INP ENSEEIHT TOULOUSE

TD n°2 : Statistique

Exercice 1

Considérons un problème de trafic, par exemple l'arrivée d'appels téléphoniques sur un faisceau de lignes d'un central téléphonique. On peut admettre que pour un faisceau particulier, le nombre d'arrivée d'appels pendant l'unité de temps suit une loi de Poisson de paramètre inconnu $\theta > 0$. On sait alors que sous cette hypothèse, la durée T séparant deux arrivées successives d'appels téléphoniques sur ce faisceau (avec l'unité de temps précédente) suit une loi exponentielle de paramètre θ soit :

$$f_T(t) = \theta e^{-\theta t} 1_{R^+}(t)$$

On désire estimer le paramètre θ inconnu à l'aide de l'observation de n durées $t_i, i = 1, ..., n$ séparant des arrivées successives d'appels téléphoniques sur ce faisceau.

1) dans un premier temps, on suppose qu'aucune information a priori n'est disponible sur θ (à part $\theta > 0$ bien sur). Déterminer à partir des observations t_i l'estimateur du maximum de vraisemblance de θ noté $\widehat{\theta}_{MV}$.

2) Il est connu que pour l'ensemble des faisceaux téléphoniques, le nombre moyen θ d'arrivées d'appels téléphoniques pendant l'unité de temps est distribuée suivant une loi exponentielle de paramètre λ connu :

$$g(\theta) = \lambda e^{-\lambda \theta} 1_{R^+}(\theta)$$

La densité $g(\theta)$ représente pour cet exemple la loi a priori sur le paramètre θ. Déterminer à partir des observations t_i les estimateurs de la moyenne a posteriori et du maximum a posteriori notés respectivement $\widehat{\theta}_{MMSE}$ et $\widehat{\theta}_{MAP}$.

3) Montrer que les estimateurs $\widehat{\theta}_{MV}$, $\widehat{\theta}_{MMSE}$ et $\widehat{\theta}_{MAP}$ sont équivalents lorsque n est grand et interpréter ce résultat.

Exercice 2

La même information binaire $\theta \in \{0, 1\}$ est transmise 2 fois consécutives vers un récepteur à travers un canal de transmission. Ces 2 informations sont perturbées par un bruit supposé Gaussien centré de variance σ^2. Le message reçu s'écrit alors $z = (z_1, z_2)$ où $z_i = \theta + e_i, i = 1, 2$ et $e_i \sim \mathcal{N}(0, \sigma^2)$. Le problème consiste à retrouver le symbole émis θ à partir du message reçu $z = (z_1, z_2)$.

1) Déterminer l'estimateur du maximum de vraisemblance du paramètre θ.

2) On suppose qu'on dispose d'une information a priori sur les bits "0" et "1" qui se traduit par $P(0) = P(1) = \frac{1}{2}$. Déterminer l'estimateur du maximum a Posteriori du paramètre θ. Représenter dans le plan (z_1, z_2) les points associés à la décision $\widehat{\theta}_{MAP} = 0$ et les points associés à la décision $\widehat{\theta}_{MAP} = 1$.

3) Comment les résultats de la question se modifient-ils lorsque $P(0) = p$ et $P(1) = q = 1 - p$?

4) Que pensez vous de l'estimateur de la moyenne a posteriori du paramètre θ ?

Exercice 3

Une source d'informations binaire émet des bits 1 et 0 avec les probabilités p et $1-p$. On désire estimer la probabilité $p = P(1)$ à l'aide de n observations émises par cette source notées $x_1, ...x_n$ avec $x_i \in \{0, 1\}$. Déterminer un intervalle de confiance du paramètre p avec un coefficient de confiance $\alpha = 0.95$ construit à partir de l'estimateur $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (on supposera que n est suffisamment grand pour pouvoir approcher la loi de $\widehat{p}$ par une loi normale en vertu du théorème de la limite centrale).

Exercice 1

1) La densité de $(T_1, ..., T_n)$ est

$$f(t_1, ..., t_n; \theta) = \theta^n \exp\left(-\theta \sum_{k=1}^{n} t_i\right)$$

Maximiser la vraisemblance de $t_1, ..., t_n$ revient à maximiser son logarithme qui s'écrit

$$\ln f(t_1, ..., t_n; \theta) = n \ln \theta - \theta \sum_{k=1}^{n} t_i$$

Mais

$$\frac{\partial \ln f(t_1, ..., t_n; \theta)}{\partial \theta} = 0 \Longrightarrow \frac{n}{\theta} - \sum_{k=1}^{n} t_i = 0$$

d'où

$$\widehat{\theta}_{\text{MV}} = \frac{n}{\sum_{k=1}^{n} T_i}$$

2) La loi a posteriori de θ s'écrit

$$\begin{aligned} f(\theta |\, t_1, ..., t_n) &\propto f(t_1, ..., t_n |\, \theta) f(\theta) \\ &\propto \theta^n \exp\left[-\theta \left(\lambda + \sum_{k=1}^{n} t_i\right)\right] \end{aligned}$$

qui est une loi gamma de paramètres $n+1$ et $\lambda + \sum_{k=1}^{n} t_i$, i.e.,

$$\theta |\, t_1, ..., t_n \sim \Gamma\left(n+1, \lambda + \sum_{k=1}^{n} t_i\right)$$

