probabilité conditionnelle

degré de croyance

principe de l'inférence bayésienne

décision bayésiennes

exemples divers (a priori, a posteriori)

distribution gaussienne multidimensionnelle

Eléments de statistiques pour les data sciences

Cours 11: inférence bayésienne

Dr. Ph. Müllhaupt

IGM - EPFL

—

plan

1. probabilité conditionnelle
2. degré de croyance
3. principe de l'inférence bayésienne
4. décision bayésiennes
5. exemples divers (a priori, a posteriori)
6. distribution gaussienne multidimensionnelle

probabilité conditionnelle

rappel de la définition

Soit deux évènements A et B

définition

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)}$$

probabilité conditionnelle

interprétation fréquentiste

interprétation

Supposons un grand nombre d'expériences n a été effectué et les résultats des évènements A et B ont été enregistrés sous la forme suivante : n_A expériences parmi les n expériences correspondent à l'évènement A, et n_B expériences correspondent à l'évènement B. On note $n_{A\cap B}$ le nombre d'expériences qui correspondent aux évènements qui ont lieu simultanément A et B. L'interprétation fréquentiste de la probabilité suggère que si les expériences ont été réalisées de manière indépendante, la fréquence $\frac{n_A}{n}$ sera proche de la probabilité $P(A)$ et la fréquence $\frac{n_B}{n}$ sera proche de la probabilité $P(B)$. Maintenant,

$$\frac{P(A\cap B)}{P(B)} \approx \frac{\frac{n_{A\cap B}}{n}}{\frac{n_B}{n}} = \frac{n_{A\cap B}}{n_B} \approx P(A|B)$$

Ceci mesure l'expectative d'avoir l'évènement A se produire lorsque l'évènement B est observé.

distribution conditionnelle

distribution continue

définition

Soit X et Y deux variables aléatoires avec densitté jointe f, distribution marginale f_1 pour X et distribution marginale f_2 pour Y. La probabilité conditionnelle de Y sachant $X = x$ est définie par

$$f_2(y|x) = \frac{f(x,y)}{f_1(x)} \qquad -\infty < y < +\infty$$

distribution conditionnelle

distribution continue

Dans le cas continu il faut être prudent car $P(X = x) = 0$. Il faut prendre un intervalle et le faire tendre vers zéro.

$$P(A|X = x) = \lim_{h \to 0} P(A|0 \leq X \leq h)$$

distribution conditionnelle

distribution continue

Une autre façon de rendre la définition plus claire est de passer par les fonctions de répartition car on peut appliquer la définition sans autre

$$\begin{aligned} P(Y \leq y | x \leq X \leq x+h) &= \frac{P(x \leq X \leq x+h,\ Y \leq y)}{P(x \leq X \leq x+h)} \\ &= \frac{F(x+h, y) - F(x, y)}{F_1(x+h) - F_1(x)} \end{aligned}$$

(remarque $F_1(0)$ n'apparaît pas dans la formule précédente car $f(x) = 0$ pour $x < 0$), et il suffit alors de prendre la limite

$$\lim_{h \to 0} P(Y \leq y | x \leq X \leq x+h) = \int_{-\infty}^{y} \frac{f(x, t)}{f_1(x)} dx \triangleq \int_{-\infty}^{y} f_2(t|x) dt$$

Exemple

distribution continue

Soit

$$f(x,y) = e^{-(x+y)} \qquad 0 \leq x < +\infty \text{ et } 0 \leq y < +\infty$$

$$P(Y \leq 2, 0 \leq X \leq h) = \int_0^2 dy \int_0^h e^{-(x+y)} dx = (1 - e^{-2})(1 - e^{-h})$$

$$P(Y \leq 2, 0 \leq X \leq hY) = \int_0^2 dy \int_0^{hy} e^{-(x+y)} dx = \frac{h}{1+h} - \frac{1 + h - e^{-2h}}{e^2(1+h)}$$

$$P(0 \leq X \leq h) = \int_0^\infty dy \int_0^h e^{-(x+y)} dx = 1 - e^{-h}$$

$$P(0 \leq X \leq hY) = \int_0^\infty dy \int_0^h e^{-(x+y)} dx = \frac{h}{1+h}$$

