Prior information

Anna Drewek und Marc Lickes
Seminar über Bayes Statistik F08

12. März 2008

1 Einleitung

Definition 1.1. In der schliessenden Statistik ist ein Bayes'sches statistisches Modell ein parametrisches Modell $f(x|\theta)$ mit einer Wahrscheinlichkeitsverteilung auf dem Parameterraum Θ. Diese Wahrscheinlichkeitsverteilung nennen wir *a priori Verteilung*. Wir nehmen an, dass die a priori Verteilung absolut stetig ist mit Dichte $\pi(\theta)$.

Die Methode zur Festlegung einer a priori Verteilung ist stark abhängig von der zur Verfügung stehenden Information über den Prior:

1. Es sind genügend Informationen bekannt, so dass man die a priori Verteilung einfach bestimmen kann. Ein Beispiel hierfür ist die Parametermethode.

2. Die Informationsmenge ist nicht ausreichend, da zum Beispiel die Theorie über ein Experiment noch nicht vollständig entwickelt ist oder die Zeit fehlte, sich genügend Informationen zu beschaffen. Hier bieten konjugierte Prior Abhilfe.

3. Falls keine Information vorhanden sind, kann man dank Noninformative Priors Bayes'sche Statistik betreiben.

2 Laplace Experiment

Laplace war historisch gesehen der Erste (1773), der eine Priormethode zur Lösung eines Problems benutzte:

In einer Urne befinden sich N schwarze und weisse Kugeln K. Über die Anzahl der Kugeln der einzelnen Farben gibt es keine Information. Es wird eine weisse (w) Kugel gezogen. Laplace fragte sich, ob man daraus die Wahrscheinlichkeit, dass die Anzahl p der weissen Kugeln gleich p_0 ist, berechnen könne.
Er nahm an, dass die möglichen Verhältnisse q $\{2, 3, \ldots, (N-1)\}$ gleichverteilt sind (Uniforme a priori Verteilung). Dies ist sehr intuitiv, denn jedes Verhältnis könnte mit gleicher Wahrscheinlichkeit eintreten. Aus der Bayesformel erhält man die zugehörige a posteriori Verteilung:

$$P(p = p_0 \mid K = w) = \frac{P(p = p_0)\, P(K = w \mid p = p_0)}{\sum\limits_q P(p = q)\, P(K = w \mid p = q)} = \frac{\frac{1}{N-2}\,\frac{p_0}{N}}{\sum\limits_q \frac{1}{N-2}\,\frac{q}{N}} = \frac{p_0}{(N(N-1)/2) - 1}$$

Die Antwort auf Laplace Frage ist also $\frac{p_0}{(N(N-1)/2)-1}$.

3 Subjektive Festlegung

Bei der subjektiven Festlegung geht es darum, wie man mit Hilfe subjektiver Einschätzungen die a priori Verteilung festlegen kann. Wir gehen davon aus, dass ausser Daten noch weitere Information vorhanden ist. Der Prior wird als Werkzeug verwendet, um alle verfügbare Information zusammenzufassen. Die Information stammt hauptsächlich von Experten oder aus Erfahrungswerten, wobei letztere nicht zwingend existieren müssen. (Beispiel: Atomkrieg)

Beispiel 3.1 Ein Chemiker führt ein Experiment durch. Er schüttet irgendwelche Substanzen zusammen und misst hinterher den Säuregehalt. Den ungefähren Messwert kann er meistens schon vor der Messung sagen, da er sich zuvor mit der Theorie beschäftigt oder sich die Ergebnisse älterer Experimente angeschaut hat.

Ein Nachteil der subjektiven Festlegung ist, dass der Mensch Wahrscheinlichkeiten sehr unterschiedlich interpretiert:

Beispiel 3.2 Eine Studie im New England Journal of Medicine zeigte, dass 44% der befragten Individuen mit einer Behandlung gegen Lungenkrebs einverstanden waren, als ihnen erzählt wurde, dass die Überlebenswahrschein-

lichkeit 68 % betragen würde. Jedoch willigten nur 18% eine Behandlung ein, als ihnen gesagt wurde, dass die Wahrscheinlichkeit des Todes 32% ist.

