Das Bayes'sche Prinzip

Olivia Gradenwitz
Patrik Kneubühler
Seminar über Bayes Statistik FS08

26. Februar 2008

1 Bayes'sches statistisches Modell

1.1 Statistische Probleme und statistische Modelle

In diesem Seminar konzentrieren wir uns auf die schliessende Statistik. Die schliessende Statistik verwendet die Realisierungen von Zufallsvariablem um damit Rückschlüsse auf die entsprechenden Verteilungen zu machen. Die unbekannte Verteilung ist bekannt bis auf einen Parameter θ.

Definition 1.1. *Ein parametrisches statistisches Modell besteht aus einer Realisierung einer Zufallsariablen x, mit Verteilungsdichte $f(x|\theta)$, wobei θ unbekannt ist und zu einem Vektorraum Θ gehöhrt. Typischerweise gilt $\Theta \subset \mathbb{R}^p$.*

Bemerkung. Wir gehen immer davon aus, dass die betrachteten Verteilungen absolutstetig sind in Bezug auf ein fixes Referenzmass μ auf dem Stichprobenraum. Ist der Stichprobenraum gleich $\mathbb{R}^n$, so wählen wir für μ meist das Lebesguemass, ist der Stichprobenraum hingegen diskret, so wählen wir das Zählmass.

1.2 Bayes'sches statistisches Modell

Wir nehmen nun an, dass ein parametrisches Modell gegeben ist und möchten auf Grund von vorliegenden Daten/Experimenten Informationen über den Parameter erhalten. Die einfachste Möglichkeit dafür ist die Likelihood Funktion.

Definition 1.2. *Sei ein parametrisches statistisches Modell mit Dichten $f(x|\theta)$ gegeben. Die Likelihood Funktion ist definiert als*

$$l(\theta|x) := f(x|\theta).$$

Intuitiv könnte man die Likelihood Funktion $l(\theta|x)$ als die Wahrscheinlichkeit interpretieren, dass θ der wahre Parameter ist, falls man die Daten x kennt. Im allgemeinen ist jedoch $l(\theta, x)$ nicht integrabel, geschweige denn $\int l(\theta, x)d\theta = 1$.

Die Basis für das Bayes'sche Modell ist die Bayes Formel, die wir in der ersten Stunde bereits kennen gelernt haben. Die stetige Version davon lautet wie folgt:

Lemma 1. Seien x, y Zufallsvariablen, wobei die Randverteilung von y durch $g(y)$ und die Verteilung von x gegeben y durch $f(x|y)$ beschrieben sei. Dann hat die bedingte Verteilung von y gegeben x die Dichte

$$g(y|x) = \frac{f(x|y)g(y)}{\int f(x|y)g(y)dy}.$$

Definition 1.3. *Ein Bayes'sches statistisches Modell ist ein parametrisches Modell $f(x|\theta)$ mit einer Wahrscheinlichkeitsverteilung auf dem Parameterraum, die wir a priori Verteilung nennen. Wir nehmen an, dass die a priori Verteilung absolutstetig ist mit Dichte $\pi(\theta)$.*

1.3 A Posteriori Verteilung

Sei ein Bayes'sches statistisches Modell $f(x|\theta)$ mit einer a priori Verteilung $\pi(\theta)$ gegeben. Die a posteriori Verteilung erhalten wir durch Einsetzen von π an der Stelle von g in Lemma 1:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta}.$$

Bemerkung.

1. Beachte, dass $\pi(\theta|x)$ proportional zur Likelihood Funktion multipliziert mit der a priori Verteilung ist.
2. Mit der a posteriori Verteilung kann man Punktschätzer, Prognosen, Vertrauensintervalle und allgemeiner statistische Verfahren definieren. Darauf möchten wir im nächten Abschnitt genauer eingehen.

1.4 Bayes'sches Prinzip

Das Bayes'sche Prinzip ist das Arbeiten mit der a posteriori Verteilung, insbesondere das Definieren statistischer Verfahren, die nur von der a posteriori Verteilung abhängen. Wir geben in den folgenden Paragraphen eine kurze Erklärung von Prognosen und vom Bayes Verfahren. Letzeres wird im Vortrag über Entscheidungstheorie ausführlicher behandelt.

Prognosen Mit der a posteriori Verteilung können wir Prognosen berechnen. Sei also eine Zufallsvariable y mit der Verteilung $g(y|\theta, x)$ gegeben. Dann gilt

$$g(y|x) = \int g(y|\theta, x)\pi(\theta|x)d\theta.$$

Weil die Prognose $g(y|x)$ nur von der a posteriori Verteilung abhängt, ist sie ein Beispiel für das Bayes'sche Prinzip.