Exemple

distribution continue

lorsque $h = 0$ les deux expressions précédentes sont différentes de zéro, et en utilisant $P(A|B) = p(A \cap B)\ P(A)$

$$p_1(h) \triangleq P(Y \leq 2 | 0 \leq X \leq h) = 1 - e^{-2}$$

$$p_2(h) \triangleq P(Y \leq 2 | 0 \leq X \ \leq hY) = 1 - \frac{1 + h - e^{-2h}}{e^2 h}$$

$$p_1 \triangleq \lim_{h \to 0} p_1(h) = 1 - e^{-2} = 0.865$$

$$p_2 \triangleq \lim_{h \to 0} p_2(h) = 1 - 3\, e^{-2} = 0.594$$

distribution a priori

rattaché au degré de croyance en un paramètre

- ensemble des échantillons $\mathcal{X}$
- échantillon x
- paramètre de la distribution θ
- notation pour la distribution : $f(x|\theta)$. Le paramètre θ est fixe mais inconnu.
- Soit H l'état des connaissances de l'expérimentateur lorsqu'il effectue l'expérience et collecte l'échantillon x.
- θ aura une distribution qui dépend de H au sens du degré de croyance

$$\boxed{\pi(\theta|H)}$$

Probabilité au sens du degré de croyance

- Plusieurs auteurs ont donné un cadre théorique solide pour les proabilités au sens du degré de croyance (Savage, Ramsey, Jeffreys, de Finetti, Carnap, Lindley, pour en citer que quelques-un)
- Il est possible de donner un systèmes d'axiomes qui rend possible l'inférence en utilisant une logique inductive, par rapport à la physique et mathématique classique qui utilise un système d'axiomes pour une logique déductive.
- Nous allons utiliser les théorèmes de Bayes.

Axiomes des probabilités

associés au degré de croyance

Axiomes :

1. $0 \leq \pi(A|B) \leq 1$ et $\pi(A|A) = 1$
2. Si les évènements $\{A_i\}$ sont exclusifs sachant B (loi d'addition des probabilités)

$$\pi\left(\cup A_i|B\right) = \sum_i \pi(A_i|B)$$

3. $\pi(C|A \cap B)\pi(A|B) = \pi(A \cap C|B)$

indépendance

définition

définition de l'indépendance (loi de multiplication des probabilités)

Deux évènements sont indépendants si connaissant C on a

$$\pi(A \cap B|C) = \pi(A|C)\pi(B|C)$$

quelques théorèmes de Bayes...

Thm 1.

$$\pi(A|B) = \pi(A \cap B|B)$$

démonstration :

Le troisième axiome donne en remplaçant A par B, B par C et C par A

$$\pi(A|B \cap C)\pi(B|C) = \pi(B \cap A|C)$$

Ensuite on remplace C par B

$$\pi(A|B \cap B)\pi(B|B) = \pi(A \cap B|B)$$

et le membre de gauche égale $\pi(A|B)$ car $B \cap B = B$ et $\pi(B|B) = 1$ par l'axiome 1.

Thm 2.

$$\text{Si } A \Rightarrow B \text{ alors}$$

$$\pi(A|B)\pi(B) = \pi(A)$$

démonstration :

L'axiome 3 peut s'écrire $\pi(A|B \cap C)\pi(B|C) = \pi(B \cap A|C)$ et en posant $C = B$, on arrive à $\pi(A|B \cap B)\pi(B|B) = \pi(B \cap A|B)$ et comme $A \Rightarrow B$, on a $A \subset B$ et donc $B \cap A = A$ et on a $\pi(B \cap A|B) = \pi(A)$.

