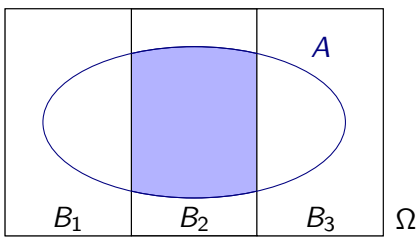
Bayessche Parameterschätzung

Florian Wickelmaier

Bayessche Schätzverfahren

Rückblick: Die Bayessche Formel

- Bilden die Ereignisse $B_1, \ldots, B_m$ eine Ereignisdisjunktion, so gilt für alle Ereignisse A und $j = 1, \ldots, m$



$$
\begin{aligned}
P(B_j \mid A) &= \frac{P(A \mid B_j) \cdot P(B_j)}{P(A)} \\
&= \frac{P(A \mid B_j) \cdot P(B_j)}{\sum_{j=1}^{m} P(A \cap B_j)} \\
&= \frac{P(A \mid B_j) \cdot P(B_j)}{\sum_{j=1}^{m} P(A \mid B_j) \cdot P(B_j)}
\end{aligned}
$$

Beispielhafte Anwendung der Bayesschen Formel

Beispiel: Aufnahmeabteilung eines psychiatrischen Krankenhauses

Die Ereignisse B_1, B_2, B_3 sollen verschiedene Diagnosen kennzeichnen, z. B.

B_1	neurotisch
B_2	psychotisch
B_3	psychisch gesund

Die *A-priori-Wahrscheinlichkeiten* $P(B_1), P(B_2), P(B_3)$ kennzeichnen die (z. B. aus langjährigen Erfahrungen resultierende) Verteilung über die Diagnosen, noch vor einer Kenntnis der individuellen Symptomatik

Beispielhafte Anwendung der Bayesschen Formel

Das Ereignis A repräsentiere ein empirisches Datum, z. B. die Beobachtung eines Symptoms (wie etwa "hört Stimmen")

Die bedingten Wahrscheinlichkeiten $P(A \mid B_j)$, $j = 1, 2, 3$, kennzeichnen dann die Auftretenswahrscheinlichkeit des Symptoms A, gegeben die Diagnose B_j trifft zu

Die Bayessche Formel erlaubt dann die Berechnung der *A-posteriori-Wahrscheinlichkeit* $P(B_j \mid A)$ der Diagnose B_j gegeben das beobachtete Symptom A

Ist das Symptom informativ, dann werden sich die A-posteriori-Wahrscheinlichkeiten von den A-priori-Wahrscheinlichkeiten unterscheiden

Parameter als Zufallsvariable

Die im Beispiel deutlich gewordene Vorgehensweise wird nun auf die Parameterschätzung angewendet

Dabei wird von folgendem Grundverständnis ausgegangen

- Da der Wert eines Parameters zunächst nicht definitiv fest liegt, wird er als Zufallsvariable Θ betrachtet
- Ein Parameterwert ϑ wird damit als Realisierung der Zufallsvariablen Θ angesehen
- Die Unsicherheit über den Wert ϑ des Parameters wird durch eine Verteilung repräsentiert, die Wahrscheinlichkeitsverteilung der Zufallsvariablen Θ

A-priori-Verteilung

Bevor nun Daten aus einer Stichprobenziehung vorliegen, wird die Unsicherheit bezüglich des Parameterwertes durch die *A-priori-Verteilung* (prior distribution) beschrieben

- Diese A-priori-Verteilung kann Vorannahmen, Vorwissen oder Überzeugungen hinsichtlich des Parameterwertes repräsentieren oder jeden möglichen Parameterwert mit gleicher Wahrscheinlichkeit zulassen (nicht informative A-priori-Verteilung; uninformative prior)
- Es wird auch davon gesprochen, dass die Bayesschen Schätzverfahren auf einem *subjektiven Wahrscheinlichkeitsbegriff* basieren

  Dieser subjektive Wahrscheinlichkeitsbegriff wird einem *frequentistischen* gegenüber gestellt, der auf einer Interpretation von Wahrscheinlichkeiten als empirischen Häufigkeiten aufbaut

A-posteriori-Verteilung

Sobald nun Daten vorliegen, wird mit Hilfe einer geeigneten Version der Bayesschen Formel die *A-posteriori-Verteilung* (posterior distribution) berechnet

- Die A-posteriori-Verteilung der Zufallsvariablen Θ wird durch die Berücksichtigung der Daten mehr oder weniger deutlich von der A-priori-Verteilung abweichen (Bayesian learning)

