Stochastische Lernalgorithmen

Gerhard Jäger

16. April 2003

1 Wahrscheinlichkeitsrechnung

1.1 Wahrscheinlichkeitsbegriff

Klassische Definition nach Laplace

Wahrscheinlichkeit eines Ereignisses ist gegeben als Verhältnis der Anzahl günstiger Ereignisse ($n(A)$) zur Anzahl möglicher Ereignisse

$$P(A) := \frac{n(A)}{n}$$

Problem: Elementarereignisse müssen gleichwahrscheinlich sein $\Rightarrow$ im Grunde zirkulär

Statistische Definition nach von Mises

Relative Häufigkeit: Verhältnis von Anzahl günstiger Ereignisse zur Gesamtzahl der Ereignisse **bei einer konkreten Ereignisfolge**

$$h(A) := \frac{n(A)}{n}$$

$$P(A) := lim_{n\to\infty} h(A)$$

Problem:

- nicht empirisch fundierbar
- letztendlich ebenfalls zirkulär

Axiomatischer Zugang nach Kolmogoroff

Grundbegriffe:

- Ω: Menge der *Elementarereignisse*
- $S \subseteq POW(\Omega)$: *Ereignisraum*
  - nicht-leer
  - abgeschlossen unter Komplement
  - abgeschlossen unter Vereinigung

Eine *Wahrscheinlichkeitsverteilung* für einen Ereignisraum ist eine Abbildung

$$S \mapsto R$$

die die folgenden Bedingungen erfüllt:

Axiome nach Kolmogoroff

1. Für beliebige Ereignisse $A \in S : P(A) \geq 0$.
2. $P(\Omega) = 1$
3. Für beliebige Sequenzen von wechselseitig disjunkten Ereignissen $A_1, A_2, A_3, ...$:

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

Einige daraus folgende Theoreme:

Theorem 1

$$P(\emptyset) = 0$$

Theorem 2 *Für beliebige endliche Sequenzen von wechselseitig disjunkten Ereignissen* $A_1, A_2, A_3, ..., A_n$*:*

$$P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$$

Theorem 3 *Für beliebige* $A \in S$*:*

$$P(-A) = 1 - P(A)$$

Theorem 4 *Für beliebige Ereignisse* $A \in S$*:*

$$0 \leq P(A) \leq 1$$

Theorem 5 *Wenn* $A \subseteq B$*, dann* $P(A) \leq P(B)$*.*

Theorem 6 *Für beliebige Ereignisse* A *und* B*:*

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Das Gesetz der großen Zahl Die statistische Definition der Wahrscheinlichkeit kann in gewissem Sinne aus der axiomatischen Definition abgeleitet werden:

Die relative Häufigkeit eines Ereignisses ist selbst eine stochastische Größe, da sie von Versuchsreihe zu Versuchsreihe variieren kann. Man kann daher sinnvoll davon sprechen, dass die relative Häufigkeit $\frac{n(A)}{A}$ mit einer bestimmten Wahrscheinlichkeit in einem bestimmten Intervall liegt. Wenn die Wahrscheinlichkeit eines Ereignisses A gleich $P(A)$ ist, dann gilt für beliebig kleine reelle Zahlen ε:

$$lim_{n\to\infty} P\left(\left|\frac{n(A)}{n} - P(A)\right| < \varepsilon\right) = 1$$

Dieser Zusammenhang heißt ***Gesetz der großen Zahl****.*

1.2 Stochastische Unabhängigkeit und Abhängigkeit

Definition 1 (Stochastische Unabhängigkeit) *Zwei Ereignisse A und B sind* unabhängig *gdw.*

$$P(A \cap B) = P(A) \cdot P(B)$$

Theorem 7 *Wenn A und B voneinander unabhängig sind, dann auch A und $-B$, $-A$ und B, sowie $-A$ und $-B$.*

