



HOCHSCHULE FÜR TECHNIK UND WIRTSCHAFT

DOKUMENTATION

Projektseminar
Optimierung und
Unsicherheitsquantifizierung mit
Bayesianischer Statistik und
MCMC-Methoden
(Prof. Schwarzenberger)

Clemens Näther, s85426

Jakub Kliemann, s85515

Contents

1	Einleitung	2
2	Theoretischer Teil	3
2.1	Grundlagen der bayesianischen Statistik und das Bayes'sche Theorem .	3
2.1.1	Einführung in die bayesianische Statistik	3
2.1.2	Das Bayes'sche Theorem und seine Bestandteile	3
2.1.3	Beispiele und praktische Anwendungen	5
2.1.4	Punktschätzer in der bayesianischen Statistik	6
2.2	Binomiale Verteilung und deren bayesianische Interpretation	7
2.3	Markov Chain Monte Carlo (MCMC) Methoden	8
2.3.1	Einführung in MCMC-Methoden	8
2.4	Konvergenzkriterien und Diagnosewerkzeuge für MCMC-Simulationen .	9
3	Praktischer Teil	10
3.1	Implementierung bayesianischer Modelle unter Verwendung in Python .	10
3.2	Anwendung der Modelle auf verschiedene Datensätze	11
3.3	Durchführung von MCMC-Simulationen	12
3.4	Interpretation der Ergebnisse	13
3.5	Vergleich mit klassischen Methoden	14
4	Zusammenfassung und Ausblick	15
5	Literaturverzeichnis	16
6	Selbstständigkeitserklärung	17

1 Einleitung

2 Theoretischer Teil

2.1 Grundlagen der bayesianischen Statistik und das Bayes'sche Theorem

2.1.1 Einführung in die bayesianische Statistik

Die bayesianische Statistik ist ein Zweig der Statistik. Sie unterscheidet sich im wesentlichen in der Interpretation der Wahrscheinlichkeit von der klassischen Statistik. Die klassische Statistik definiert die Wahrscheinlichkeit als die **relative Häufigkeit** in einem Zufallsexperiment [3, p. 2]. In der bayesianischen Statistik hingegen wird die Wahrscheinlichkeit als Grad des Glaubens respektiv als **Plausibilität** eines Ereignisses oder einer Aussage interpretiert [1, p. 1].

Kern der bayesianischen Statistik ist es Wissen über ein Ereignis zu verfeinern, sobald neue Informationen vorliegen. Dazu nutzt man hauptsächlich das **Bayes'sche Theorem**, welches erlaubt das Vorwissen (Prior) mit neuen Daten (Likelihood) zu kombinieren und daraus eine aktualisierte Wahrscheinlichkeit (Posterior) zu berechnen.

Mit Hilfe des Bayes'schen Theorems kann man unbekannte Parameter schätzen, ein Konfidenzintervall für diese Parameter angeben und Hypothesen prüfen. Die klassische Statistik benötigt hingegen dafür Testgrößen, weshalb die bayesianische Statistik als flexibler und intuitiver gilt. [1, p. 1].

Problem der bayesianischen Statistik ist jedoch, dass die Berechnung der Posterioriverteilung analytisch oft nicht möglich ist. Da es nun aber gute numerische Methoden wie die **Markov Chain Monte Carlo (MCMC)** Methoden gibt, findet die bayesianische Statistik immer mehr Anwendungen. So zum Beispiel in der Medizin oder für künstliche Intelligenzen. [3, p. 1].

2.1.2 Das Bayes'sche Theorem und seine Bestandteile

Das Bayes'sche Theorem ist ein fundamentales Konzept der bayesianischen Statistik. Es beschreibt, wie man vorhandenes Vorwissen durch neue Daten aktualisiert.

Die **Prioriverteilung** beschreibt die anfänglichen Annahmen oder das Vorwissen über einen Parameter oder ein Ereignis, bevor neue Daten berücksichtigt werden. Dabei "enthält die Prioriverteilung eines Parameters θ , ausgedrückt durch $f(\theta)$, was man vor Auswertung der Stichprobe über θ weiß." [3, p. 90].

Als Priori-Wahrscheinlichkeit wird somit die Wahrscheinlichkeit $P(A)$ bezeichnet.

Die **Posterioriverteilung** beschreibt das Wissen über einen Parameter oder ein Ereignis, nachdem alle vorhandenen Daten berücksichtigt wurden. Durch die neuen Daten, meist einer Stichprobe, wird die anfängliche Annahme, die durch die Prioriverteilung ausgedrückt wird, aktualisiert. Dies führt zu einer neuen Verteilung die widerspiegelt, wie wahrscheinlich verschiedene Werte des Parameters auf Grundlage sowohl des Vorwissens als auch der neuen Informationen sind. [3, p. 109]

Die Posteriori-Wahrscheinlichkeit wird somit als $P(A|B)$ bezeichnet.

