Bayessche Statistik

Leonard Pleschberger

30. Oktober 2019

Ziele des Vortrags

Mit dem Vortrag sollen einige Grundprinzipien der Bayesschen Inferenz dargestellt werden, insbesondere:

1. Die Schritte des Verfahrens.
2. Die Auswahl von geeigneten A-priori-Verteilungen unter verschiedenen Gesichtspunkten.

Bayessche Statistik

Bayessche Statistik

1. ist eine alternative Methodik zur Datenanalyse.
2. verwendet explizit das Vorwissen des Statistikers.
3. basiert auf einem alternativen Wahrscheinlichkeitsbegriff.
4. ermöglicht Aussagen über nicht-wiederholbare Ereignisse.

Bayessche vs. frequentistische Statistik

Hauptunterschied zwischen den Modellen:

Frequentistische Statistik: Parameter θ ist fest, aber unbekannt.

Bayessche Statistik: Auch θ besitzt eine Verteilung.

Schritte der Bayesschen Methode

Die Bayessche Methode ist ein Lernprozess:

1. Stelle ein vernünftiges Modell auf. ($\Rightarrow$ **Likelihood**)
2. Wähle eine Startverteilung. ($\Rightarrow$ **A-priori-Verteilung**)
3. Erhalte ein neues Modell. ($\Rightarrow$ **A-posteriori-Verteilung**)
4. Ist das neue Modell sinnvoll oder nicht wirklich?

Schritte wiederholen mit A-posteriori- als neuer A-priori-Verteilung.

Notation

θ : Parameter des Modells

y : Daten

$p(\cdot)$: Verteilung / Randverteilung / Dichte

$\propto$: Verteilungen sind proportional zueinander, d.h.

$$x \propto y \quad :\Leftrightarrow \quad x = C \cdot y, \quad \text{für } C > 0.$$

Satz von Bayes

Für die **A-posteriori-Verteilung** $p(\theta|y)$ gilt

$$p(\theta|y) = \frac{p(\theta) \cdot p(y|\theta)}{p(y)}.$$

$p(y)$ ist eine (Normierungs-)Konstante. Wir schreiben:

$$p(\theta|y) \propto p(\theta) \cdot p(y|\theta).$$

$$\text{A-posteriori-Verteilung} \propto \text{A-priori-Verteilung} \cdot \text{Likelihood}.$$

Beispiel: Rezessive Erbkrankheit

Geschlechtschromosome:

XY : ♂, XX : ♀
Vater und Mutter geben je ein Chromosom weiter ($p = 1/2$).

Hämophilie (Bluterkrankheit):

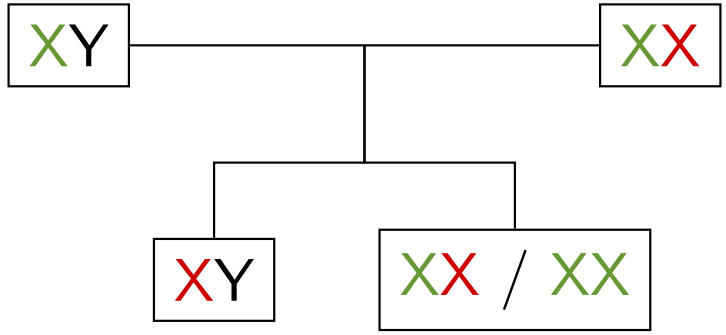
Wird über X-Chromosom vererbt.
X : Normales X-Chromosom, X : X-Chromosom mit gen. Variation.

Hämophilie ist **rezessiv**:

XX : Gesunde ♀, XX : Erkrankte ♀, XY : Erkrankter ♂

Ausgangssituation

Wir betrachten eine Frau mit gesunden Eltern und einem an Hämophilie erkrankten Bruder:



Ist die Frau Kondukorin, d.h. Trägerin der genetischen Variation?

Parameter: $\theta = 0$: XX , $\theta = 1$: XX

A-priori-Verteilung: $p(\theta = 1) = p(\theta = 0) = 0.5$.