Die A-posteriori-Verteilung bildet dann den Ausgangspunkt für die Bestimmung des (Punkt-)Schätzers für den unbekannten Parameter, z. B. durch deren

- Erwartungswert (im Mittel erwarteter Wert)
- Modus ("wahrscheinlichster Wert")

Notation

$p(y_1, \ldots, y_n \mid \vartheta)$ bezeichnet die Wahrscheinlichkeitsdichte (bzw. Wahrscheinlichkeitsfunktion) für die Daten (also für $Y_1 = y_1, \ldots, Y_n = y_n$), gegeben es gilt $\Theta = \vartheta$

$p(y_1, \ldots, y_n)$ bezeichnet die Randdichte (bzw. Randverteilung) der Daten

$p(\vartheta)$ bezeichnet die A-priori-Dichte (bzw. A-priori-Verteilung) von Θ (also die Randdichte von Θ)

$p(\vartheta \mid y_1, \ldots, y_n)$ bezeichnet die A-posteriori-Wahrscheinlichkeitsdichte (bzw. Wahrscheinlichkeitsfunktion) von Θ, gegeben die beobachteten Daten $y_1, \ldots, y_n$

Bayessche Inferenz

Die Formel von Bayes lautet nun allgemein

$$p(\vartheta \mid y_1, \ldots, y_n) = \frac{p(y_1, \ldots, y_n \mid \vartheta) \cdot p(\vartheta)}{p(y_1, \ldots, y_n)}$$

Der Nenner hängt nur von den Daten, aber nicht vom Parameter ab!

Darüber hinaus ist

$$p(y_1, \ldots, y_n \mid \vartheta) = p(y_1 \mid \vartheta) \cdot \ldots \cdot p(y_n \mid \vartheta) = L(\vartheta)$$

die Likelihoodfunktion

Randdichte der Daten

Da die Zufallsvariable Θ im Allgemeinen stetig ist (z. B. wenn der Parameter Werte in einem reellen Intervall annehmen kann), erhalten wir die Randdichte $p(y_1, \ldots, y_n)$ durch Integration über alle möglichen Werte ϑ

$$p(y_1, \ldots, y_n) = \int p(y_1, \ldots, y_n \mid \vartheta) \cdot p(\vartheta) d\vartheta$$

Dieses Vorgehen wird auch so beschrieben

- Die Abhängigkeit vom unbekannten Parameterwert ϑ wird "herausintegriert"

Die Charakterisierungen der Dichten $p(y_1, \ldots, y_n \mid \vartheta)$ und $p(y_1, \ldots, y_n)$ lassen sich in die Formel von Bayes einsetzen

Bayessche Inferenz

Insgesamt erhält man also folgende alternative Darstellungen der *Bayesschen Inferenz*

$$
\begin{aligned}
p(\vartheta \mid y_1, \ldots, y_n) &= \frac{p(y_1, \ldots, y_n \mid \vartheta) \cdot p(\vartheta)}{\int p(y_1, \ldots, y_n \mid \vartheta) \cdot p(\vartheta) d\vartheta} \\
&= \frac{p(y_1 \mid \vartheta) \cdot \ldots \cdot p(y_n \mid \vartheta) \cdot p(\vartheta)}{\int p(y_1 \mid \vartheta) \cdot \ldots \cdot p(y_n \mid \vartheta) \cdot p(\vartheta) d\vartheta} \\
&= \frac{L(\vartheta) \cdot p(\vartheta)}{\int L(\vartheta) \cdot p(\vartheta) d\vartheta}
\end{aligned}
$$

EAP-Schätzer

Wählt man als Punktschätzer für den unbekannten Parameter den Erwartungswert der A-posteriori-Verteilung, dann bildet man

$$\hat{\vartheta}_{EAP} = E(\vartheta \mid y_1, \ldots, y_n) = \int \vartheta \cdot p(\vartheta \mid y_1, \ldots, y_n) d\vartheta$$

Zur expliziten Berechnung der A-posteriori-Verteilung muss das Integral $\int L(\vartheta) \cdot p(\vartheta) d\vartheta$ berechnet werden

Diese Berechnung ist analytisch meist nicht möglich, sondern es werden numerische Integrationsverfahren eingesetzt oder Simulationstechniken benutzt