Bedingte Wahrscheinlichkeit

Definition 2 *Vorausgesetzt sei dass $P(B) > 0$. Dann ist die* bedingte Wahrscheinlichkeit von A bezüglich B *gegeben durch*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Theorem 8 *Angenommen $P(B) > 0$. Dann sind A und B unabhängig gdw.*

$$P(A|B) = P(A)$$

Bsp.: Das Ziegenproblem *In einer Gewinnshow haben Sie die Chance, ein Auto zu gewinnen. Vor Ihnen sehen Sie drei verschlossene Türen. Hinter zwei Türen sind Ziegen versteckt, hinter der dritten Ihr Traumauto. Der Moderator instruiert Sie folgerndermaßen: "Entscheiden Sie sich für eine der drei Türen (1. Wahl). Anschließend öffne ich eine der beiden anderen Türen, und zwar immer eine, hinter der eine Ziege steht. Danach dürfen Sie nochmals entscheiden (2. Wahl), welche der beiden übrig gebliebenen Türen Sie öffnen wollen."*
Ist das Auto, nachdem der Moderator die Tür geöffnet hat, eher hinter der Tür, die Sie als erste gewählt haben, zu erwarten oder hinter der anderen nicht geöffneten Tür? Anders ausgedrückt: Sollten Sie die erste Wahl beibehalten (Strategie 1) oder besser zur anderen Tür wechseln (Strategie 2)?

Bayes'sches Theorem Der Begriff der bedingten Wahrscheinlichkeit liegt probabilistischen Schlussfolgerungen zugrunde: $P(A|B)$ ist die Wahrscheinlichkeit, die man A zuweist, wenn man weiß, dass B mit Sicherheit eintritt.
$P(A|B)$ und $P(B|A)$ sind streng zu unterscheiden, hängen aber zusammen:

Theorem 9 (Bayes'sches Theorem)

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

Bsp.: *Angenommen man interessiert sich für eine seltene syntaktische Konstruktion wie z.B. parasitäre Lücken, die im Schnitt einmal in 100 000 Sätzen auftreten. Es gibt ein Perl-Skript, das diese Konstruktionen erkennt. Allerdings ist es nicht hundertprozentig zuverlässig: Wenn ein Satz eine parasitäre Lücke enthält, wird er mit nur 95%iger Wahrscheinlichkeit erkannt. Außerdem wird ein Satz, der die fragliche Konstruktion nicht enthält, mit 0,5%iger Wahrscheinlichkeit fälschlich indiziert. Angenommen das Skript markiert Satz S. Wie hoch ist die Wahrscheinlichkeit, dass S tatsächlich die fragliche Konstruktion enthält?*

Zufallsvariablen

- *Axiomatische Wahrscheinlichkeitsrechnung: Ω kann beliebige (nicht-leere) Menge sein*
- *Beschränkung auf Teilmengen von R häufig mathematisch einfacher*
- *Daher: Abbildung von Ω auf R*

Definition 3 (Zufallsvariable) *Sei S ein Wahrscheinlichkeitsraum über die Menge Ω, und P eine Wahrscheinlichkeitsverteilung über S. Eine* Zufallsvariable *(„random variable") ist eine Abbildung*

$$X : \Omega \mapsto R$$

so dass für alle $x \in R : \{\omega \in \Omega | X(\omega) = x\} \in S$

Beachte: Der Begriff „Variable" ist irreführend; Zufallsvariablen sind Funktionen und haben nichts mit Variablen im Sinne der Logik oder von Programmiersprachen zu tun!

Beispiel 1: *Sei $\Omega = \{Kopf, Zahl\}$. Wenn $X(Kopf) = 0$ und $X(Zahl) = 1$ eine Zufallsvariable.*

Beispiel 2: *Werfen zweier Würfel:* $\Omega = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$. *Die Summe der Augen auf beiden Würfeln ist eine Zufallsvariable.*

Eine Zufallsvariable X und eine Wahrscheinlichkeitsverteilung P über dem Definitionsbereich von X determinieren zusammen eine Wahrscheinlichkeitsverteilung über R.

Definition 4 (Wahrscheinlichkeitsfunktion) *Sei X eine Zufallsvariable. Die zugehörige* Wahrscheinlichkeitsfunktion *(„probability function“ oder „probability mass function“) p_X is gegeben durch*

$$p_X(x) = P(X = x) = P(\{\omega \in \Omega | X(\omega) = x\})$$

Es lässt sich zeigen, dass p_X immer eine Wahrscheinlichkeitsverteilung über R ist.