3.1 Parametrische Approximation

Die subjektive Einschätzung ergibt gewisse charakteristische Werte, die die Wahl von einer a priori Verteilung einschränkt. Meistens sind Momente oder Quantile gegeben. Da viele in der Praxis beutzte Verteilungen schon durch zwei Parameter eindeutig bestimmt sind, kann mit Hife der Priorinformation eine a priori Verteilung festlegt werden. Diese Methode wird *Parametrische Approximation* genannt.

Beispiel

- Für die Normalverteilung reicht zum Beispiel der Median und das 75% Quantil oder der Erwartungswert und die Standardabweichung.
- Sei $X \sim B(n,p)$ die Anzahl der Studenten von insgesamt n Studenten, die einen Anfängermathematikkurs bestehen. Im vorherigen Jahr lag der arithmetische Durchschnitt für p bei 0.7 mit einer Standardabweichung von 0.1. Wenn wir annehmen die p seien alle mit derselben Betaverteilung $Be(\alpha,\beta)$ erzeugt, dann lassen sich die Parameter α und β wie folgt schätzen:

$$E[p] = \frac{\alpha}{\alpha+\beta} = 0.7 \qquad Var(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = 0.1$$
$$\Rightarrow \alpha = 0.77 \text{ und } \beta = 0.33$$

4 Konjugierte Prior

Konjugierte Prior werden meist verwendet, falls nur beschränkt Information vorhanden ist.

Definition 4.1. Eine Familie $\mathcal{F}$ von Wahrscheinlichkeitsverteilungen auf dem Parameterraum Θ heisst *konjugiert* für eine Likelihoodfunktion $f(x|\theta)$, falls $\forall \pi \in \mathcal{F}$ die a posteriori Verteilung $\pi(\theta \mid x)$ auch zu $\mathcal{F}$ gehört.

Beispiele für $\mathcal{F}$

- Menge $\mathcal{F}_0$ aller möglichen Verteilungen auf Θ.
- Die exponentielle Familie (vgl. Def 4.2.)

Bei der konjugierten Prior Methode sucht man nach möglichst kleinen und parametrisierten Familien $\mathcal{F}$. Die Parametrisierung bietet den Vorteil, den Posterior einfach durch Updaten der Parameter des Priors zu erhalten.

Mit den konjugierten Priorn werden hauptsächtlich die exponentielle Familien assoziiert.

4.1 Exponentielle Familie

Definition 4.2. Sei μ ein σ-endliches Mass auf dem Raum der Beobachtungen $\mathcal{X}$ und Θ der Parameterraum. $C : \Theta \to \mathbb{R}_+$, $h : \mathcal{X} \to \mathbb{R}_+$, $R : \Theta \to \mathbb{R}^k$ und $T : \mathcal{X} \to \mathbb{R}^k$. Die Familie von Verteilungen mit Dichten

$$f(x \mid \theta) = C(\theta)h(x)\exp(R(\theta)T(x))$$

heisst *exponentielle Familie* der Dimension k.

Bemerkung Die exponentielle Familie lässt sich mittels Umparametrisieren auch auf folgende zwei Art schreiben:

1. Kanonische Form: $z := T(x)$ und $\eta := R(\theta)$
   $\Rightarrow f(z \mid \eta) = C^{'}(\eta)h^{'}(z)\exp(\eta z)$
2. Gewöhnliche Form: $\vartheta := \eta$ und $e^{-\psi(\eta)} := C^{'}(\eta)$
   $\Rightarrow f(z \mid \vartheta) = h^{'}(z)\exp(\vartheta z - \psi(\vartheta))$

Beispiele

1. Sei $x \sim Poisson(\theta)$. Dann ist $f(x \mid \theta) = e^{-\theta}\frac{\theta^x}{x!} = \frac{1}{x!}\exp(x\log(\theta) - \theta)$.
   Also $T(x) = x$, $R(\theta) = \log(\theta)$, $C(\theta) = \exp(-\theta)$ und $h(x) = \frac{1}{x!}$

2. Sei $x \sim \mathcal{N}(\mu, \sigma^2)$.
   Dann ist $f(x|\theta) = \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}\frac{x^2}{\sigma^2} + \frac{\mu}{\sigma^2}x - (\frac{1}{2}\frac{\mu^2}{\sigma^2} + \frac{1}{2}\log\sigma^2)]$
   Hier ist $R(\theta) = (\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2})$ und $T(x) = (x^2, x)$, wobei $\theta = (\mu, \sigma^2)$

4.2 Exponentielle Familie und Konjugierte Prior

Proposition 4.1. Sei $f(x|\theta) = h(x)\exp(\theta x - \psi(\theta))$ eine gewöhnliche exponentielle Familie. Dann ist eine konjugierte Familie für $f(x|\theta)$ gegeben durch

$$\pi(\theta|\mu, \lambda) = K(\mu, \lambda)\exp(\theta\mu - \lambda\psi(\theta)) \tag{1}$$

wobei $K(\mu, \lambda)$ eine Normalisierungskonstante der Dichte ist. Die zugehörige a posteriori Verteilung ist $\pi(\theta \mid \mu + x, \lambda + 1)$.