Die **Likelihood-Funktion** enthält die Informationen, die die Daten über den Parameter oder das Ereignis liefern. Dabei beschreibt die Likelihood die Informationen aus den neuen Daten, die zur Aktualisierung der Prioriverteilung beitragen. [3, p. 88] Die Likelihood-Wahrscheinlichkeit wird somit als $P(B|A)$ bezeichnet.

Die Wahrscheinlichkeit $P(B)$ wird als Normierungskonstante bezeichnet. Sie sorgt dafür, dass die Posterioriverteilung korrekt normiert ist, das heißt, dass die Summe der Wahrscheinlichkeiten aller möglichen Werte des Parameters 1 ergibt. [3, p. 109]

Das Bayes'sche Theorem lässt sich somit wie folgt darstellen:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)} \quad (1)$$

Das Bayes'sche Theorem lässt sich auch rekursiv anwenden [1, p. 17].

Gegeben sei das Ereignis A sowie die Teilergebnisse B_1, B_2, \dots, B_n . Dann ergibt sich die Wahrscheinlichkeit $P(A|B_1)$ zu:

$$P(A|B_1) = \frac{P(A) \cdot P(B_1|A)}{P(B_1)} \quad (2)$$

Nun wird die Information B_2 hinzugefügt. Die Wahrscheinlichkeit $P(A|B_1, B_2)$ ergibt sich bei Unabhängigkeit von den Teilereignissen B_1, B_2, \dots, B_n zu:

$$P(A|B_1, B_2) = \frac{P(A) \cdot P(B_1|A) \cdot P(B_2|A)}{P(B_1) \cdot P(B_2)} \quad (3)$$

Weiterhin lässt sich diese Formel umstellen, wodurch deutlich wird, dass beim Hinzufügen von neuen Informationen die Posterioriverteilung aktualisiert wird:

$$P(A|B_1, B_2) = \frac{P(A) \cdot P(B_1|A) \cdot P(B_2|A)}{P(B_1) \cdot P(B_2)} = P(A|B_1) \cdot \frac{P(B_2|A)}{P(B_2)} \quad (4)$$

Dies lässt sich allgemein formulieren für:

$$P(A|B_1, B_2, \dots, B_n) = P(A|B_1, B_2, \dots, B_{n-1}) \cdot \frac{P(B_n|A)}{P(B_n)} \quad (5)$$

Die Wahl der Prioriverteilung ist ein wichtiger Aspekt der bayesianischen Statistik. Sie wird immer so gewählt, dass die Entropie maximal ist. Die Entropie ist ein Maß für die Unsicherheit, was bedeutet, dass nur Informationen enthalten sind, die vor der Beobachtung bekannt sind. [1, p. 57]. Unter folgenden Bedingungen ist die Prioriverteilung optimal [1, p. 59]:

- Zufallsvariablen, die in $[a, b]$ definiert sind, sind **gleichverteilt**
- Zufallsvariablen mit gegebenen Mittelwert und Varianz sind **normalverteilt**
- Zufallsvariablen mit gegebenem Mittelwert sind **exponentialverteilt**
- Zufallsvariablen mit gegebenem Mittelwert und Varianz im Intervall $[0, \infty]$ besitzen eine **abgeschnittene Normalverteilung**

Wenn keine Informationen über den Parameter vorliegen, wird eine **uninformative** Prioriverteilung gewählt. Es handelt sich dabei um eine uneigentliche Verteilung. [1, p. 57].

2.1.3 Beispiele und praktische Anwendungen

Beispiel 1: m gleichgeformte Kugeln, unter denen sich k rote Kugeln und $m - k$ schwarze Kugeln befinden. Eine Kugel wird zufällig gezogen. Die Wahrscheinlichkeit, dass die gezogene Kugel rot ist, beträgt

$$P(A) = \frac{k}{m} = p \quad (6)$$

Der Versuch wird erweitert, sodass n -mal eine Kugel mit Zurücklegen gezogen wird. Die Wahrscheinlichkeit, dass x -mal eine rote Kugel bei n -maligem Ziehen gezogen wird, beträgt

$$P(x|n, p) = \binom{n}{x} \cdot p^x \cdot (1 - p)^{n-x} \quad (7)$$

Sei nun p unbekannt. Dieses p ist nun zu schätzen. Die Binomialverteilung wird nun als Likelihood-Funktion verwendet:

$$P(n, x|p) = \binom{n}{x} \cdot p^x \cdot (1 - p)^{n-x} \quad (8)$$

wobei $0 \leq p \leq 1$. Als Prioridichte wird die Gleichverteilung verwendet, da es keine Informationen über p gibt.