Likelihood

Wir betrachten als **Daten** die Söhne y_1 und y_2 der Frau, mit

$$y_i = \begin{cases} 1 : \text{Sohn erkrankt } (\mathrm{XY}), \\ 0 : \text{Sohn gesund } \quad (\mathrm{XY}), \end{cases} \qquad i \in 1, 2.$$

Es ergeben sich die **Likelihoods**:

$$\begin{aligned} p(y_1 = 0, y_2 = 0 | \theta = 1) &= p(y_1 = 0 | \theta = 1) \cdot p(y_2 = 0 | \theta = 1) \\ &= 0.5 \cdot 0.5 = 0.25. \end{aligned}$$

$$p(y_1 = 0, y_2 = 0 | \theta = 0) = 1 \cdot 1 = 1.$$

A-posteriori-Verteilung

Mit welcher Wahrscheinlichkeit ist die Frau Konduktorin, wenn ihre beiden Söhne gesund sind?

$$p(\theta = 1|y) = \frac{0.5 \cdot 0.25}{0.25 \cdot 0.5 + 1 \cdot 0.5} = 0.20.$$

A-posteriori- als neue A-priori-Verteilung

Falls es einen weiteren gesunden Sohn gibt:

Die A-posteriori-Verteilung kann als neue A-priori-Verteilung verwendet werden.

$$p(\theta = 1 | y_1, y_2, y_3)$$

$$= \frac{p(\theta = 1 | y_1, y_2) \cdot p(\theta = 1)}{p(y_1, y_2, y_3 | \theta = 1) \cdot p(\theta = 1) + p(y_1, y_2, y_3 | \theta = 0) \cdot p(\theta = 0)}$$

$$= \frac{0.20 \cdot 0.50}{0.50 \cdot 0.20 + 1 \cdot 0.8}$$

$$\approx 0.111.$$

Auswertung der Daten

Die A-posteriori-Verteilung muss noch auf Plausibilität überprüft werden:

1. Passen die Ergebnisse zu den Daten?
2. Gibt es weitere A-priori-Informationen, die noch nicht verwendet wurden?

Komplexität des Modells

Man soll für das Modellfitting kein riesiges Modell bauen, das über Nacht läuft.

Lieber kleinschrittig vorgehen:

1. Mit einfachen Modellen und wenigen verschiedenen Datentypen beginnen.

2. Allmählich die Komplexität des Modells steigern.

Beispiel: Rechtschreibekorrektur von Google

Wir betrachten einen Satz wie: The data are totally radom.

„radom“ kann ein Rechtschreibefehler sein. Mögliche Wörter sind:

1. „random“.
2. „radon“ (radioaktives Edelgas).
3. „radom“ (bewusster Fehler wie oben).

Aufstellen eines Wahrscheinlichkeitsmodells

Erfassen der Daten und Wahl des Parametergrundraums:

y: Getipptes Wort „radom“.

θ: Gewünschte Wort, zur Vereinfachung nur „random“, „radon“oder „radom“möglich.

Mit dem Satz von Bayes gilt für die A-posteriori-Verteilung

$$p(\theta|y = \text{„radom“}) \propto p(\theta) \cdot p(y = \text{„radom“}|\theta)$$

A-priori-Verteilung und Likelihood

A-priori-Verteilung $p(\theta)$: Relative Häufigkeiten der möglichen Wörter in der Google-Datenbank.

Likelihood $p(y|\theta)$: Aus Rechtschreibekorrektur-Modell von Google.

θ	$p(\theta)$	$p(y = \text{„radom“}|\theta)$
„random“	$7,60 \cdot 10^{-5}$	0.00193
„radon“	$6,05 \cdot 10^{-6}$	0.000143
„radom“	$3,12 \cdot 10^{-7}$	0.975

Die bedingte Wahrscheinlichkeit für „radom“erscheint sehr hoch. Radom ist eine Großstadt in Polen.

A-posteriori-Verteilung

Mit dem Satz von Bayes ergibt sich die **A-posteriori-Verteilung**:

θ	$p(\theta) \cdot p(\text{„radom“}\|\theta)$	$p(\theta\|\text{„radom“})$
„random“	$1.47 \cdot 10^{-7}$	0.325
„radon“	$8.65 \cdot 10^{-10}$	0.002
„radom“	$3.04 \cdot 10^{-7}$	0.673

Auswertung der Ergebnisse

Die A-posteriori-Wahrscheinlichkeit für ein bewusstes „radom“ erscheint sehr hoch.

Möglichkeit 1: „radom“ wird als korrekt akzeptiert.

Möglichkeit 2: Weitere A-priori-Informationen werden einbezogen.

Weitere A-priori-Informationen

Wir fügen die Textquelle x als weitere A-priori-Information hinzu.
Der Satz von Bayes liefert:

$$p(\theta|y,x) \propto \; p(\theta|x) \cdot p(y|\theta,x).$$

Radom ist eine polnische Großstadt. Zu erwarten wäre:

x	$p(\theta = \text{„radom“}|x)$
Statistik-Buch	Sehr unwahrscheinlich
Polnischer Reiseführer	Möglich

A-priori-Verteilungen für Standard-Modelle

Mögliche einparametrige Wahrscheinlichkeitsmodelle mit Parameter θ sind:

Binomialverteilung	$\text{Bin}(n, \theta)$
Normalverteilung	$\mathcal{N}(\theta, \sigma^2)$ bzw. $\mathcal{N}(\mu, \theta)$
Poissonverteilung	$\text{Poi}(\theta)$

Die Anzahl n in der Binomialverteilung wird durch die Daten vorgegeben.

