Stochastische Lernalgorithmen

Gerhard Jäger

23. April 2003

Wahrscheinlichkeitsrechnung (Forts.)

Gängige Wahrscheinlichkeitsfunktionen

Bernoulli-Verteilung

Besitzt ein Zufallsexperiment nur zwei mögliche Ergebnisse, so spricht man von einem Bernoulli-Experiment (Jakob Bernoulli 1654-1705). Diese werden i. A. mit Zufallsvariablen beschrieben, die die Werte 0 oder 1 annehmen können. Wenn p die Wahrscheinlichkeit für das Ergebnis 1 ist, dann ist die Verteilungsfunktion durch den Parameter p schon eindeutig bestimmt. Diese heißt Bernoulliverteilung (Schreibweise $(b(1,p))$.
Die Wahrscheinlichkeitsfunktion ist gegeben durch

$$b(1;1,p) = p; b(0;1,p) = 1-p$$

Erwartungswert:

$$E(b(1,p) = p$$

Varianz:

$$Var(b(1,p)) = p(1-p)$$

Binomial-Verteilung

Wir betrachten den n-fachen Münzwurf. Die Wahrscheinlichkeit, genau k-mal "Zahl" zu werfen, ist nicht gleichmäßig verteilt. Die hier vorliegende Wahrscheinlichkeitsverteilung heißt *Binomialverteilung.*
Bezeichnung: $b(n,p)$
Dabei gibt der Parameter n die Gesamtzahl der Versuche an und p die Wahrscheinlichkeit des Ereignisses, das k-mal auftreten soll.
Beispiel: Beim zweifachen Münzwurf (also $n = 2$) einer fairen Münze ist die Wahrscheinlichkeit $P(k)$ für k-mal "Zahl":

$$\begin{aligned} P(k=0) &= 0.25 \\ P(k=1) &= 0.5 \\ P(K=2) &= 0.25 \end{aligned}$$

Die Ungleichverteilung ergibt sich daraus, dass es zwei Möglichkeiten gibt, die Summe "1" zu erzielen, aber nur jeweils eine für "0" und "2".
Als Wahrscheinlichkeitsfunktion ausgedrückt, ergibt sich also

$$\begin{aligned} b(0;2,0.5) &= 0.25 \\ b(1;2,0.5) &= 0.5 \\ b(2;2,0.5) &= 0.25 \end{aligned}$$

Allgemein gesprochen ergibt sich die Binomialverteilung $b(n,p)$ aus der n-fachen stochastisch unabhängigen Wiederholung eines Bernoulli-Experiments mit Verteilung $b(1,p)$.
Was ist i.A. der Wert für $b(k;n,p)$?

- Die Wahrscheinlichkeit einer konkreten Folge aus k positiven und $n-k$ negativen Ausgängen ist $p^k(1-p)^{n-k}$ (da die einzelnen Experimente stochastisch unabhängig sind).
- Es gibt ingesamt $\left(\begin{array}{c} n \\ k \end{array} \right)$ verschiedene Folgen aus k Einsen und $n-k$ Nullen.
- Also gilt

$$b(k;n,p) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k}$$

Erwartungswert:

$$E(b(n,p)) = np$$

Varianz:

$$Var(b(n,p)) = np(1-p)$$

Normalverteilung

(Generell sehr wichtig für alle möglichen Anwendungen der Statistik; in der Computerlinguistk aber nicht so zentral).

Exkurs: Stetige Verteilungen: *Wenn der Wertebereich einer Zufallsvariablen überabzählbar ist (z.B. alle reellen Zahlen oder alle reellen Zahlen im Intervall [0,1]), ist es nicht möglich, jeder Zahl einen Wert größer 0 zuzuweisen, auch wenn alle Zahlen mögliche Werte der Variablen sind. In diesem Fall stellt man die Wahrscheinlichkeitsverteilung nicht durch eine Wahrscheinlichkeitsfunktion, sondern eine Wahrscheinlichkeits**dichte**funktion f_X dar. Die Fläche zwischen x-Achse und Funktionskurve im Intervall $[a,b]$—also formal das bestimmte Integral $\int_a^b f_X(x)dx$ entspricht der Wahrscheinlichkeit, dass $a \leq X \leq b$.*

Wenn sich eine Vielzahl unabhängiger zufälliger und ungerichteter Effekte summieren, ergibt sich eine *Normalverteilung*. Bei vielen statistischen Verfahren wird davon ausgegangen, dass die Daten die Werte normalverteilter Zufallswerte darstellen, oder kurz gesagt „normalverteilt sind".

$$N(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Es gilt:

Erwartungswert:

$$E(N(\mu,\sigma)) = \mu$$

Varianz:

$$Var(N(\mu,\sigma)) = \sigma^2$$

$\Rightarrow$ Für große Werte von n nähert sich die Binomialverteilung der Normalverteilung an; die Normalverteilung kann deshalb auch als Näherung für diskrete Zufallsvariablen verwendet werden.