$$P(p) = \begin{cases} 1, & \text{für } 0 \leq p \leq 1 \\ 0, & \text{sonst} \end{cases} \quad (9)$$

Die Posterioridichte ergibt sich somit zu:

$$P(p|n, x) = \frac{\binom{n}{x} p^x \cdot (1 - p)^{n-x}}{P(n, x)} \quad (10)$$

Vergleicht man dies mit der Dichtefunktion der Beta-Verteilung, so erkennt man dass die Posterioridichte einer Beta-Verteilung entspricht.

$$P(p|n, x) = \frac{(n+1)!}{x! \cdot (n-x)!} \cdot p^x \cdot (1-p)^{n-x} = \frac{\Gamma(n+1)}{\Gamma(x+1) \cdot \Gamma(n-x+1)} \cdot p^x \cdot (1-p)^{n-x} \quad (11)$$

Somit suche nach Maximum der Posterioridichte, um den Schätzer für p zu finden.

$$\frac{d}{dp} P(p|n, x) = x p^{x-1} \cdot (1-p)^{n-x} - (n-x) p^x \cdot (1-p)^{n-x-1} = 0 \quad (12)$$

$$\Rightarrow x p^{x-1} \cdot (1-p)^{n-x} = (n-x) p^x \cdot (1-p)^{n-x-1} \quad (13)$$

Vereinfacht ergibt sich:

$$\Rightarrow x(1-p) = (n-x)p \quad (14)$$

$$\Rightarrow \frac{x}{p} - \frac{n-x}{1-p} = 0 \quad (15)$$

$$\Rightarrow p = \frac{x}{n} \quad (16)$$

Der Schätzer für p ist somit der relative Anteil der roten Kugeln an der Gesamtanzahl der Kugeln.

Beispiel 2: Beispiel 1 wird erweitert. Es wird nun eine zweite Stichprobe gezogen. Stichprobe 1: $n_1 = 10$, $x_1 = 4$, Stichprobe 2: $n_2 = 20$, $x_2 = 6$. Daten sind unabhängig voneinander. Die Daten (Posterioriverteilung) der 1. Stichprobe dienen nun als Prioridichte für die 2. Stichprobe. Man erhält somit:

$$P(p|n_1, x_1, n_2, x_2) = \frac{P(p|n_1, x_1) \cdot P(p|n_2, x_2)}{P(n_1, x_1, n_2, x_2)} \quad (17)$$

Dabei ist die Prioridichte identisch zur Posterioridichte der 1. Stichprobe, siehe Gleichung (10). Die Posterioridichte der 2. Stichprobe ergibt sich somit zu:

$$P(p|n_1, x_1, n_2, x_2) = \frac{\binom{n_1+n_2}{x_1+x_2} \cdot p^{x_1} \cdot (1-p)^{n_1-x_1} \cdot p^{x_2} \cdot (1-p)^{n_2-x_2}}{P(n_1, x_1, n_2, x_2)} \quad (18)$$

Oder mithilfe der Beta-Verteilung:

$$P(p|n_1, x_1, n_2, x_2) = \frac{\Gamma(n_1 + n_2 + 1)}{\Gamma(x_1 + x_2 + 1) \cdot \Gamma(n_1 + n_2 - x_1 - x_2 + 1)} \cdot p^{x_1+x_2} \cdot (1-p)^{n_1+n_2-x_1-x_2} \quad (19)$$

Die Daten der 1. und 2. Stichprobe können somit kombiniert werden. Für die Daten $n_1 = 10$, $x_1 = 4$, $n_2 = 20$, $x_2 = 6$:

$$P(p|10, 4, 20, 6) = 931395465p^{10} \cdot (1-p)^{20} \quad (20)$$

2.1.4 Punktschätzer in der bayesianischen Statistik

Im folgenden wird die Schätzung eines Parameters mithilfe der Bayes-Strategie erläutert. Die möglichen Schätzwerte der Parameter x werden als \hat{x} bezeichnet. Die wahren Parameter werden als x bezeichnet.