Beweis: Aus der Bayes'schen Formel wissen wir: $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$.
$\pi(\theta|\mu, \lambda)f(x|\theta) = K(\mu, \lambda)\exp(\theta\mu - \lambda\psi(\theta))h(x)\exp(\theta x - \psi(\theta))$
$\propto K(\mu + x, \lambda + 1)\exp(\theta(\mu + x) - \psi(\lambda + 1)) = \pi(\theta|\mu + x, \lambda + 1)$
Die zugehörige a posteriori Verteilung liegt offensichtlich in derselben Familie wie die a priori Verteilung. □

Bemerkungen

- Die a priori Verteilung in (1) wird oft als *natürliche konjugierte Verteilung* bezeichnet.
- Genau dann, wenn $\lambda > 0$ und $\frac{\mu}{\lambda} \in N^\circ$, wobei $N = \{\theta; \int_{\mathcal{X}} e^{\theta x}h(x)d\mu(x) < \infty\}$ der *natürliche Parameterraum*, ist die a priori Verteilung in (1) ein Wahrscheinlichkeitsmass. Nur in diesem Fall ist auch der Normalisierungsfaktor $K(\mu, \lambda)$ wohldefiniert.
- Mit Hilfe dieser Proposition kann für gegebenen Likelihood $f(x \mid \theta)$ die natürliche konjugierte Verteilung berechnet werden.

4.3 Gemischte natürliche a priori Verteilungen

Lemma 4.2. Sei $\mathcal{F}$ eine natürliche konjugierte Familie einer exponentiellen Familie. Dann ist die diskrete Mischung von N konjugierten Verteilungen $\pi(\theta \mid \lambda_i, \mu_i)$, also $\tilde{\mathcal{F}}_N = \{\sum_{i=1}^{N} \omega_i \pi(\theta \mid \lambda_i, \mu_i); \sum_{i=1}^{N} \omega_i = 1, \omega_i > 0\}$ auch eine konjugierte Familie.

Beweis: Mit Hilfe von Proposition 4.1 erhält man eine a posteriori Verteilung $\pi(\theta \mid x) = \sum_{i=1}^{N} \omega_i'(x) \pi(\theta \mid \lambda_i + 1, \mu_i + x)$. □

Im Folgenden soll gezeigt werden, dass sich durch Mischen von natürlichen a priori Verteilungen jede a priori Verteilung beliebig genau approximieren lässt.

Definition 4.3. Die Prohorov Distanz zwischen zwei Wahrscheinlichkeitsmassen ist definiert als $d^p(\pi, \tilde{\pi}) = \inf\{\epsilon > 0; \forall \text{A Borel gilt } \pi(A) \leq \tilde{\pi}(A^\epsilon) + \epsilon\}$, wobei A^ϵ die Menge aller Punkte ist, die maximale Distanz ϵ von A haben.

Bemerkung Die Prohorov Distanz induziert die Topologie der schwachen Konvergenz.

Theorem 4.3. Wenn Θ ein natürlicher Parameterraum für die exponentielle Familie $f(x|\theta)$ ist und π eine a priori Verteilung auf Θ, dann existiert $\forall \epsilon > 0$ ein N und ein $\tilde{\pi} \in \tilde{\mathcal{F}}_N$, so dass $d^p(\pi, \tilde{\pi}) < \epsilon$

4.4 Kritik an der konjugierten Priormethode

Der Einfluss der Priorwahl auf den Posterior kann relativ gross sein für kleine Stichproben.