Parameter-Schätzung

Ausgangspunkt der statistischen Analyse ist eine beobachtbare stochastische Größe X, deren Verteilungsfunktion $F(x|\theta)$ von unbekannten Parametern $\theta \in \Xi$ abhängt, wobei Ξ, der Parameterraum, alle möglichen Werte von θ beinhaltet.

$$\textit{Beispiel: } X \sim b(1,p), \quad \theta = p$$

Die unbekannte numerische Größe θ kann dabei interpretiert werden als Repräsentant der Unsicherheit des Modells. Um Information über θ zu erhalten, betrachtet man eine Beobachtung von X, die dadurch gewonnen wird, dass ein zu X gehöriges stochastisches Experiment durchgeführt und der erhaltene Zahlenwert X festgehalten wird.

Maximum-Likelihood-Schätzung

Maximum-Likelihood-Schätzer (ML-Schätzer, nach Fisher, 1890 - 1962) basieren auf der Annahme, dass wahrscheinlich das Wahrscheinlichste wahr ist. Diese Methode ist offensichtlich fehlbar. Trotzdem neigt man auch im Alltag gerne dazu, nach diesem Prinzip zu schätzen. Weiß man von einem Koch z.B. dass er wesentlich wahrscheinlicher eine Suppe versalzt, wenn er verliebt ist, als wenn er es nicht ist, so wird man sich während dem Essen einer versalzenen Suppe denken: „wahrscheinlich ist der Koch (mal wieder) verliebt". Genau das ist eine ML-Schätzung. Um ML-Schätzungen durchführen zu können, braucht man eine Likelihood-Funktion $L(\theta;\vec{x})$, die angibt, welcher Wert des gesuchten Parameters θ bei der vorgegebenen Stichprobe $\vec{x} = (x_1, x_2, \cdots, x_n)$ wie wahrscheinlich ist. Das Maximum

dieser Funktion in Abhängigkeit von θ bei gegebener Stichprobe $\vec{x}$ ist der Wert der ML-Schätzung. Um die Likelihood-Funktion zu erzeugen, entscheidet man sich aufgrund seiner Erfahrung (oder ähnlich überzeugender Argumente) dafür, dass die Stichprobe aus einem bestimmten Modell stammt. Das Modell kann dann Stichproben erzeugen. Die Likelihood-Funktion ist nun die Wahrscheinlichkeit für eine Realisation $\vec{x}$ in Abhängigkeit von den m Parametern $\vec{\theta} = \theta_1, \theta_2, \cdots, \theta_m$ des angenommenen Modells. Die Likelihood-Funktion braucht dann nur noch uminterpretiert zu werden. Man betrachtet nicht die Wahrscheinlichkeit für die x_i in Abhängigkeit von $\hat{\theta}$, sondern $\hat{\theta}$ als Variablen und die x_i als Parameter.

Beispiel 1:
10 Würfe einer Münze haben 8 mal Kopf und 2 mal Zahl ergeben. Aufgrund unseres Wissens über Münzfürfe gehen wir davon aus, dass die einzelnen Würfe stochastisch unabhängig voneinander sind. Die zugrundeliegende Wahrscheinlichkeitsverteilung wird also durch die Wahrscheinlichkeit p, bei einem Wurf Kopf zu werfen, vollständig festgelegt. Demnach gilt

$$\begin{aligned} \theta &= p \\ \Xi &= [0,1] \\ \Sigma_{i=1}^{10} x_i &= 8 \\ L(\theta; \vec{x}) &= p^8 \cdot (1-p)^2 \\ \arg_{\theta} \max L(\theta; \vec{x}) &= 0.8 \end{aligned}$$

Beispiel 2:

Ist der Löwe hungrig, müde oder ist ihm schlecht? *Der Einfachheit halber wollen wir uns vorstellen, dass ein Löwe nur in einem von drei Zuständen sein kann: Er kann hungrig sein, er kann müde sein oder ihm kann schlecht sein. Je nach Zustand pflegt er, eine unterschiedliche Gierigkeit im Verspeisen von Menschen in der Nacht an den Tag zu legen. Da die Menschenjagd stochastischen Einflüssen unterworfen ist, ergibt sich für jeden Gefräßigkeitszustand eine Verteilung für die in einer Nacht vertilgten Menschen. Diese Verteilungen lassen sich in der folgenden Tabelle ablesen:*

Zustand	*0*	*1*	*2*	*3*	*4*
hungrig	*0.00*	*0.05*	*0.05*	*0.80*	*0.10*
müde	*0.05*	*0.05*	*0.80*	*0.10*	*0.00*
magenkrank	*0.90*	*0.08*	*0.02*	*0.00*	*0.00*

Die Zeilen zeigen also die verschiedenen Verteilungen, die zu den drei verschiedene Zuständen gehören. Etwas formaler lässt sich die Tabelle so hinschreiben:

θ_i	*0*	*1*	*2*	*3*	*4*	$p(j \mid \theta_i)$
θ_1	*0.00*	*0.05*	*0.05*	*0.80*	*0.10*	$p(j \mid \theta_1)$
θ_2	*0.05*	*0.05*	*0.80*	*0.10*	*0.00*	$p(j \mid \theta_2)$
θ_3	*0.90*	*0.08*	*0.02*	*0.00*	*0.00*	$p(j \mid \theta_3)$

Jetzt stellen Sie sich vor, vor Ihnen taucht ein Löwe auf! Dann sehen Sie sich sofort der Frage gegenüber: In welchem Zustand befindet sich der Löwe? Sie können die Situation auch so beschreiben, dass zunächst zu klären ist, welche der drei Verteilungen der Tabelle zutreffend ist, oder wenn Sie die drei als eine Verteilung mit einem Löwenaktivitätsparameter ansehen: Welches wird der zutreffende Löwenaktivitätsparameter sein?
Jetzt zeigt sich wieder einmal, dass Entscheidungen ohne Informationshintergrund, sprich: Daten, unbefriedigend sein müssen. Woher soll man den Aktivitätszustandsparameter erahnen können? Zu Ihrem Glück sind Sie in Begleitung eines Löwenparkwächters. Und dieser Wächter teilt Ihnen (zu Ihrem Glück?) mit, dass der Löwe in der letzten Nacht keinen einzigen Menschen gefressen hat! Nun die Frage: Was meinen Sie, in welchem Aktivitätszustand sich der Löwe befindet?
Wenn Sie geantwortet haben: magenkrank! haben Sie offensichtlich das Maximum-Likelihood-Prinzip bereits verstanden und unbewusst angewendet. Sie müssen sich nur noch darüber klar werden, was Sie warum bei Ihrer Antwortfindung gemacht haben!

Der ML-Algorithmus im diskreten Fall *Die Beobachtung 0 verzehrte Menschen in der letzten Nacht führt zu der Betrachtung der ersten Spalte der Tabelle. Aus ihr entnimmt man, dass der Zustand hungrig auf jeden Fall nicht vorliegen kann. Also bleiben die Alternativen müde und magenkrank. Wäre der Löwe müde gewesen, hätte sich die Beobachtung mit 5 % Wahrscheinlichkeit realisiert. Im Gegensatz dazu führt eine Magenverstimmung in 90 % aller Fälle zu einer Diätnacht. Damit ist der dritte Zustand viel naheliegender, und wenn die Frage im Raum steht: Welchen Aktivitätszustand vermuten Sie? lautet die Antwort: Ich schätze der Löwe hatte eine Magenverstimmung!*
Algorithmus: Der Algorithmus zu dieser Antwort lässt sich schnell formulieren:

1. *Suche die Spalte, die zu der Beobachtung gehört.*
2. *Suche in der Spalte die Zeile mit dem maximale Wert.*
3. *Lese am Rand den Parameter der gefundenen Zeile ab.*

Es sei noch einmal betont, dass zu einem festen Parameterwert in der entsprechenden Zeile die zugehörige Verteilung (Wahrscheinlichkeitsfunktion) zu finden ist. Demgegenüber wird in dem dargelegten Algorithmus aufgrund der Beobachtung eine Spalte betrachtet, in der auf jeden Fall keine Wahrscheinlichkeitsverteilung zu finden ist.