Bei der Normalverteilung seien entweder die Varianz σ^2 oder der Erwartungswert μ bekannt.

A-priori-Verteilungen für Standard-Modelle

Die Wahl des Wahrscheinlichkeitsmodells ergibt die **Likelihood**.

Ziel: Eine für das Modell geeignete **A-priori-Verteilung** wählen.

Manche A-priori-Verteilungen sind geeigneter als andere.
Wünschenswert sind:

- Einfache Rechnungen.
- A-posteriori-Verteilung in schöner Form (bekannte Verteilung).
- Viel Information allein durch die Daten.

A-posterior-Verteilung

Die A-postiori-Verteilung hat im Schnitt eine **geringere Varianz** als die A-priori-Verteilung:

$$\mathrm{Var}(\theta) = \mathbb{E}[\mathrm{Var}(\theta|y)] + \mathrm{Var}(\mathbb{E}[\theta|y])$$

Im Schnitt konzentriert sich die A-posteriori-Verteilung also stärker um den Erwartungswert.

Binomialmodell: Mädchenanteil

Wie groß ist der Anteil weiblicher Babys bei n Geburten?

Modell: Zwei mögliche Ausgänge $\Rightarrow$ Bin(n, p)

y: Anzahl der Mädchen bei n Geburten.

θ: Wahrscheinlichkeit p für ein Mädchen.

Binomialmodell: Bestimmung der Likelihood

Mit obiger Notation ergibt sich als **Likelihood**:

$$p(y|\theta) = \binom{n}{y} \cdot \theta^y \cdot (1-\theta)^{n-y} \propto \theta^y \cdot (1-\theta)^{n-y}.$$

Da $\binom{n}{y}$ bereits aus den Daten vollständig bestimmt ist, hängt es nicht von θ ab.

Es kann als Konstante behandelt werden, was $\propto$ rechtfertigt.

Binomialmodell: Wahl einer A-priori-Verteilung

Mögliche **A-priori-Verteilung** ist $\text{Unif}_{[0,1]}$ mit zwei Begründungen:

Bayes hat gezeigt, dass mit $p(\theta) = \text{Unif}_{[0,1]}$ gilt:

$$p(y) = \int_0^1 \binom{n}{y} \theta^y (1-\theta)^{n-y} d\theta = \frac{1}{n+1} = \text{Unif}_{\{0,1,\cdots,n\}}, \text{ a priori.}$$

Laplace folgt dem „Prinzip des unzureichenden Grundes“:

Man soll die Gleichverteilung wählen, wenn sonst kein Vorwissen vorliegt.

Binomialmodell: Wahl einer A-priori-Verteilung

Es ergibt sich mit $p(y) = 1/(n+1)$ die **A-posteriori-Verteilung**

$$p(\theta|y) = \frac{\Gamma(n+2)}{\Gamma(y+1)\cdot\Gamma(n-y+1)}\theta^y\cdot(1-\theta)^{n-y} = \text{Beta}(y+1, n-y+1).$$

Vorteil dieser A-posteriori-Verteilung:

Erwartungswert, Median, Standardabweichung und Quantile sind bekannt.

Binomialmodell: Verallgemeinerung der A-priori-Verteilung

Wir wählen nun als **A-priori-Verteilung**:

$p(\theta) = \text{Beta}(\alpha, \beta)$ mit Hyperparameter (α, β).

„Hyper“bezieht sich auf die Parametrisierung des Parameters θ.

Da Beta(1,1) $= \text{Unif}_{[0,1]}$, stellt dies eine Verallgemeinerung zu obiger Situation dar.

Binomialmodell: Berechnung der A-posteriori-Verteilung

Mit der **Likelihood** : Bin(n, θ) und der **A-priori-Verteilung** : Beta(α, β) gilt für die **A-posteriori-Verteilung**:

$$
\begin{aligned}
p(\theta|y) &\propto p(\theta) \cdot p(y|\theta) \\
&= \theta^{\alpha-1}(1-\theta)^{\beta-1} \cdot \theta^{y}(1-\theta)^{n-y} \\
&= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \\
&= \text{Beta}(\theta|\alpha+y, \beta+n-y).
\end{aligned}
$$

Konjugierte Verteilung

A-priori-Verteilung und A-posteriori-Verteilung sind beide Beta-verteilt.