Bayes Verfahren Sei $\mathbb{X} = \mathbb{R}^n$ ein Stichprobenraum, $\mathbb{A}$ ein Aktionsraum und $d : \mathbb{X} \to \mathbb{A}$ ein statistisches Verfahren. Wir definieren eine Verlustfunktion $L : \Theta \times \mathbb{A} \to \mathbb{R}$, das Risiko $R(\theta, d) := \int_{\mathbb{X}} L(\theta, d(x)) f(x|\theta) dx$ des Verfahrens bei wahrem Parameter θ, und schliesslich das über alle θ gemittelte Risiko $r(\pi, \theta) = \int_{\Theta} R(\theta, d)\pi(d\theta)$. Ein Bayes Verfahren ist nun ein Abbildung d, welche das gemittelte Risiko r, auch Bayes Risiko genannt, minimiert. Dank Satz 2.1 aus [2] gilt dann

$$d(x) = \underset{a}{\operatorname{argmin}}\ E[L(\theta, a)|x],$$

falls das Minimum existiert. Damit ist das Bayes Verfahren nur von der a posteriori Verteilung abhängig.
Setzt man $\mathbb{A} = \Theta = \mathbb{R}$, so ist das Verfahren d ein Punktschätzer. Man kann zeigen, dass für quadratischen Verlust $L(\theta, a) = (\theta - a)^2$ der Punktschätzer $d(x) = E[\theta|x]$, auch Posterior Mean genannt, resultiert.

Das folgende Beispiel soll die Idee der a posteriori Verteilung und der Prognose konkretsieren.

Beispiel 1. Wir betrachten das Interval $[0, 1]$ und wählen in einem ersten Schritt mit einer a priori Verteilung π einen Punkt darin aus, den wir θ nennen. Weil wir keine weiteren Informationen haben, nehmen wir die Gleichverteilung als a priori Verteilung an. In einem zweiten Schritt wählen wir, ebenfalls mit der Gleichverteilung, zufällig n Punkte in $[0, 1]$ aus. Die Anzahl Punkte X links von θ ist dann $Bin(n, \theta)$-verteilt. Wir erhalten also mit der a priori Verteilungsdichte $\pi(\theta) \equiv 1$

$$\begin{aligned} f(x|\theta) &:= P(X = x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ \pi(\theta|x) &= \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x}}{\int_0^1 \binom{n}{x}\theta^x(1-\theta)^{n-x}d\theta} = \frac{\theta^x(1-\theta)^{n-x}}{Beta(x+1, n-x+1)}, \end{aligned}$$

wobei wir folgende Identität benutzt haben

$$\int_0^1 t^l(1-t)^m dt = \frac{\Gamma(l+1)\Gamma(m+1)}{\Gamma(l+m+2)} = Beta(l+1, m+1).$$

Wir gehen davon aus, dass wir nicht θ, aber X kennen. Wir wählen nun zusätzlich einen weiteren Punkt p in $[0, 1]$ wie oben und setzen $Y \in \{0, 1\}$ genau dann eins, wenn p links von θ liegt. Eine Prognose für Y kann nun aus der a posteriori

Verteilung wie folgt bestimmt werden:

$$\begin{aligned} g(y|\theta,x) &:= P(Y=y|\theta,X=x) = \theta 1_{\{y=1\}} + (1-\theta)1_{\{y=0\}} \\ P(Y=1|X=x) &= \int_0^1 g(y|\theta,x)\pi(\theta|x)d\theta = \int_0^1 \theta \frac{\theta^x(1-\theta)^{n-x}}{\int_0^1 \theta^x(1-\theta)^{n-x}d\theta}d\theta \\ &= \frac{\int_0^1 \theta^{x+1}(1-\theta)^{n-x}}{\int_0^1 \theta^x(1-\theta)^{n-x}d\theta}d\theta = \frac{x+1}{n+2} \\ &= \left(1-\frac{n}{n+2}\right)\cdot\frac{1}{2} + \frac{n}{n+2}\cdot\frac{x}{n} \\ P(Y=0|X=x) &= \frac{n-x+1}{n+2} = \left(1-\frac{n}{n+2}\right)\cdot\frac{1}{2} + \frac{n}{n+2}\cdot\frac{n-x}{n}, \end{aligned}$$

folglich erhalten wir die Prognose

$$g(y|x) := P(Y|X=x) = \frac{x+1}{n+2}1_{\{y=1\}} + \frac{n-x+1}{n+2}1_{\{y=0\}}.$$

Die Wahrscheinlichkeit für das Ereignis $\{Y=1\}$ ist eine konvexe Kombination der a priori Wahrscheinlichkeit $\frac{1}{2}$ und der bisherigen Erfolgsquote $\frac{k}{n}$, die wir aus den Daten X erhalten haben. Dabei wird bei wachsendem n die empirische Erfolgsquote immer stärker gewichtet.