Beispiel Wir versetzen eine Münze in eine Kreiselbewegung auf ihrem Rand. Naiverweiser könnte man annehmen, dass die Wahrscheinlichkeit für Kopf oder Zahl sei 1/2. Beachtet man aber die Unebenheit des Randes, dann

könnte die Wahrscheinlichkeit für Kopf genauso gut 1/3 oder 2/3 sein. Deshalb wären die folgenden drei a priori Verteilungen plausibel:

π_1 $\quad Be(1,1)$

π_2 $\quad \frac{1}{2}[Be(10,20) + Be(20,10)]$

π_3 $\quad 0.5Be(10,20) + 0.2Be(15,15) + 0.3Be(20,10))$

Man lässt die Münze 10mal kreiseln und beobachtet 3 Mal Kopf. Daraus lassen sich die entsprechenden a posteriori Verteilungen berechnen:

$\pi_1(\theta \mid x)$ $\quad Be(1+x, 1+n-x) = Be(4,8)$

$\pi_2(\theta \mid x)$ $\quad 0.84Be(13,27) + 0.16Be(23,17)]$

$\pi_3(\theta \mid x)$ $\quad 0.77Be(13,27) + 0.16Be(18,22) + 0.07Be(23,17))$

Die a posteriori Verteilungen sind verschieden.

Bemerkung Lässt man die Stichprobenzahl gegen unendlich streben, so führen die meisten Prior zu denselben a posteriori Verteilungen.

5 Noninformative Prior

Noninformative Prior werden benützt, falls keine Information vorhanden ist. Sie liefern bessere Ergebnisse für die a posteriori Verteilung als klassische Methoden, wie zum Beispiel der Maximum Likelihood oder eine Approximation der Hyperparameter mithilfe der Daten. Die einzige Informationsquelle ist die Stichprobenverteilung.

5.1 Laplace Prior

Wie im einführenden Beispiel gesehen, benutzte Laplace eine a priori Verteilung, die allen Parametern die gleiche Wahrscheinlichkeit zuordnet, den *Uniform Prior*.

Probleme und Kritik

- Falls der Parameterraum Θ nicht kompakt ist, bekommt man improper a priori Verteilungen. Mit diesen improper Prior ist es im Allgemeinen schwierig, wenn auch nicht unmöglich zu arbeiten.

- Eine Aufteilung des Parameterraums Θ in verschiedene Partitionen kann zu Unstimmigkeiten führen. Beispielsweise könnte Θ von der Form $\{\theta_1, \theta_2\}$ sein, dann wäre nach Laplace $\pi(\theta_1) = \pi(\theta_2) = 1/2$. Verfeinert man Θ zu $\{\theta_1, \omega_1, \omega_2\}$, dann erhält jeder Parameter nur noch eine Wahrscheinlichkeit von $\pi(\theta_1) = \pi(\omega_1) = \pi(\omega_2) = 1/3$. Diesen Kritikpunkt könnte man abschwächen, indem man einen maximalen (fixen) Partitionsgrad einführt.

- Bei einem Parameterwechsel können Probleme auftreten:
  Sei $\theta \in \Theta$. Dann gehen wir von Θ auf $g(\Theta)$ mit $g : \theta \mapsto \eta$. Der Prior sollte durch die Transformation nicht verändert werden. Jedoch ist, falls $\pi(\theta) = 1$ auf Θ, die Verteilung für η nicht konstant, sondern $\pi^*(\eta) = |\frac{d}{d\eta} g^{-1}(\eta)|$.

5.2 Invariante Prior

Ein erster Lösungsansatz, um die Invarianzeigenschaft des Priors zu berücksichtigen, bilden Gruppen $\mathcal{G}$, die auf $\mathcal{X}$ wirken und Gruppen $\mathcal{G}^*$ induzieren, die auf Θ wirken.

Beispiel: Die Familie $f(x-\theta)$ ist translationsinvariant, d.h. die Verteilung von $y = x - x_0$ liegt in derselben Familie. θ heisst Lokationsparameter. In diesem Fall sollte der Prior invariant sein bezüglich Translationen. Man wählt den Prior so, dass $\pi(\theta) = \pi(\theta - \theta_0)\ \forall \theta_0$ gilt. Dies ist nur möglich, falls $\pi(\theta)$ konstant ist. Somit ist der Prior für alle Lokationsparameter θ gegeben durch $\pi(\theta) = c$.