Man sagt: A-priori-Verteilung und A-posteriori-Verteilung sind **konjungiert**.

Beta-Verteilung : **Konjugationsfamilie** des Binomialmodells.

Konjugierte Verteilung: Vorteile

Die Konjugation ist nützlich:

Bei wiederholtem Bayesschen Schließen dient die A-posteriori-Verteilung als A-priori-Verteilung.

$\Rightarrow$ Wir bleiben stets in derselben Verteilungsfamilie.

Ist die A-posteriori-Verteilung eine bekannte Verteilung, erhält man sofort: Erwartungswert, Median, Standardabweichung und Quantile, sofern diese existieren.

Konjugation: Verallgemeinerung

Nur Verteilungen der **Exponentialfamilie** haben nat. Konjugation:

$$p(y_1, \cdots, y_n|\theta) \propto g(\theta)^n \exp(\phi(\theta)t(y)).$$

Dies hängt von $\boldsymbol{y}$ nur ab die durch **suffiziente Statistik**

$$t(\boldsymbol{y}) = \sum_{i=1}^{n} u(y_i).$$

$\phi(\theta)$ bezecihnet den **natürlichen Parameter** der Familie.

Konjugation: A-priori- und A-posteriori-Verteilung

Wir wählen als **A-priori-Verteilung** aus der Exponentialfamilie

$$p(\theta) \propto g(\theta)^{\eta} \exp(\phi(\theta) \cdot \nu).$$

Wir erhalten als **A-posteriori-Verteilung**

$$p(\theta|y) \propto g(\theta)^{n+\eta} \exp(\phi(\theta) \cdot (\nu + t(y))).$$

Diese stammt wieder aus der Exponentialfamilie.

Beispiel: Normalverteilung bei bekannter Varianz

Nun betrachten wir verschiedene Standardmodelle und bestimmten die konjugierten A-priori-Verteilungen.

Wir nehmen eine normalverteilte **Likelihood** $\sim \mathcal{N}(\theta, \sigma^2)$ mit bekannter Varianz σ^2 an und wollen für eine Beobachtung y den Erwartungswert θ schätzen.

Dies ist ist oft durch den **ZGS** gerechtfertigt.

Die zugehörige Lebesgue-Dichte lautet

$$p(y|\theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(y-\theta)^2\right).$$

Beispiel: Normalverteilung bei bekannter Varianz

Die **konjugierte A-priori-Verteilung** $\sim \mathcal{N}(\mu_0, \tau_0^2)$ ist:

$$p(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right).$$

Algebraische Umformungen ergeben die **A-posteriori-Verteilung**

$$p(\theta|y) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right) \cdot \exp\left(-\frac{1}{2\sigma^2}(y - \theta)^2\right)$$

$$= \exp\left(-\frac{1}{2\left(\frac{\tau_0^2 \cdot \sigma^2}{\sigma^2 + \tau_0^2}\right)}\left(\theta - \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}\right)^2\right).$$

Beispiel: Normalverteilung bei bekannter Varianz

Somit gilt für die A-posteriori-Verteilung

$$\theta|y \sim \mathcal{N}(\mu_1, \tau_1^2) \text{ mit } \mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}} \text{ und } \frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}.$$

Wir definieren die **Präzision** einer Verteilung als

$$\text{Präzision}_y = \frac{1}{\sigma^2} \text{ und } \text{Präzision}_\theta = \frac{1}{\tau_0^2}.$$

Der **A-posteriori-Erwartungswert** ist das mit Präzisionen gewichtete Mittel aus **A-priori-Erwartungswert** μ_0 und **Beobachtung** y.

Schätzen des A-posteriori-Erwartungswerts

Für die Verteilung einer weiteren $\mathcal{N}(\theta, \sigma^2)$-verteilten ZV $\tilde{y}$ gilt:

$$
\begin{aligned}
p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y) d\theta \\
&= \int p(\tilde{y}|\theta, y) p(\theta|y) d\theta \\
&= \int p(\tilde{y}|\theta) p(\theta|y) d\theta \\
&\propto \int \exp\left(-\frac{1}{2\sigma^2}(\tilde{y} - \theta)^2\right) \exp\left(-\frac{1}{2\tau^2}(\theta - \mu_1)^2\right) d\theta.
\end{aligned}
$$

Schätzen des A-posteriori-Erwartungswerts

$\tilde{y}$ und θ sind gemeinsam a-posteriori-normalverteilt.

Somit ist die Randdichte von $\tilde{y}$ ebenfalls normalverteilt.