Ein Hinweis zur Bezeichnung „ML" Dieser Algorithmus setzt das Maximum-Likelihood-Prinzip (ML-Prinzip) um und erklärt in gewisser Weise auch den Namen. M steht für Maximum zur Erinnerung an den zweiten Schritt, in dem das Maximum einer Spalte ermittelt wird. L steht für Likelihood oder Mutmaßlichkeit oder Wahrscheinlichkeit. Es wird das Maximum von Werten gesucht, die in unserem Fall Wahrscheinlichkeiten der Spalten darstellen. Jedoch bilden diese Wahrscheinlichkeiten keine Wahrscheinlichkeitsverteilung, wie man durch Summation der Werte schnell überprüfen kann. (In jeder Zeile enthält die Tabelle jedoch eine Wahrscheinlichkeitsverteilung.) Aus diesem Grund und auch für die

Übertragbarkeit auf kontinuierliche Problemfälle ist die Bezeichnung Likelihood und nicht zum Beispiel Probability gewählt worden.

Bayes-Statistik

Die klassischen statistischen Methoden behandeln den Parameter θ als unbekannte Konstante und leiten Information über θ lediglich aus den Beobachtungen ab. Im Gegensatz dazu wird bei den Bayes-Methoden der Parameter θ als Realisation einer stochastischen Größe Θ betrachtet, die eine Wahrscheinlichkeitsverteilung auf Ξ , die *A-priori-Verteilung* $G(\theta)$, besitzt. Der Gewinn, der durch die Einführung der A-priori-Verteilung erzielt wird, liegt darin, dass auch bereits vorhandenes Wissen über θ (etwa in Form von früheren Erfahrungen, Expertenwissen oder plausiblen Annahmen über Parameter) in der Analyse verwertet werden kann.
Die Bayes-Analyse wird durchgeführt, indem man die A-priori-Information $G(\theta)$ über den Parameter und die Information aus einer Beobachtung x in Form der sogenannten *A-posteriori-Verteilung* von Θ, bedingt durch x, kombiniert, und aus ihr Entscheidungen und Folgerungen θ betreffend ableitet.

Beispiel 1: *Angenommen, wir kennen den Wert p für die Wahrscheinlichkeit von „Kopf" zwar nicht, aber wir halten ein faire Münze wahrscheinlicher als eine unfaire – je fairer, umso wahrscheinlicher. Dieses „Vorurteil" lässt sich sogar quantifizieren (evt. aufgrund unserer bisherigen Erfahrungen mit Münzen aus diesem Land). Und zwar gilt:*

$$G(\theta) = \begin{cases} 6\theta(1-\theta) & \textit{wenn } 0 < \theta < 1 \\ 0 & \textit{sonst} \end{cases}$$

Aufgrund der Beobachtung müssen wir unsere A-priori-Annahme korrigieren. Die A-posteriori-Verteilung *ist die bedingte Wahrscheinlichkeit eines Wertes von θ unter Voraussetzung der gemachten Beobachtung:*

$$\begin{aligned} P(\theta|s) &= \frac{P(s|\theta)G(\theta)}{P(s)} \\ &= \frac{\theta^8(1-\theta)^2 \cdot 6\theta(1-\theta)}{P(s)} \\ &= \frac{6\theta^9(1-\theta)^3}{P(s)} \end{aligned}$$

$P(s)$, die A-priori-Wahrscheinlichkeit der Beobachtung s, kennen wir nicht, aber sie hängt nicht von θ ab. Daher können wir auch so den Wert von θ berechnen, für den die A-Posteriori-Wahrscheinlichkeit von θ maximal wird:

$$arg_{\ \theta}\ max\ P(\theta|s) = arg_{\ \theta}\ max\ 6\theta^9(1-\theta)^3 = 0.75$$

Vor dem Hintergrund unseres Vorwissens über θ ist auf der Basis der Beobachtung $\theta = 0.75$ die plausibelste Hypothese, obwohl $\theta = 0.8$ die Wahrscheinlichkeit der Beobachtung maximieren würde.

Hausaufgaben

Manning und Schütze, S. 60, Aufgabe 2.7 und 2.8