ECN 7060, cours 10

William McCausland

2019-11-12

Fonction de risque, risque de Bayes

- Pour une fonction de perte $L(\theta, a)$ donnée et un estimateur $\delta(X)$ donné, la fonction de risque (une fonction de θ) est, dans la notation de Casella et Berger :

$$R(\theta, \delta) = E_\theta[L(\theta, \delta(X))]$$

- L'espérance est par rapport à la loi de X pour θ donné.
- Pour un bayésien, θ est aléatoire et on peut écrire

$$R(\theta, \delta) = E[L(\theta, \delta(X))|\theta].$$

- Le risque de Bayes, pour une densité *a priori* $\pi(\theta)$ donnée, est

$$\begin{aligned} r(\pi, \delta) &\equiv \int \pi(\theta) E[L(\theta, \delta(X))|\theta]\, d\theta = E[E[L(\theta, \delta(X))|\theta]] \\ &= E[L(\theta, \delta(X))] \end{aligned}$$

- En même temps,

$$r(\pi, \delta) = E[E[L(\theta, \delta(X))|X]].$$

Règles (de décision) de Bayes

- Rappel : $r(\pi, \delta) = E[E[L(\theta, \delta(X))|X]]$.
- Une *règle de Bayes* est une fonction de décision δ^* qui minimise $r(\pi, \delta)$ pour π et $L(\theta, a)$ donné.
- Difficultés possibles
  - non-unicité de δ
  - absence d'une solution parce que $R(\theta, \delta) = \infty$ pour tous δ
- Même si $r(\pi, \delta)$ est toujours infini, on peut souvent trouver, pour x donné, $\delta(x)$ qui minimise la perte *a posteriori* espérée $E[L(\theta, \delta(X))|X]$ à $\{X = x\}$.
  - C'est une règle de Bayes *généralisée*.
  - En pratique, on le fait pour x observée seulement; $\delta(x)$ a souvent la même dimension que θ.
  - Pour la perte quadratique, $\delta(x)$ est la moyenne *a posteriori*.
  - Pour la perte valeur absolue, $\delta(x)$ est la médiane *a posteriori*.
  - Pour une autre perte, on peut approximer $\delta(x)$ par simulation.

Dominance et admissibilité

- La fonction de décision δ^* *domine* la fonction de décision δ par rapport à la fonction de perte $L(\theta, a)$ si $R(\theta, \delta^*) \leq R(\theta, \delta)$, avec une inégalité stricte pour au moins une valeur de θ.
- Une fonction de décision est admissible s'il n'y a pas d'autre fonction de décision qui la domine.
- Supposons que $\delta(x)$ minimise $r(\pi, \delta) = \int R(\theta, \delta)\pi(\theta)\, d\theta$, pour une fonction $\pi \colon \Theta \to \mathbb{R}_+$.
  - Si $\delta(x)$ est inadmissible, il existe une $\delta^*(x)$ qui la domine : il y a un ensemble $\bar{\Theta}$ où $R(\theta, \delta) > R(\theta, \delta)$. Il faut que $\pi(\bar{\Theta}) = 0$. Sinon, $\delta(x)$ ne minimise $r(\pi, \delta)$.
- À quelques conditions techniques près, un estimateur admissible est une règle de Bayes généralisée (avec possiblement une loi *a priori* impropre).

Biais, EMQ

- Notation, définitions
  - W est un estimateur de θ ou plus généralement de $\tau(\theta)$
  - Le biais de W est $E_\theta[W] - \theta$ ou $E_\theta[W] - \tau(\theta)$
  - L'espérance moyenne quadratique est $E_\theta[(W-\theta)(W-\theta)^\top] = \mathrm{Var}_\theta[W] + \mathrm{biais}_\theta[W]\mathrm{biais}_\theta[W]^\top$.
- L'importance du biais et l'EMQ est largement due à la solubilité des problèmes.
- Rappelons que la perte quadratique est seulement un choix parmi plusieurs. Quelques problèmes :
  - paramètres d'échelle toujours positifs,
  - impossibilité de la perte asymétrique,
  - non-existance de la moyenne ou la variance d'un estimateur.
- Le non-biais n'est pas un principe fiable, si on considère l'exemple suivant.

Un estimateur non-biaisé ridicule (RUBE)

- $X_i \sim \mathrm{Po}(\lambda)$, $n = 1$.
- On veut estimer $\tau(\lambda) = e^{-3\lambda}$
- Considérons la statistique $T(X) = (-2)^X$
- $E[T] = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(-2\lambda)^x}{x!} = e^{-3\lambda}$
- Par complétion de la famille de loi Poisson, T est l'estimateur unique non-biaisé de $\tau(\lambda)$.
  - Si $E_\theta[g(X)] = 0$ pour tous θ, $P(\{g(X) = 0\}) = 1$.
  - Soit $g(x) = T(x) - T'(x)$ la différence entre deux candidats pour un estimateur non-biaisé.
- Pour $x = 9, 10, 11$, $T(x) = -512, 1024, -2048$
- Pour $\lambda = 10$, $e^{-3\lambda} \approx 9.357623 \times 10^{-14}$.