Bsp 1: *Angenommen die Münze ist fair. Dann $P(Kopf) = P(Zahl) = 0,5$. Also $p_X(0) = 0,5$ und $p_X(1) = 0,5$, sowie $p_X(x) = 0$ für alle $x \notin \{0,1\}$.*

Bsp 2:

x	$X^{-1}(x)$	$p_X(x)$
2	1, 1	1/36
3	1, 2	2/36
	2, 1	
4	1, 3	3/36
	2, 2	
	3, 1	
5	1, 4	4/36
	2, 3	
	3, 2	
	4, 1	
6	1, 5	5/36
	2, 4	
	3, 3	
	4, 2	
	5, 1	
7	1, 6	6/36
	2, 5	
	3, 4	
	4, 3	
	5, 2	
	6, 1	

x	$X^{-1}(x)$	$p_X(x)$
8	2, 6	5/36
	3, 5	
	4, 4	
	5, 3	
	6, 2	
9	3, 6	4/36
	4, 5	
	5, 4	
	6, 3	
10	4, 6	3/36
	5, 5	
	6, 4	
11	6, 5	2/36
	5, 6	
12	6, 6	1/36

*(Für die Zwecke der statistischen Sprachverarbeitung kann man sich auf **diskrete** Zufallsvariable beschränken, also Funktionen, deren Wertebereich endlich oder abzählbar ist. Im allgemeinen Fall können Zufallsvariable auch kontinuierlich sein. Statt von einer Wahrscheinlichkeitsfunktion spricht man dann von einer **Wahrscheinlichkeitsdichtefunktion** („probability density function"). Für den allgemeinen Fall wärem im Folgenden alle Summenzeichen durch bestimmte Integrale zu ersetzen.)*

Kennwerte einer Zufallsvariablen

Erwartungswert

Der Erwartungswert einer Zufallsvariable ist das zu erwartende arithmetische Mittel des Wertes der Variablen bei sehr vielen Versuchen.

Definition 5 (Erwartungswert)

$$E(X) = \sum_x x \cdot p_X(x)$$

Bsp. 1: *Also*

$$E(X) = 0,5 \cdot 0 + 0,5 \cdot 1 = 0,5$$

Bsp 2:

$$\begin{aligned} E(X) &= 1/36 \cdot 2 + 2/36 \cdot 3 + 3/36 \cdot 4 + 4/36 \cdot 5 + \\ &\quad 5/36 \cdot 6 + 6/36 \cdot 7 + 5/36 \cdot 8 + 4/36 \cdot 9 + \\ &\quad 3/36 \cdot 10 + 2/36 \cdot 11 + 1/36 \cdot 12 \\ &= 7 \end{aligned}$$

Bsp. 3: *Einfacher Wurf mit einem fairen Würfel; X gibt den Zahlenwert der gewürfelten Augen an.*

$$E(X) = 1/6 \cdot 1 + 1/6 \cdot 2 + 1/6 \cdot +1/6 \cdot 4 + 1/6 \cdot +1/6 \cdot 5 + 1/6 \cdot 6 = 3,5$$

Jede Funktion $g : R \mapsto R$ bildet eine Zufallsvariable auf eine neue Zufallsvariable ab. I.a.W., wenn X eine Zufallsvariable ist, dann auch $Y = g(X)$. Der Erwartungswert der abgeleiten Variable ergibt sich durch

$$E(Y) = \sum_x g(x) \cdot p_X(x)$$

Insbesondere gilt:

Theorem 10 *Wenn g eine lineare Funktion ist, also $g(x) = ax + b$, dann*

$$E(g(X)) = aE(X) + b$$

Summe mehrerer Zufallsvariablen: *Der Erwartungswert linearer Kombinationen von Zufallsvariablen berechnet sich nach*