Markov-Chain-Monte-Carlo-Schätzung

- Mit den Ziehungen aus der A-posteriori-Verteilung $\vartheta_1, \ldots, \vartheta_m \sim p(\vartheta \mid y_1, \ldots, y_n)$ ist der MCMC-Schätzer

$$\hat{E}(\vartheta \mid y_1, \ldots, y_n) = \frac{1}{m} \sum \vartheta_i$$

MAP-Schätzer

Diese Problematik umgeht man, wenn man als Punktschätzer für den unbekannten Parameter den Modus der A-posteriori-Verteilung verwendet, den *Maximum-a-posteriori-Schätzer* (MAP)

Da der Nenner ja nicht mehr von ϑ abhängt, kann der Modus durch Maximieren des Zählerausdrucks $L(\vartheta) \cdot p(\vartheta)$ (bzw. des Logarithmus $\log L(\vartheta) + \log p(\vartheta)$) erfolgen

Man wählt also den Parameterwert $\hat{\vartheta}_{MAP}$ so, dass gilt

$$L(\hat{\vartheta}_{MAP}) \cdot p(\hat{\vartheta}_{MAP}) = \max_{\vartheta} L(\vartheta) \cdot p(\vartheta)$$

Damit wird auch die Beziehung zur Maximum-Likelihood-Schätzung deutlich

- Anstatt der Likelihood $L(\vartheta)$ wird hier die mit der A-priori-Verteilung $p(\vartheta)$ gewichtete Funktion $L(\vartheta) \cdot p(\vartheta)$ maximiert

Beziehung zur Maximum-Likelihood-Schätzung

Ist die A-priori-Dichte $p(\vartheta)$ sehr flach, so drückt dies wenig Vorwissen bezüglich des Wertes ϑ aus

Im Extremfall ist $p(\vartheta)$ beispielsweise eine Gleichverteilung über einem Intervall $[-c, c]$ mit einer sehr großen reellen Zahl c

Dann sind der ML-Schätzer $\hat{\vartheta}$ und der MAP-Schätzer $\hat{\vartheta}_{MAP}$ kaum unterscheidbar und stimmen für den Grenzfall $c \to \infty$ sogar völlig überein

Bayessche Schätzung bei der Binomialverteilung

Als Beispiel betrachten wir die Präsentation eines Reizpaares (x, s) im Kontext der Konstanzmethode

Zur Bestimmung der Wahrscheinlichkeit ϑ, mit der eine Versuchsperson den Vergleichsreiz x über den Standardreiz s bevorzugt, erheben wir über eine Stichprobe binärer Zufallsvariablen $Y_1, \ldots, Y_n$ die Daten $y_1, \ldots, y_n$

Die Likelihoodfunktion ist dann durch die Bernoulli-Verteilung gegeben

$$L(\vartheta) = \vartheta^{\sum_{i=1}^{n} y_i} \cdot (1 - \vartheta)^{n - \sum_{i=1}^{n} y_i}$$

Beispiel

Wir nehmen an, dass die Versuchsperson in 20 von 30 Darbietungen den Vergleichsreiz als intensiver als den Standardreiz beurteilt

Da der ML-Schätzer die relative Häufigkeit ist, erhält man dann den Schätzwert

$$\hat{\vartheta} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{20}{30} = \frac{2}{3}$$

Der ML-Schätzer nutzt keinerlei Vorinformation, die z. B. darin bestehen kann, dass bekannt ist, dass der Vergleichsreiz physikalisch intensiver (oder nicht intensiver) als der Standardreiz ist

Beispiel

Das Vorwissen über den Wert des Parameters ϑ beschreiben wir durch die Dichtefunktion der Beta-Verteilung

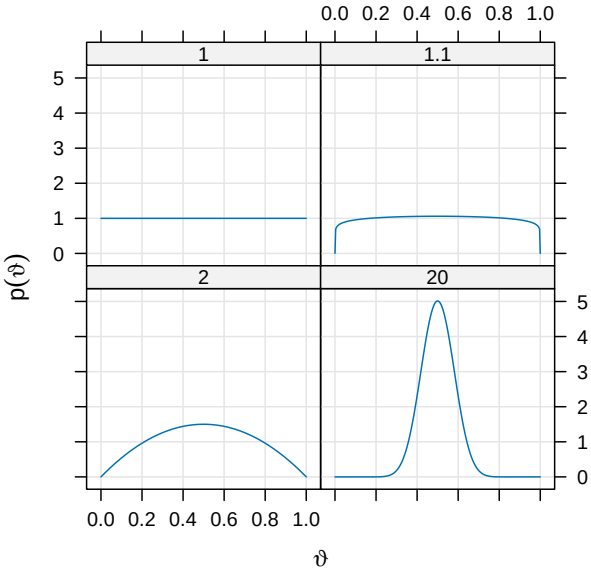
$$\begin{aligned} p(\vartheta) &= \text{beta}(a, b) \\ &= \frac{\vartheta^{a-1} \cdot (1-\vartheta)^{b-1}}{B(a, b)} \\ &= \frac{\vartheta^{a-1} \cdot (1-\vartheta)^{b-1}}{\int_0^1 \vartheta^{a-1} \cdot (1-\vartheta)^{b-1} d\vartheta} \\ &= \frac{\vartheta^{a-1} \cdot (1-\vartheta)^{b-1}}{\frac{\Gamma(a) \cdot \Gamma(b)}{\Gamma(a+b)}} \end{aligned}$$