Thm 3. (thm. de Bayes)

soit $\{A_i\}$ une séquence d'évènements et B un évènement quelconque avec $\pi(B) \neq 0$ alors

$$\pi(A_n|B) \propto \pi(B|A_n)p(A_n)$$

démonstration :

L'axiome 3 donne $\pi(C|A)\pi(A) = \pi(A \cap C)$ et $\pi(A|C)\pi(C) = \pi(C \cap A)$ et en combinant on a

$$\pi(C|A) = \pi(A|C)\pi(C)/\pi(A)$$

principe de l'inférence bayésienne

a priori, a posteriori

- $\mathrm{x} = (x_1, \ldots, x_n)$ un échantillon de variables aléatoires indépendantes posons

$$p(\mathrm{x}|\theta, H) = \prod_{i=1}^{n} f(x_i|\theta, H)$$

- utilisation du théorème de Bayes

$$\pi(\theta|\mathrm{x}, H) \propto p(\mathrm{x}|\theta, H)\, \pi(\theta|H)$$

La constante de proportionalité

$$\left\{ \int p(\mathrm{x}|\theta, H)\pi(\theta|H)d\theta \right\}^{-1} \triangleq \frac{1}{\pi(\mathrm{x}|H)}$$

ne dépend pas des paramètres.

- on omettra H par la suite sauf si c'est important

- Vocabulaire : distribution a priori

$$\pi(\theta|H)$$

- Vocabulaire : distribution a posteriori

$$\pi(\theta|\mathrm{x}, H)$$

.

- Vocabulaire : vraisemblance

$$p(\mathrm{x}|\theta, H)$$

vue comme une fonction de θ

hypothèse et décision bayésienne

fonction de coût

- $C_{H_0\,H_0} = C_{00}$: coût d'accepter H_0 alors que H_0 est vraie.
- $C_{H_1\,H_0} = C_{10}$: coût d'accepter H_1 alors que H_0 est vraie.
- $C_{H_1\,H_1} = C_{11}$: coût d'accepter H_1 alors que H_1 est vraie.
- $C_{H_0\,H_1} = C_{01}$: coût d'accepter H_0 alors que H_1 est vraie.

décision bayésienne

coût attendu

l'espérance mathématique du coût entraîne le coût attendu (expected cost)

$$\mathscr{R} \triangleq \mathbb{E}\{ \text{ coût } \}$$

$$\mathscr{R} = \mathbb{E}\{ \text{ coût } | H_0\} \, \pi(H_0) + \mathbb{E}\{ \text{ coût } | H_1\} \, \pi(H_1)$$

régions d'acceptation

- $\mathcal{A}_0$: région d'acceptation de H_0
- $\mathcal{A}_1$: région d'accepation de H_1

définition de la décision bayésienne

Il faut sélectionner $\mathcal{A}_0$ et $\mathcal{A}_1$ de telle sorte à minimiser $\mathscr{R}$.

décision bayésienne

On décide H_0 lorsque

$$\lambda(x) = \frac{f(x|H_0)}{f(x|H_1)} > \frac{\pi(H_1)(C_{01} - C_{11})}{\pi(H_0)(C_{10} - C_{00})} \triangleq k$$

et on décide H_1 lorsque $\lambda(x) < k$

Exemple

illustratif de la décision bayésienne

- $$H_0 : f(x|H_0) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$
- $$H_1 : f(x|H_1) = \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{1}{8}x^2\right)$$
- On suppose a priori que les deux hypothèses sont d'égale probabilité $P\{H_0\} = P\{H_1\} = 0.5$.
- $C_{01} = C_{10} = 1$ et $C_{00} = C_{11} = 0$

le rapport de vraisemblance

$$\lambda(x) = \frac{1}{2} \exp\left(\frac{1}{8}x^2 - \frac{1}{2}x^2\right) = \frac{1}{2} \exp\left(-\frac{3}{8}\right)$$

Le niveau

$$k = \frac{\frac{1}{2}(1-0)}{\frac{1}{2}(1-0)} = 1$$

Si $1/2 \exp(-3/8x^2) < 1$ on privilégie H_1

théorème

variable aléatoire gaussienne avec a priori gaussien

Thm.