La moyenne de cette loi est

$$\widehat{\theta}_{\text{MMSE}} = E\left[\theta |\, T_1, ..., T_n\right] = \frac{n+1}{\lambda + \sum_{k=1}^{n} T_i}$$

tandis que son mode est

$$\widehat{\theta}_{\text{MAP}} = \frac{n}{\lambda + \sum_{k=1}^{n} T_i}$$

3) On remarque que

$$\widehat{\theta}_{\text{MMSE}} = \frac{n}{\sum_{k=1}^{n} T_i} \frac{1 + \frac{1}{n}}{1 + \frac{\lambda}{\sum_{k=1}^{n} T_i}}$$

et

$$\widehat{\theta}_{\text{MAP}} = \frac{n}{\sum_{k=1}^{n} T_i} \frac{1}{1 + \frac{\lambda}{\sum_{k=1}^{n} T_i}}$$

Quand $n \to \infty$, on sait que

$$\frac{1}{n} \sum_{k=1}^{n} T_i \to E\left[T_i\right] = \frac{1}{\theta} \text{ donc } \sum_{k=1}^{n} T_i \to \infty$$

Les estimateurs $\widehat{\theta}_{\mathrm{MV}}, \widehat{\theta}_{\mathrm{MMSE}}$ et $\widehat{\theta}_{\mathrm{MAP}}$ sont donc équivalents pour n "grand".

Exercice 2

1)

$$f(z_1, z_2; \theta) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{(z_1-\theta)^2 + (z_2-\theta)^2}{2\sigma^2}\right]$$

Donc

$$\widehat{\theta}_{\mathrm{MV}} = 0 \text{ si } f(z_1, z_2; 0) \geq f(z_1, z_2; 1)$$

c'est-à-dire

$$\widehat{\theta}_{\mathrm{MV}} = 0 \text{ si } z_1^2 + z_2^2 \leq (z_1-1)^2 + (z_2-1)^2$$

c'est-à-dire qu'on décide que le bit transmis est 0 si (z_1, z_2) est plus proche de $(0,0)$ que de $(1,1)$.

2)

$$f(\theta|\, z_1, z_2) \propto \exp\left[-\frac{(z_1-\theta)^2 + (z_2-\theta)^2}{2\sigma^2}\right]$$

Comme la loi a priori est uniforme, on a

$$\widehat{\theta}_{\mathrm{MAP}} = \widehat{\theta}_{\mathrm{MV}}$$

3)

$$f(\theta|\, z_1, z_2) \propto (1-p)^{\theta}\, p^{1-\theta} \exp\left[-\frac{(z_1-\theta)^2 + (z_2-\theta)^2}{2\sigma^2}\right]$$

donc

$$\widehat{\theta}_{\mathrm{MAP}} = 0 \text{ si } f(0|\, z_1, z_2) \geq f(1|\, z_1, z_2)$$

c'est-à-dire si

$$p \exp\left[-\frac{z_1^2 + z_2^2}{2\sigma^2}\right] \geq (1-p) \exp\left[-\frac{(z_1-1)^2 + (z_2-1)^2}{2\sigma^2}\right]$$

soit

$$\ln p - \frac{1}{2\sigma^2}\left(z_1^2 + z_2^2\right) \geq \ln(1-p) - \frac{(z_1-1)^2 + (z_2-1)^2}{2\sigma^2}$$

c'est-à-dire

$$z_1^2 + z_2^2 \leq (z_1-1)^2 + (z_2-1)^2 - 2\sigma^2 \ln\left(\frac{1-p}{p}\right)$$

Par exemple, si $p = 1/4$, on plus de chance d'avoir le bit 1 que le bit 0 et la région de décision est

$$\widehat{\theta}_{\mathrm{MAP}} = 0 \text{ si } z_1^2 + z_2^2 \leq (z_1-1)^2 + (z_2-1)^2 - 2\sigma^2 \ln 3$$

qui est plus réduite que celle obtenue à la question 2), ce qui est normal.

4) L'estimateur MMSE n'est pas à valeur dans $\{0,1\}$ donc il n'est pas adapté à ce problème.

Exercice 3

Pour n grand, en vertu du théorème de la limite centrale, on peut approcher la loi de

$$T = \frac{1}{n}\sum_{k=1}^{n} X_i$$

par une loi normale de moyenne

$$E[T] = \frac{1}{n}\sum_{k=1}^{n} E[X_i] = p$$

et de variance

$$\text{var}[T] = \frac{\text{var}[X_1]}{n} = \frac{p(1-p)}{n}$$

donc

$$\frac{T-p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0,1)$$

On en déduit avec les tables de la loi normale

$$P\left[\left|\frac{T-p}{\sqrt{\frac{p(1-p)}{n}}}\right| \geq 1.96\right] = 0.05$$

Donc on a

$$P\left[\left|\frac{T-p}{\sqrt{\frac{p(1-p)}{n}}}\right| \leq 1.96\right] = 0.95$$

L'intervalle de confiance est donc défini par

$$\left|\frac{T-p}{\sqrt{\frac{p(1-p)}{n}}}\right| \leq 1.96$$

Il suffit de résoudre cette inégalité en fonction de p pour voir l'intervalle de confiance désiré.