Es wird eine Kostenfunktion $L(\hat{x}, x)$ definiert, die die Kosten für die Schätzung \hat{x} des wahren Parameters x angibt. Dies bedeutet, dass die Kostenfunktion die Differenz zwischen dem wahren Parameter x und der Schätzung \hat{x} angibt. Dabei gibt es verschiedene Kostenfunktionen, die verwendet werden können. [1, p. 65]

Die **quadratische Kostenfunktion** ist definiert als: $L(x - \hat{x}) = (x - \hat{x})\Sigma^{-1}(x - \hat{x})$. Diese gibt den quadratischen Abstand zwischen dem wahren Parameter x und der Schätzung \hat{x} an. Die zu erwartenden Kosten werden berechnet mit dem Erwartungswert der Kostenfunktion. Diese Schätzung führt zu dem Erwartungswert von x , das heißt $\hat{x} = E(x)$. [1, pp. 65–66]

Die **Kostenfunktion der absoluten Fehler** ist definiert als: $L(x, \hat{x}) = |x - \hat{x}|$. Diese gibt den absoluten Abstand zwischen x und \hat{x} an. Die Schätzung mit dem absoluten Fehler ergibt den Median der Verteilung, das heißt $F(\hat{x}_{med}) = 0.5$ [1, pp. 67–68]

Die **Null-Eins-Kostenfunktion** bedeutet, dass es entweder Kosten oder keine Kosten gibt. Diese ist definiert durch: $L(x - \hat{x}) = \begin{cases} 0 & \text{für } |x - \hat{x}| < b \\ a & \text{für } |x - \hat{x}| \geq b \end{cases}$

wobei a und b als Konstanten angenommen werden. Wenn der Fall $b \rightarrow 0$ betrachtet wird, ergibt sich als Schätzer das Argument des Maximums der Posterioriverteilung, das heißt $\hat{x}_M = \arg \max p(x|y)$. [1, pp. 68–69]

2.2 Binomiale Verteilung und deren bayesianische Interpretation

2.3 Markov Chain Monte Carlo (MCMC) Methoden

2.3.1 Einführung in MCMC-Methoden

Bei einer direkten Simulation wird vorausgesetzt, dass die Verteilung der Zufallsvariablen bekannt ist. Dies ist jedoch in der Praxis nicht immer gegeben. Die Berechnung der Posterioriverteilung ist analytisch oft nicht möglich, vor allem bei komplexen Modellen oder hohen Dimensionen.

Die Markov Chain Monte Carlo (MCMC) Methoden sind eine Klasse von Algorithmen, die es ermöglichen, eine Stichprobe aus einer Verteilung zu ziehen, ohne die Verteilung zu kennen. [2, p. 179]

Diese Methoden verwenden zwei Konzepte: Markov-Ketten und Monte Carlo-Methoden. Eine Markov-Kette ist eine Folge von Zufallsvariablen, die die Markov-Eigenschaft erfüllen. Die Markov-Eigenschaft sagt aus, dass die nächste Zufallsvariable nicht von den vorherigen Zufallsvariablen, sondern nur von der letzten Zufallsvariable abhängt. Das bedeutet, dass die Wahrscheinlichkeit, im nächsten Zustand $X_n + 1$ zu landen, nur von X_n abhängt. Die Übergangswahrscheinlichkeit zwischen den Zuständen kann in einer Übergangsmatrix dargestellt werden. [2, 188f.]

Die Monte Carlo-Methoden sind eine Gruppe von Algorithmen, die es ermöglichen, Zufallsvariablen zu schätzen, indem Zufallszahlen generiert werden. Sie erzeugen zufällige Stichproben, um eine Näherung der Verteilung zu erhalten. [2, 14f.]

Die MCMC-Methoden nutzen die Monte Carlo-Methoden, um eine Markov-Kette zu simulieren. Diese Technik ist besonders nützlich, um eine Posterioriverteilung zu schätzen, wenn direkte Berechnungen nicht möglich sind. [2, p. 179]

Im folgenden wird der Metropolis-Hastings-Algorithmus erläutert, der eine der bekanntesten MCMC-Methoden ist.

2.4 Konvergenzkriterien und Diagnosewerkzeuge für MCMC-Simulationen

3 Praktischer Teil

3.1 Implementierung bayesianischer Modelle unter Verwendung in Python

3.2 Anwendung der Modelle auf verschiedene Datensätze

3.3 Durchführung von MCMC-Simulationen

3.4 Interpretation der Ergebnisse

3.5 Vergleich mit klassischen Methoden

4 Zusammenfassung und Ausblick

5 Literaturverzeichnis

References

- [1] Karl-Rudolf Koch. *Einführung in die Bayes-Statistik*. Berlin [u.a.]: Springer, 2000. ISBN: 3540666702. URL: <https://katalog.slub-dresden.de/id/0-306244284>.
- [2] Thomas Müller-Gronbach, Erich Novak, and Klaus Ritter. *Monte Carlo-Algorithmen*. Berlin: Springer, 2012. ISBN: 9783540891406. URL: <https://katalog.slub-dresden.de/id/0-618339728>.
- [3] Wolfgang Tschirk. *Statistik: Klassisch oder Bayes zwei Wege im Vergleich*. Berlin , , © 2014. ISBN: 3642543847. URL: <https://katalog.slub-dresden.de/id/0-160866449X>.

6 Selbstständigkeitserklärung