- $X \sim \mathcal{N}(\theta, \sigma^2)$ où σ^2 est connu
- moyenne θ inconnue mais supposée distribuée a priori $\pi(\theta) \sim \mathcal{N}(\mu_0, \sigma_0^2)$.

La distribution a posteriori sera alors $\pi(\theta|x) \sim \mathcal{N}(\mu_1, \sigma_1^2)$ avec

$$\mu_1 = \frac{x/\sigma^2 + \mu_0/\sigma_0^2}{1/\sigma^2 + 1/\sigma_0^2}, \qquad \sigma_1^{-2} = \sigma^{-2} + \sigma_0^{-2}$$

commentaire

- La distribution initiale influence le résultat par rapport à l'échantillon
- La moyenne est une pondération entre l'échantillon x et la valeur a priori de la moyenne μ_0.
- La variance de la distribution a posteriori et la moyenne harmonique de la variance apriori σ_0 et de celle de la variance de la distributin de l'échantillon σ^2

$$\sigma_1^2 = \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

- L'échantillon x remet à jour notre croyance initiale en θ donné sous la forme de $\pi(\theta)$ pour constituer un nouveau degré de croyance en θ donné sous la forme de la distribution a posteriori $\pi(\theta|x)$ pour le paramètre θ.

démonstration

la vraisemblance

$$p(x|\theta) \propto \exp[-(x-\theta)^2/(2\sigma^2)]$$

a priori

$$\pi(\theta) \propto \exp[-(\theta-\mu_0^2)/(2\sigma_0^2)]$$

a posteriori

$$\begin{aligned}
\pi(\theta|x) &\propto \exp\left\{-\frac{(x-\theta)^2}{2\sigma^2} - \frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right\} \\
&\propto \exp\left\{-\frac{1}{2}\theta^2(1/\sigma^2 + 1/\sigma_0^2) + \theta(x/\sigma^2 + \mu_0/\sigma_0^2)\right\} \\
&= \exp\left\{-\frac{1}{2}\theta^2/\sigma_1^2 + \theta\mu_1/\sigma_1^2\right\} \\
&\propto \exp\{-1/2(\theta-\mu_1)^2/\sigma_1^2\} \qquad \text{CQFD}
\end{aligned}$$

corollaire

corollaire

Soit $\mathrm{x} = (x_1, x_2, \ldots, x_n)$ un échantillon de n variables aléatoires indépendantes $X_i \sim \mathscr{N}(\theta, \sigma^2)$ où σ^2 est connu et avec la distribution a priori $\pi(\theta) \sim \mathscr{N}(\mu_0, \sigma_0^2)$. La distribution a posteriori est alors $\pi(\theta|\mathrm{x}) \sim \mathscr{N}(\mu_n, \sigma_n^2)$ avec

$$\mu_n = \frac{n\,\bar{x}/\sigma^2 + \mu_0/\sigma_0^2}{n/\sigma^2 + 1/\sigma_0^2/}, \qquad \sigma_n^{-2} = n\,\sigma^{-2} + \sigma_0^{-2}$$

et $\bar{x} = 1/n \sum x_i$ la moyenne d'échantillon.

démonstration

$$
\begin{aligned}
p(\mathrm{x}|\theta) &\propto \exp\left[-\sum_{i=1}^{n}(x_i-\theta)^2/(2\sigma^2)\right] \\
&\propto \exp[-1/2\,\theta^2(n/\sigma^2)+\theta\bar{x}(n/\sigma^2)] \\
&\propto \exp[-1/2\,(\bar{x}-\theta)^2(n/\sigma^2)] \qquad \text{CQFD}
\end{aligned}
$$

Encore un théorème...

