{d(x)=1} f(x|H_0) dx$$
$$= c_{01} P(H_1) p_n + c_{10}(1 - P(H_1)) p_f$$
$$= c_{10} p_f + P(H_1) \left(c_{01} p_n - c_{10} p_f \right)$$

Test du Minimax (2)

Les coefficients c_{01} et c_{10} sont fixés (choisis). Par contre $P(H_1)$ est inconnu. La fonction de risque minimum en fonction de $P(H_1)$ est par contre connue. Nous allons donc essayer de **réduire le risque potentiel** le plus important au minimum :



Remarque : La pente de la droite est nulle lorsque

$$\boxed{(c_{01}p_n - c_{10}p_f) = 0}$$

Test de Newman-Pearson

Si les probabilités $P(H_0)$ et $P(H_1)$ sont inconnues, alors on peut définir un critère basé sur les **probabilités conditionnelles** de **non détection** et de **fausse alarme**.

Le critère de *Neyman-Pearson* consiste à ne considérer que les tests dont l'erreur de première espèce maximale se situe au-dessous d'un seuil fixer un seuil arbitraire $\alpha \in [0, 1]$. Les valeurs typiques d'α sont 0.005 (0.5%), 0.01 (1%), 0.05 (5%), suivant la "tolérance" de l'expérience à l'erreur. Une fois le seuil fixé c'est l'erreur de deuxième espèce qui devra être minimisée et la probabilité de détection ($d(x) = 0$ lorsque H_0 est vraie) maximisée (appelée également *puissance du test*).

Test de Newman-Pearson (2)

La règle de décision est basée sur le rapport de vraisemblance. Il s'agit de sélectionner un niveau λ tel que

$$\int_{x:\Lambda(x)>\lambda} f(x|H_0)dx = \alpha$$

La règle optimale (test uniformément le plus puissant) est de choisir

$$\begin{array}{lll} H_0 & \text{si} & x < \lambda \\ H_1 & \text{si} & x > \lambda \\ H_0 \text{ ou } H_1 & \text{si} & x = \lambda \end{array}$$

Exercice 3