Statistiques suffisantes dans un modèle gaussien

- Modèle : $X_i \sim \mathrm{iid}\, N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$
- Densité des données :

$$
\begin{aligned}
f(x|\theta) &= \sum_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right] \\
&= (2\pi)^{-n/2}\sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right] \\
&= (2\pi)^{-n/2}\sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}((n-1)S^2 + n(\bar{x}-\mu)^2)\right]
\end{aligned},
$$

où $S^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

- Une statistique suffisante minimale pour (μ, σ^2) : $(\bar{x}, S^2)$.

EMQ de $\hat{\sigma}^2$ et S^2 dans le modèle $X_i \sim \mathrm{iid}\, N(\mu, \sigma^2)$

- Rappel: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$.
- L'estimateur EMV de (μ, σ^2) est $(\hat{\mu}, \hat{\sigma}^2) = (\bar{x}, \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2)$.
- S^2 est non-biaisé ; $\hat{\sigma}^2$ est biaisé mais sa EMQ est moins grande, peu importe la valeur de σ^2. (exemples 7.3.3, 7.3.4)

La fonction de score

- Soit $L(\theta; x)$ une vraisemblance, $f(x|\theta)$ la densité des données.
- La fonction de score est le gradient :

$$V(\theta, x) = \frac{\partial \log L(\theta; x)}{\partial \theta^\top} = \frac{1}{L(\theta; x)} \frac{\partial L(\theta; x)}{\partial \theta}.$$

- Si on peut changer l'ordre de l'intégrale et la dérivée,

$$E\left[\frac{\partial \log L(\theta; x)}{\partial \theta^\top}\right] = \int \frac{f(x|\theta)}{f(x|\theta)} \frac{\partial f(x|\theta)}{\partial \theta^\top}\, dx = \frac{\partial \int f(x|\theta)\, dx}{\partial \theta^\top} = 0.$$

- Conditions suffisantes pour pouvoir changer l'ordre de l'intégrale et la dérivée
  1. La densité $f(x|\theta)$ a un support borné et ce support ne dépend pas de θ.
  2. La densité $f(x|\theta)$ a un support infini et est continument différentiable en θ ; l'intégral converge uniformement sur Θ.

Inégalité Cramér-Rao

- Échantillon $X_1, \ldots, X_n$, pas nécessairement iid, densité $f(x|\theta)$.
- Supposons que $E[V(\theta, X)] = 0$, $\mathrm{Var}_\theta[W(X)] < \infty$,
$$\frac{d}{d\theta} E_\theta[W(X)] = \int \frac{\partial}{\partial\theta}[W(x) f(x|\theta)]\, dx.$$
- Alors
$$\mathrm{Var}_\theta[W(X)] \geq \frac{\left(\frac{d}{d\theta} E_\theta[W(X)]\right)^2}{E_\theta\left[V(\theta, X)^2\right]}.$$
- Preuve I :
$$\frac{d}{d\theta} E_\theta[W(X)] = \int W(x) \left[\frac{\partial}{\partial\theta} f(x|\theta)\right] dx$$
$$= E_\theta\left[W(X) V(\theta, X)\right] = \mathrm{Cov}_\theta[W(X), V(\theta, X)]$$
$$\mathrm{Var}_\theta[V(\theta, X)] = E_\theta[V(\theta, X)^2]$$
- Preuve, II : le reste par l'inégalité de covariance $\mathrm{Var}_\theta[W(X)]\mathrm{Var}_\theta[V(\theta, X)] \geq \mathrm{Cov}_\theta[W(X), V(\theta, X)]^2$

Remarques, inégalité Cramér-Rao

- Le dénominateur est l'information Fisher, qui dépend du modèle et non l'estimateur (pourvu que c'est non-biaisé)
- Plus utile dans le cas où $W(X)$ est non-biaisé : $E_\theta[W(X)] = \theta$, numérateur $= 1$, la variance a une borne qui ne dépend pas de l'estimateur.
- Toujours une fonction de θ, par contre.

Théorème Rao-Blackwell

- Soit W un estimateur non-biaisé de $\tau(\theta)$, T une statistique suffisante pour θ. Alors $\phi(T) = E[W|T]$ est un estimateur de $\tau(\theta)$ qui est non-biaisé et uniformement meilleur en termes de variance.
- Preuve: $\phi(T)$ est une fonction de T et alors une fonction de l'échantillon seulement.
- Non-biais :

$$E_\theta[\phi(T)] = E_\theta[E[W|T]] = E_\theta[W] = \tau(\theta).$$

- Uniformement meilleur en termes de variance :

$$\begin{aligned}\mathrm{Var}_\theta[W] &= \mathrm{Var}_\theta[E[W|T]] + E_\theta[\mathrm{Var}[\mathrm{W}|\mathrm{T}]] \\ &= \mathrm{Var}_\theta[\phi(T)] + E_\theta[\mathrm{Var}[\mathrm{W}|\mathrm{T}]]\end{aligned}$$

Devoirs et lectures

Devoirs, Casella et Berger (matière du cours 10)

1. Exercise 7.10
2. Exercise 7.38
3. Exercise 7.40
4. Exercise 7.41

Préparation du cours 11, Casella et Berger

1. Chapitre 8 (pas tous les détails des exemples)
2. Questions suggérées : 8.6, 8.16