2 Likelihood-Prinzip

2.1 Definition der Prinzipien

Definition 2.1. *Sei $x \sim f(x|\theta)$. Eine Funktion T von x (auch Statistik genannt) heisst suffizient für das Modell $f(x|\theta)$, wenn die Verteilung von x bedingt auf $T(x)$ unabhängig von θ ist.*

Das Faktorisierungslemma (siehe [2]) sagt aus, dass eine Statistik genau dann suffizient ist, wenn sie wie folgt zerlegt werden kann

$$f(x|\theta) = g(T(x)|\theta)h(x|T(x)). \tag{1}$$

Das folgende Prinzip ist wichtig:

Prinzip 2.2 (Suffizienz)**.** *Für zwei Beobachtungen x und y mit $T(x) = T(y)$, müssen die Rückschlüsse auf θ gleich sein.*

Korollar 2.3. *Das Bayes-Verfahren erfüllt das Suffizienz-Prinzip.*

Beweis. Sei eine suffiziente Statistik T gegeben. Wegen dem Faktorisierungslemma hängt die a posteriori Verteilung nur durch $T(x)$ von den Daten x ab

$$\pi(\theta|x) \propto \pi(\theta)f(x|\theta) = \pi(\theta)g(T(x)|\theta)h(x) \propto \pi(\theta)g(T(x)|\theta). \tag{2}$$

Weil das Bayes-Verfahren nur von der a posteriori Verteilung abhängt, folgt damit die Behauptung.

Während dieses Prinzip generell anerkannt wird, weil nach dem Theorem von Rao-Blackwell zulässige Verfahren bei konvexen Verlustfunktionen nur von suffizienten Statistiken abhängen (siehe [2]), wird das folgende Prinzip in der frequentistischen Statistik manchmal verletzt.

Prinzip 2.4 (Likelihood)**.** *Die Information einer Beobachtung x über θ ist vollständig in der Likelihood-Funktion $l(\theta|x)$ enthalten. Ausserdem, falls x_1 und x_2 zwei Beobachtungen sind, die vom gleichen Parameter θ abhängen, so dass eine Konstante c existiert, mit*

$$l_1(\theta|x)_1 = cl_2(\theta|x_2)$$

für jedes θ, dann enthalten sie die gleichen Informationen über θ und sollten deshalb die gleichen Rückschlüsse bewirken.

Das nächste Prinzip ist das letzte was wir vorstellen müssen, um unser Hauptresultat präsentieren zu können.

Prinzip 2.5 (Conditionality)**.** *Wenn zwei Experimente ϵ_1 and ϵ_2, für einen Parameter θ, zu Verfügung stehen und wenn eines der Experimente mit Wahrscheinlichkeit p gewählt wird, dann sollte der resultierende Rückschluss auf θ nur vom gewähltem Experiment abhängen.*

2.2 Äquivalenz des Likelihood-Prinzips zum Suffizienz- und Conditionality-Prinzip

Theorem 2.6. *Das Likelihood-Prinzip ist äquivalent zum Suffizienz- plus Conditionality-Prinzip.*

Aus dem generell akzeptierten Suffizienz-Prinzip und dem vertretbaren Conditionality-Prinzip folgt also das umstrittene Likelihood-Prinzip. Die Annahme des Likelihood-Prinzips wird dadurch gerechtfertigt.

2.3 Maximum Likelihood Estimate

Gegeben ein statistisches Modell $f(x, \theta)$ wir möchten einen Punktschätzer für den Wert θ konstruieren, der das Likelihood-Prinzip erfüllt. Wir wissen bereits, dass alle Bayes Punktschätzer, zum Beispiel der Posterior Mean, eine Lösung zu diesem Problem sind. Eine natürlichere und intuitivere Implementation des Likelihood-Prinzips ist aber auch der folgende Punktschätzer.

Definition 2.7. *Sei $f(x, \theta)$ ein statistisches Modell. Wir erinnern an die Likelihoodfunktion $l(\theta, x) := f(x, \theta)$ und definieren den Punktschätzer als*

$$\widehat{\theta} := \operatorname*{argmax}_{\theta} l(\theta|x).$$

Offensichtlich hängt dieser Schätzer nur von der Likelihood Funktion ab.

Beispiel 2. Sei $x = (x_1, \ldots, x_n)$ eine Familie von i.i.d. Zufallsvariablen mit der gemeinsamen Dichte

$$f(x|\theta) = \frac{1}{(2\pi)^n} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2}.$$

Anstatt $l(\theta|x)$ zu maximieren, maximieren wir durch Ableiten und Null setzen die Funktion

$$\log l(\theta|x) = n \log(\frac{1}{2\pi}) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2.$$

Wir erhalten als Schätzer den Mittelwert $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Die Maximum Likelihood Methode wird oft verwendent, weil sie sehr intuitiv ist und gute asymptotische Eigenschaften hat. Die Methode hat jedoch keinen entscheidungstheoretischen Hintergrund und kann oft nur numerisch (z.B. mit dem EM-Algorithmus) berechnet werden.

Literatur

[1] *The Bayesian Choice* Robert, Springer NY 2007

[2] *Mathematische Statistik* Künsch, October 2007