Probleme: Da nicht vorausgesetzt wird, dass Θ kompakt ist, kann dies zu *improper Priorn* führen, d.h. $\pi(\theta)$ ist ein Mass, für das $\int \pi(\theta) d\theta = \infty$ mit $-\infty < c < \infty$ gilt. Jedoch kann man trotz des Improper Prior eine proper a posteriori Verteilung erhalten. Ein hinreichendes Kriterium hierfür ist, dass die Randverteilung
$m(x) = \int\limits_{\Theta} f(x|\theta)\pi(\theta) d\theta$ für alle x endlich ist. $\pi(\theta) = c$ wird dann als *flat prior* bezeichnet.

Kritik: Diese Invarianz Prior Methode ist aus verschiedenen Gründen nur teilweise zufriedenstellend:

- Es gibt mehrere Wege, eine a priori Verteilung zu wählen.
- Es existiert nicht immer eine invariante a priori Verteilung.
- Die geforderte Invarianzstruktur ist möglicherweise für den Entscheidungsträger nicht relevant.

5.3 Jeffreys Prior

Der Jeffreys Prior entstand aus der Kritik an den Invarianten Priorn. Er umgeht den Zwang einer Invarianzstruktur, ist aber dem Invarianzprior sehr ähnlich, falls Invarianz gefordert wird. Im Folgenden sei $\Theta \subset \mathbb{R}$ offen.

Definition Der *Jeffreys Prior* für eine Likelihoodfunktion $f(x|\theta)$ ist gegeben durch $\pi(\theta) \propto \sqrt{I(\theta)}$, wobei $I(\theta) = -E^{(x|\theta)}[\frac{\partial^2 \log(f(x|\theta))}{\partial \theta^2}]$ die Fisher Information ist.

Herleitung Sei $f(x \mid \theta)$ eine Likelihoodfunktion. Die zugehörige Fisher Information $I(\theta) = -E^{(x|\theta)}[\frac{\partial^2 \log(f(x|\theta))}{\partial \theta^2}]$ misst die Sensitivität des Likelihoods in der Nähe des MLE und ist proportional zu der erwarteten Krümmung des Likelihoods beim MLE.
Die Wurzel ist wichtig wegen des Invarianzprinzips:
Sei $\phi = h(\theta)$, wobei h Diffeomorphismus, und $g = h^{-1}$. Somit ist $g(\phi) = \theta$. Dann gilt

$$\pi(\phi) = \pi(g(\phi)) \, |\frac{dg(\phi)}{d\phi}| = \pi(\theta) \, |\frac{d\theta}{d\phi}| \tag{2}$$

Aber:

$$I(\phi) = -E^{(x|\phi)}[\frac{\partial^2 \log(f(x|\phi))}{\partial \phi^2}] = -E^{(x|\theta)}[\frac{\partial^2 \log(f(x|\theta))}{\partial \theta^2} |\frac{d\theta}{d\phi}|^2] = I(\theta)|\frac{d\theta}{d\phi}|^2$$

Daher wie in (2): $I^{1/2}(\phi) = I^{1/2}(\theta) \, |\frac{d\theta}{d\phi}|$

Beispiel $X_1, \ldots, X_n$ haben eine Likelihoodfunktion proportional zu $e^{-n(\bar{x}-\theta)^2/(2\sigma^2)}$ mit bekanntem σ^2. Da $\frac{\partial^2 \log(f(x|\theta))}{\partial \theta^2} = \frac{-n}{\sigma^2}$, ist der Jeffreys Prior $\pi(\theta) \propto (\frac{-n}{\sigma^2})^{1/2} \propto 1$.

Wenn für dieselben Daten das Mittel μ bekannt ist, aber die Varianz der Parameter of interest wäre, würde der Likelihood die folgende Gestalt haben: $f(x|\theta) = \theta^{n/2} e^{-s/(2\theta)}$, wobei $s = \sum_i (x_i - \mu)^2$ suffiziente Statistik.

Dann ist $\frac{\partial^2 \log(f(x|\theta))}{\partial \theta^2} = \frac{n}{(2\theta^2)} - \frac{s}{\theta^3}$. Da $E[s|\theta] = n\theta$ ist der Jeffreys Prior $\pi(\theta) = \frac{1}{\theta}$.

6 Literatur

[1] *The Bayesian Choice*, Robert, Ch. ,2nd ed., Springer NY

[2] *Handout 5* von www2.isye.gatech.edu/∼brani/isyebayes/handouts.html

[3] *Mathematische Statistik*, Vorlesungsmitschrift von Prof. Sara van de Geer, HS 2007