gaussienne avec moyenne connue, mais variance inconnue

théorème

Soit $\mathrm{x} = (x_1, x_2, \ldots, x_n)$ un échantillon de n variables aléatoires indépendantes $X_i \sim \mathcal{N}(\mu, \theta)$, avec la moyenne connue mais la variance $\theta = \sigma^2$ inconnue. On considère l'apriori

$$\nu_0 \sigma_0^2 / \theta \sim \chi^2_{(\nu_0)}$$

alors la distribution a posteriori de

$$(\nu_0 \sigma_0^2 + s^2)/\theta \sim \chi^2_{(\nu_0 + n)}$$

avec la variable

$$s \triangleq \sum_{i=1}^{n} (x_i - \mu)^2$$

lemme utile

lors d'intégrale de fonctions gamma

lemme

$$\int_0^\infty e^{-A/\theta}\,\theta^{-m}\,d\theta = (m-2)!/A^{m-1} \qquad (A > 0, m > 1)$$

La démonstration peut être obtenue par intégration par partie itérée après la substitution $x = A/\theta$ et $dx = -Ad\theta/\theta^2$.

démonstration

Rappel de la distribution du $\chi^2_{(2m)}$

$$\exp^{-1/2x} x^{m-1}/(2^m(m-1)!)$$

Si $X = \nu_0\sigma_0/\theta \sim \chi_{(\nu_0=2m)}$ alors

$$\begin{aligned}
\pi(\theta) &= \exp\left\{-\frac{\nu_0\sigma_0^2}{2\theta}\right\}\left(\left(\frac{\nu_0\sigma_0^2}{\theta}\right)^{\frac{1}{2}\nu_0-1}\frac{\nu_0\sigma_0^2}{\theta^2}/[2^{\frac{1}{2}\nu_0}(\frac{1}{2}\nu_0-1)!]\right) \\
&\propto \exp\left\{-\frac{\nu_0\sigma_0^2}{2\theta}\right\}\theta^{-\frac{1}{2}\nu_0-1}
\end{aligned}$$

car $dx = -\nu_0\sigma_0^2 d\theta/\theta^2$.

La vraisemblance de l'échantillon est

$$p(\mathrm{x}|\theta) \propto \exp\left\{-\sum_{i=1}^{n}(x_1-\mu)^2/(2\theta)\right\} \propto e^{-s^2/2\theta}\,\theta^{-1/2n}$$

Distribution a posteriori

$$\pi(\theta|\mathrm{x}) \propto e^{-(\nu_0\sigma_0^2+s^2)/(2\theta)}\,\theta^{-1/2(n+\nu_0)-1} \qquad \mathit{CQFD}$$

deux échantillons gaussiens

$X_{1i} \sim \mathcal{N}(\theta_1, \sigma_1^2)$ et $X_{2j} \sim \mathcal{N}(\theta_2, \sigma_2^2)$

théorème

Si

- $\mathrm{x}_1 = (x_{11}, x_{12}, \ldots, x_{1n_1})$ un échantillon avec $X_{1i} \sim \mathcal{N}(\theta_1, \sigma_1^2)$
- $\mathrm{x}_2 = (x_{21}, x_{22}, \ldots, x_{2n_2})$ un échantillon avec $X_{2j} \sim \mathcal{N}(\theta_2, \sigma_2^2)$.
- la distribution a priori pour θ_1 et θ_2 est uniforme sur $]-\infty; +\infty[$

alors la distribution a posteriori de $\delta \triangleq \theta_1 - \theta_2$ est donnée par

$$\delta \sim \mathcal{N}\left(\bar{x}_1 - \bar{x}_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

deux distributions gaussiennes

$X_{1i} \sim \mathcal{N}(\theta_1, \phi)$ et $X_{2j} \sim \mathcal{N}(\theta_2, \phi)$