mit den "Hyperparametern" $a, b > 0$ und der Gamma-Funktion $\Gamma(x) = (x-1)!$ (falls x positiv ganzzahlig)

Beispiel

Mit den Hyperparametern steuern wir die Spezifität der Vorannahmen über den Parameter ϑ

- Für $a = b = 1$ gilt $p(\vartheta) = 1/B(1, 1)$, sodass alle Parameterwerte gleich wahrscheinlich sind (uninformative prior), während sich mit steigenden a- und b-Werten die Dichtefunktion zunehmend konzentriert (für $a = b$ um den Wert 0.5)
- A-priori-Verteilungen $p(\vartheta)$ für $a, b \in \{1, 1.1, 2, 20\}$



Beispiel

Die Konzentration der Wahrscheinlichkeitsmasse um 0.5 kann z. B. durch das Vorwissen motiviert sein, dass der Vergleichsreiz hinsichtlich der physikalischen Reizintensität relativ nahe beim Standardreiz liegt

Zur Bestimmung des MAP-Schätzers $\hat{\vartheta}_{MAP}$ betrachten wir dann die Funktion

$$\begin{aligned} L(\vartheta) \cdot p(\vartheta) &= \vartheta^{\sum_{i=1}^{n} y_i} \cdot (1-\vartheta)^{n-\sum_{i=1}^{n} y_i} \cdot \frac{\vartheta^{a-1} \cdot (1-\vartheta)^{b-1}}{B(a,b)} \\ &= \frac{\vartheta^{\sum_{i=1}^{n} y_i + a - 1} \cdot (1-\vartheta)^{n-\sum_{i=1}^{n} y_i + b - 1}}{B(a,b)} \end{aligned}$$

Beispiel

Für deren Logarithmus gilt

$$\log L(\vartheta) + \log p(\vartheta) = \log \vartheta \left(\sum_{i=1}^{n} y_i + a - 1 \right) + \log(1-\vartheta) \left(n - \sum_{i=1}^{n} y_i + b - 1 \right) - \log B(a,b)$$

Ableiten nach ϑ und Nullsetzen liefert dann den MAP-Schätzer

$$\hat{\vartheta}_{MAP} = \frac{\sum_{i=1}^{n} y_i + a - 1}{n + a + b - 2}$$

Beispiel

Die Hyperparameter steuern in dieser Schätzgleichung die Gewichtung des Einflusses der Daten bzw. des Vorwissens

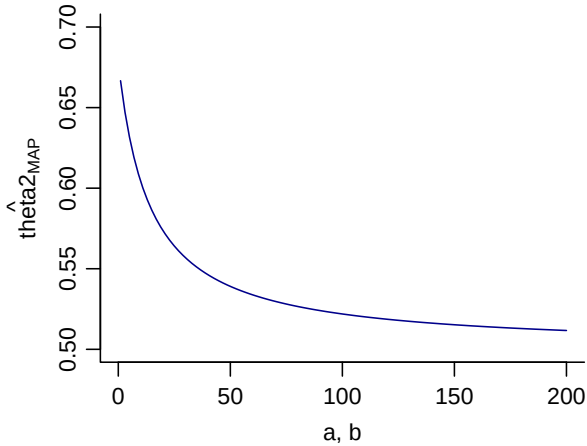
- Für $a = b = 1$ ist der MAP-Schätzer identisch mit dem ML-Schätzer, also der relativen Häufigkeit, was mit dem Fehlen jeglichen Vorwissens korrespondiert ("flache" A-priori-Verteilung)
- Für steigende a- und b-Werte wird der Quotient zunehmend durch das Verhältnis der "A-priori-Treffer" a im Zähler und der "A-priori-Durchgänge" $a + b$ im Nenner dominiert und erreicht im Grenzfall $a, b \to \infty$ weitgehend unabhängig von den Daten den Wert $\frac{a}{a+b}$