Theorem 11

$$E(a \cdot X + b \cdot Y) = aE(X) + bE(Y)$$

Beachte: *Dieses Gesetz gilt auch, wenn X und Y nicht unabhängig sind!*

Produkt mehrerer Zufallsvariablen

Theorem 12 *Wenn X und Y stochastisch unabhängig sind, dann gilt*

$$E(X \cdot Y) = E(X) \cdot E(Y)$$

Varianz

Als Kennwert, wie unterschiedlich *die Werte einer Zufallsvariable X sind, wird die* Varianz *der variable berechnet. Sie gib and,* wie weit *die WErte von X vom Erwrtungswert abweichen. Genauer: Die Varianbz gibt die* durchschnittliche quadrierte Abweichung *vom Erwartungswert an.*

Definition 6 (Varianz einer Zufallsvariable)

$$Var(X) = E((X - E(X))^2)$$

Die Varianz ist ein Maß für die Streuung *einer Zufallsvariable. Sie entspricht der zue erwartenden durchschnittlichen quadratischen Abweichung vom Erwartungswert.*
Es gilt:

$$E((X - E(X))^2) = E(X^2) - E^2(X)$$

Bsp 1:

$$Var(X) = E((X - 0,5)^2) = \sum_x (x - 0,5)^2 \cdot p_X(x) = 0,25 \cdot 0,5 + 0,25 \cdot 0,5 = 0,25$$

Bsp 2:

$$\begin{aligned} Var(X) &= E(X^2) - E^2(X) \\ &= 4 \cdot 1/36 + 9 \cdot 2/36 + 16 \cdot 3/36 + 25 \cdot 4/36 + \\ &\quad 36 \cdot 5/36 + 49 \cdot 6/36 + 64 \cdot 5/36 + 81 \cdot 4/36 + \\ &\quad 100 \cdot 3/36 + 121 \cdot 2/36 + 144 \cdot 1/36 \\ &\quad -49 \\ &= 35/6 \end{aligned}$$

Hausaufgaben

1. *Beim Münzwurf spielen Sie um Geld: Werfen Sie eine Münze und es fällt „Kopf“, so zahlen Sie 1 Euro, fält „Zahl“, so erhalten Sie 1,50 Euro. Wie groß ist der Erwartungswert, wenn Sie 10 mal die Münze werfen? Wie groß ist die Varianz der Gewinnausschüttung? Wie ändert sich dies Varianz, wenn Gewinn und Verlust verdoppelt werden?*

   Würden Sie das Spiel noch mitmachen, wenn die Münze nicht fair wäre, sondern die Wahrscheinlichkeiten P(„Kopf“) = 1/3, P(„Zahl“)=2/3 vorliegen würden?

2. *Sie warten auf den wichtigen Anruf eines Freundes, der ganz sicher irgendwann zwischen 17:00 und 18:00 erfolgen soll, jeder Zeitpunkt dazwischen ist gleich wahrscheinlich. Sie wollen aber von 17:20 bis 17:50 ihre Lieblingssendung im Fernsehen ungestört sehen. Wie wahrscheinlich ist das Ereignis, dass Ihr Freund sich bis 17:20 gemeldet hat? Wie wahrscheinlich ist es, dass Sie wegen des Anrufs nicht ungestört fernsehen können? Mit welcher durchschnittlichen Wartezeit von 17:00 an müssen Sie rechnen? Wie wahrscheinlich ist ein Anruf in den letztenn 10 Minuten, wenn es bereits nach halb sechs ist und der Anruf noch nicht erfolgt ist?* (Hinweis: Nehmen Sie der Einfachheit halber an, dass der Anruf zu einer vollen Minute erfolgt.)

3. *Für die Varianz einer Zufallsvariable gilt die Rechenregel:*

$$Var(aX) = a^2 Var(X)$$

   Beweisen Sie diesen Sachverhalt mit Hilfe der oben angegebenen Rechenregeln für den Erwartungswert!

4. *Wenn X und Y stochastisch unabhängig sind, gilt die Regel*

$$Var(X + Y) = Var(X) + Var(Y)$$

   Leiten Sie dieses Gesetz aus den oben gegebenen Rechenregeln für den kombinierten Erwartungswert von unabhängigen Zufallsvariablen ab!

5. Lektüre: *Manning und Schütze, Abschnitt 2.1*