Nun lässt sich der **A-posteriori-Erwartungswert** berechnen:

$$\mathbb{E}[\tilde{y}|y] = \mathbb{E}[\underbrace{\mathbb{E}[\tilde{y}|\theta, y]}_{=\theta}|y] = \mathbb{E}[\theta|y] = \mu_1.$$

Für die **A-posteriori-Varianz** gilt:

$$\mathrm{Var}(\tilde{y}|y) = \sigma^2 + \tau_1^2 = \mathrm{Var}_y + \mathrm{Var}_{\mathrm{Posterior}}.$$

Beispiel: Poisson-Modell

Nun gehen wir von u.i.v. Posson-verteilten ZV $y_1, \ldots, y_n$ mit Parameter θ aus.

Wir betrachten den multivariaten Zufallsvektor $\boldsymbol{y} = (y_1, \cdots, y_n)$.

Die zugehörige **Likelihood** lautet

$$p(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} \frac{1}{y_i!} \theta^{y_i} e^{-\theta}$$

$$\propto \theta^{t(\boldsymbol{y})} e^{-n\theta}, \text{ mit suffizienter Statistik } t(\boldsymbol{y}) = \sum_{i=1}^{n} y_i$$

$$\propto e^{-n\theta} e^{t(\boldsymbol{y}) \log(\theta)}, \text{ mit nat. Parameter } \phi(\theta) = \log(\theta).$$

Beispiel: Poisson-Modell

Als **konjugierte A-priori-Verteilung** wählen wir

$$p(\theta) \propto (e^{-\theta})^{\eta} e^{\nu \log \theta}, \text{ mit Hyperparameter } (\eta, \nu).$$

Durch Reparametrisierung erhalten wir

$$p(\theta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} e^{-\beta\theta} \theta^{\alpha-1} = \text{ Gamma}(\alpha, \beta).$$

Es ergibt sich die **A-posteriori-Verteilung**

$$\theta | y \sim \text{Gamma}(\alpha + n\bar{\boldsymbol{y}}, \beta + n).$$

Informative vs. nichtinformative A-priori-Verteilungen

Bisher haben wir A-priori-Verteilungen betrachtet, die Informationen beinhalten.

$\Rightarrow$ **Informative A-priori-Verteilungen**.

Nun möchten wir A-priori-Verteilungen wählen, die die A-posteriori-Verteilung möglichst wenig beeinflussen.

$\Rightarrow$ **Nichtinformative A-priori-Verteilungen**.

Jeffreys Invarianzprinzip

Wir betrachten Transformationen $\phi = h(\theta)$ des Parameters θ.

Für eine bijektive Abbildung h erhalten wir die Identität

$$p(\phi) = p(h(\theta)) = p(\theta) \cdot |J_h(\theta)|^{-1} = p(\theta) \cdot |h'(\theta)|^{-1}.$$

Jeffreys Prinzip lautet nun:

Der Informationsgehalt der A-priori-Verteilung $p(\theta)$ soll gleich dem der transformierten A-priori-Verteilung $p(\phi) = p(h(\theta))$ sein.

Jeffreys Invarianzprinzip

Ziel: Wahl einer A-priori-Verteilung, die invariant gegenüber Reparametrisierung ist.

Dies wird mit $p(\theta) \propto \sqrt{J(\theta)}$ erreicht, wobei $J(\theta)$ die **Fisher-Information** von θ bezeichnet, definiert durch

$$J(\theta) = \mathbb{E}\left[\left(\frac{d}{d\theta}\log p(y|\theta)\right)^2 \middle| \theta\right]$$

$$= -\mathbb{E}\left[\frac{d^2}{d\theta^2}\log p(y|\theta)\middle|\theta\right].$$

Jeffreys Invarianzprinzip

Dass Jeffreys A-priori-Modell invariant unter Reparametrisierung ist, sehen wir durch Auswertung von $J(\phi)$ in $\theta = h^{-1}(\phi)$:

$$
\begin{aligned}
J(\phi) &= -\mathbb{E}\left[\frac{d^2}{d\phi^2}\log p(y|\phi)\right] \\
&= -\mathbb{E}\left[\frac{d^2}{d\theta^2}\log p\left(y|\theta = h^{-1}(\phi)\right)\cdot|h'(\theta)|^{-2}\right] \\
&= J(\theta)\cdot|h'(\theta)|^{-2}.
\end{aligned}
$$

Damit ist $\sqrt{J(\phi)} = \sqrt{J(\theta)}\cdot|h'(\theta)|^{-1}$ gezeigt.

Literatur

Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D (2013): Bayesian Data Analysis (3. ed.), CRC Press.