Je spezifischer also die Vorannahmen sind, desto mehr Einfluss gewinnen sie auf den Wert des MAP-Schätzers

Andererseits je informativer die Daten sind, desto weniger Einfluss haben die Vorannahmen (prior overwhelmed by data)

Beispiel

Beurteilt die Versuchsperson den Vergleichsreiz bei 30 Darbietungen insgesamt 20-mal als intensiver als den Standardreiz, dann ergeben sich in Abhängigkeit von den Hyperparametern a, b folgende Werte für den MAP-Schätzer



Bayessche Konfidenzintervalle

Vorsicht bei der Interpretation "frequentistischer" Konfidenzintervalle!

Das Konfidenzniveau $1 - \alpha$ kennzeichnet

- *nicht* die Wahrscheinlichkeit, mit der der Parameter in einem resultierenden Intervall liegt
- die Wahrscheinlichkeit, mit der das verwendete Verfahren (bei wiederholter Anwendung auf mögliche Datensätze) ein Intervall liefert, in dem der Parameter enthalten ist

Für Konfidenzintervalle, die im Rahmen der Bayesschen Statistik abgeleitet werden, kann das Konfidenzniveau tatsächlich als die Wahrscheinlichkeit interpretiert werden, mit der der Parameter im resultierenden Intervall liegt

Konstruktion Bayesscher Konfidenzintervalle

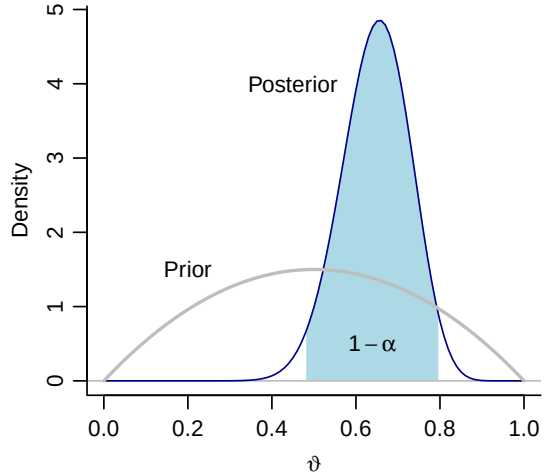
Die Konstruktion eines Bayesschen Konfidenzintervalls für den Parameter ϑ bezieht sich auf die A-posteriori-Verteilung bzw. deren Dichtefunktion $p(\vartheta \mid y_1, \ldots, y_n)$

Für ein zweiseitiges Konfidenzintervall zum Niveau $1 - \alpha$ bestimmt man

- die untere Grenze g_u des Intervalls als $(\alpha/2)$-Quantil der A-posteriori-Verteilung, d. h.
$$\int_{-\infty}^{g_u} p(\vartheta \mid y_1, \ldots, y_n) d\vartheta = \frac{\alpha}{2}$$
- die obere Grenze g_o des Intervalls als $(1 - \alpha/2)$-Quantil der A-posteriori-Verteilung, d. h.
$$\int_{-\infty}^{g_o} p(\vartheta \mid y_1, \ldots, y_n) d\vartheta = 1 - \frac{\alpha}{2}$$

Beispiel

Für die oben besprochene Anwendung der Konstanzmethode (20 Ja-Antworten, 30 Durchgänge) mit den Hyperparametern $a, b = 2.0$ der A-priori-Verteilung ergibt sich das 0.95-Konfidenzintervall für ϑ $[0.48, 0.80]$



Zusammenfassung

Die Bayessche Schätzung kennzeichnet die Unsicherheit über den Wert eines Parameters durch entsprechende Verteilungen

- Die A-priori-Verteilung ermöglicht die Einbeziehung von Vorwissen bzw. subjektiven Einschätzungen in die Parameterschätzung
- Die A-posteriori-Verteilung kennzeichnet die Unsicherheit nach Auswertung der in den Daten verfügbaren Information und bildet die Grundlage für die Parameterschätzung

MAP-Schätzung (Maximum-a-posteriori) ist im Grenzfall fehlenden Vorwissens äquivalent zur ML-Schätzung

Verfahren wird in der psychologischen Forschung zunehmend eingesetzt (z. B. Testtheorie, Wahrnehmungspsychologie), insbesondere auch im Rahmen einer rekursiven Anwendung (adaptive Verfahren)