- $X_{1i} \sim \mathcal{N}(\theta_1, \phi),\ i = 1, \ldots, n_1$
- $X_{2j} \sim \mathcal{N}(\theta_2, \phi),\ j = 1, \ldots, n_2$
- distributions a priori pour θ_1 et θ_2 uniforme sur $]-\infty; +\infty[$
- distribution uniforme pour $\log \phi$ sur $]-\infty; +\infty[$

alors

$$\frac{\nu s^2}{\phi} \sim \chi^2_{(\nu)}$$

avec

$$\begin{aligned} \nu_i s_i^2 &= \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2, \qquad \nu_1 = n_i - 1 \\ \nu s^2 &= \nu_1^2 s_1^2 + \nu_2^2 s_2^2, \qquad \nu = \nu_1 + \nu_2 \end{aligned}$$

esquisse de la démontration

- en multipliant les vraisemblances par les a priori, cela conduit à la distribution a posteriori
- $$\pi(\theta_1, \theta_2, x_1, x_2) \propto \phi^{-1/2(n_1+n_2+2)} \exp[-\{n_1(\bar{x}_1 - \theta_1)^2 + n_2(\bar{x}_1 - \theta_2)^2 + \nu_1 s_1^2 + \nu_2 s_2^2\}/(2\phi)]$$
- Pour obtenir la distribution a posteriori de ϕ il faut intégrer l'expression précédente par rapport à θ_1 et θ_2 (distribution mariginale) et cela donne

$$\pi(\phi|x_1, x_2) \propto e^{-\nu s^2/(2\phi)}\, \phi^{-1/2\nu-1}$$

et on découvre le $\chi^2_{(\nu_1+\nu_2)}$

distribution normale multidimensionnelle

$$\boxed{p(x) = \frac{1}{(2\pi)^{n/2}|C|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)\right)}$$

- vecteur de moyenne

$$\mu = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix} = \mathbb{E}[X]$$

- matrice de covariance (matrice semi–définie positive), appelée également parfois matrice des variances et matrice de covariance

$$C = \mathbb{E}[(X - \mathbb{E}[X])\,(X - \mathbb{E}[X])^T]$$

- en composantes $C = [\sigma_{ij}]$ avec $\sigma_{ij} = \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)]$

C est notée également Σ ou Γ et K_{XX} selon les auteurs et les sources.

distribution normale multidimensionnelle

Thm.

- $x \sim \mathcal{N}(A\theta, C)$
- $\pi(\theta) \sim \mathcal{N}(\mu_0, C_0)$

sous ces conditions la matrice A est connue (i.e. telle que $\mathbb{E}(x) = A\theta$), le vecteur de moyennes μ_0 est oconnu et la matrice de covariance C_0 est connue mais les vecteur des paramètres θ n'est pas connu, la distribution a posteriori s'écrit

$$\begin{aligned}
\pi(\theta|x) &= \mathcal{N}(\mu_1, \theta_1) \\
\mu_1 &\triangleq (C_0^{-1} + A^T C^{-1} A)^{-1} (C_0^{-1} \mu_0 + A^T C^{-1} x) \\
C_1 &\triangleq (C_0^{-1} + A^T C^{-1} A)^{-1}
\end{aligned}$$

esquisse de démonstration

$$
\begin{aligned}
p(\mathrm{x}|\theta) &\propto \exp\left(-\frac{1}{2}(\mathrm{x}-A\,\theta)^T\,C^{-1}\,(x-A\,\theta)\right)\\
\pi(\theta) &\propto \exp\left(-\frac{1}{2}(\theta-\mu_0)^T\,C_0^{-1}\,(\theta-\mu_0)\right)
\end{aligned}
$$

Alors $\pi(\theta|\mathrm{x}) \propto p(\mathrm{x}|\theta)\,\pi(\theta)$. En éliminant les termes constants multiplicatifs on arrive après quelques manipulations

$$
\begin{aligned}
\pi(\theta|x) &\propto \exp\left(-\frac{1}{2}[\theta^T(\mathrm{C_0}^{-1}+A^TC^{-1}A)\,\theta-2\theta^T(C_0^{-1}\mu_0+A^TC^{-1}\mathrm{x})]\right)\\
&\propto \exp\left(-\frac{1}{2}[(\theta-\mu_1)^T\,(C_0^{-1}+A^TC^{-1}A)\,(\theta-\mu_1)\right)
\end{aligned}
$$

distance de Mahalanobis

r est la distance de x à μ

$$\boxed{r^2 = (x-\mu)^T\, C^{-1}\, (x-\mu)}$$